В. Л. Борщ, Ю. П. Совіт

МАТЕМАТИЧНІ МЕТОДИ В БІОЛОГІЇ. ОСНОВИ СТАТИСТИКИ

2008

Міністерство освіти і науки України
Дніпропетровський національний університет
імені Олеся Гончара

90-річчю ДНУ присвячується

В. Л. Борщ, Ю. П. Совіт

МАТЕМАТИЧНІ МЕТОДИ В БІОЛОГІЇ. ОСНОВИ СТАТИСТИКИ

Конспект лекцій

Ухвалено вченою радою університету
протокол № 6 від 20.12.07

Дніпропетровськ
РВВ ДНУ
2008

УДК 519.22/.25 (075.8)
Б83

Рецензенти: д-р фіз.-мат. наук, проф. М. В. Поляков
д-р техн. наук, проф. Б. І. Мороз

Б83 Борщ, В. Л. Математичні методи в біології. Основи статистики [Текст]: конспект лекцій / В. Л. Борщ, Ю. П. Совіт. – Д.: РВВ ДНУ, 2008. – 144 с.

Видання містить продовження лекцій з дисципліни „Математичні методи в біології“, початок яких – „Математичні методи в біології. Основи теорії ймовірностей“ – виданий в 2007 р. Лекції узагальнюють досвід викладання цієї дисципліни студентам біолого-екологічного факультету і факультету заочної та дистанційної освіти ДНУ. Матеріалами конспекту лекцій можна послуговуватися в процесі вивчення інших дисциплін, де застосовується числова обробка дослідних даних.

© Борщ В. Л., Совіт Ю. П., 2008

Передмова

Пропоноване видання продовжує виклад конспекту лекцій „Математичні методи в біології. Основи теорії ймовірностей“. Фактично ці видання вміщують повний лекційний курс дисципліни „Математичні методи в біології“. Для зручності користування конспектом лекцій нумерацію розділів, формул, таблиць і рисунків, прийняту в „Основах теорії ймовірностей“, ми продовжили й в „Основах математичної статистики“.

На початку нашого конспекту ми докладно репрезентували власні погляди на викладення дисципліни „Математичні методи в біології“ для студентів біолого-екологічного факультету та факультету заочної та дистанційної освіти, відтак обмежимося лише вказівкою на індуктивний стиль викладення, близький до запропонованого Мізесом (1930). Індуктивний підхід ґрунтується на узагальненні фактів у вигляді теорії, тобто на прямуванні від „часткового“ до „загального“. В даному лекційному курсі немає чіткого розмежування на теорію ймовірностей і статистику, отже, назва „Основи теорії ймовірностей“ для початку конспекту обумовлена введенням такого поняття, як імовірність, з якого починається звичайний курс теорії ймовірностей, а назва „Основи статистики“ для завершення конспекту ґрунтується на тому, що в останньому розділі викладаються алгоритми перевірки статистичних гіпотез, тобто тема, якою звичайно завершуються стандартні курси математичної статистики.

Логічна й тематична побудова лекцій є така.

У розділі 7 ми розглядаємо системи випадкових величин, включаючи дискретні, неперервні й інтервальні. Інтервальні випадкові величини, власне, є такі, що спостерігаються в досліді замість ідеалізованих неперервних, через обмежену точність вимірювальних приладів. Одночасно вони є проміжні між дискретними і неперервними. Отже, багато понять, які вводяться для дискретних величин, можна легко перенести і на інтервальні. Якщо розглядати границю інтервальних величин за зменшення ширини інтервалів, то можна наблизитися як завгодно близько до неперервних. Такий підхід, на наш погляд, найбільш придатний для викладення дуже непростих понять, пов’язаних з діями над неперервними величинами. Дійсно, навіть для введення суми двох неперервних величин, заданих своїми функціями розподілу, необхідно використовувати таке поняття, як згортка (Турчин, 2006). На нашу думку, викладення „згортки“ нашим студентам є недоцільним і майже неможливим з огляду на обмеженість програми з вищої математики. Отже, застосування переходу від неперервних величин до інтервальних, виконання дій над інтервальними величинами, граничне повернення від інтервальної величини-результата до неперервної величини-результата є, можливо, єдиним способом викладення дій над неперевними величинами. До речі, саме в такий спосіб ми вводимо поняття незалежності для неперевних величин, узагальнюючи наше поняття незалежності для випадкових подій, уведене в розділі 3. Далі в розділі 7 ми вводимо числові характеристики дій над випадковими величинами, приділяючи увагу різниці між числовими характеристиками залежних і незалежних величин, яка втілюється в важливому понятті коваріації. Адже саме коваріація дозволяє відрізняти залежні і незалежні випадкові величини під час перевірки статистичних гіпотез. Наприкінці розділу 7 ми вводимо такі

поняття, як кореляція і регресія. Перше дозволяє „вимірювати“ лінійні зв’язки між випадковими величинами, друге – спрощувати ці зв’язки, через залежність в середньому одної випадкової величини від значень іншої. Докладно ми розглядаємо лише лінійну регресію. Крім методу обчислення коефіцієнтів лінійної регресії обговорюється і вирівнювання даних, яке формально, під кутом зору застосованих формул, не відрізняється від регресії, але насправді є зовсім іншою задачею.

У розділі 8 ми знову повертаємося до нормального розподілу неперервних випадкових величин. Це зумовлено тим, що дії над випадковими величинами вже викладені, отже, можна ввести знамениту центральну граничну теорему. На нашу думку, якщо в подібних курсах її й уводять, то лише для назви, оскільки без уведення дій над випадковими величинами, особливо неперервних, її введення недоцільне. Далі на основі нормального розподілу вводимо (звичайним способом) стандартні розподіли – Пірсона, Фішера, Стьюдента. Водночас ми робимо спробу ввести ці розподіли і як граничні вібіркові розподіли. І це можливо, оскільки всі поняття, засновані на первинній обробці вибіркових даних (вибіркова функція густини і вибіркова функція розподілу), були введені ще в розділі 4. На нашу думку, у такий спосіб ми певним чином відходимо від формального і досить поширеного підходу, який майже не сприймається нашими студентами.

У розділі 9 ми вводимо поняття, пов’язані з перевіркою статистичних гіпотез, в загальних рисах використовуючи ідеї, які виклав у своєму відомому підручнику Тутубалін (1986). Крім того, наводимо всі основні види гіпотез, що зустрічаються за первинної обробки гіпотез на рівні, достатньому для виконання курсових і дипломних робіт.

В останньому розділі мова йде про інтервальні оцінки, іншими словами, про точність перевірки статистичних гіпотез. Ми зробили спробу дати алгоритми інтервальних оцінок для усіх параметрів, розглянутих у попередніх розділах.

Насамкінець зазначимо: ми не ставили за мету, навіть стосовно важливих результатів вдаватися до класичних доведень, натомість обмежувалися достатнім, на наш погляд, „здоровим сенсом“. У тих нечисленних випадках, де доведення зустрічаються, ми позначали їх завершення знаком ■. Знак ● застосовували для позначки завершення прикладів, а знак ▲ – завершення зауважень. На зауваження, подані у конспекті лекцій, доцільно звернути увагу з огляду на те, що вони не є штучними, тобто їх не слід вважати навмисними виділеннями в тексті. Переважна їх більшість є відповідями на запитання студентів під час лекцій або аналізом їх помилок на практичних заняттях чи у відповідях під час іспитів.

На всі запозичення, включаючи і змістовні, в тексті лекцій зроблені відповідні посилання.

7. Системи двох випадкових величин

7.1. Поняття системи випадкових величин

Системи випадкових величин утворюються за відповідної організації досліду. Найпростіший приклад такого досліду – вимірювання у кожного об'єкта генеральної сукупності, що витягається у вибіркову сукупність, двох або більше числових ознак. У цьому випадку результатом досліду будуть заздалегідь невідомі значення числових ознак. Але можна припустити, що між цими випадковими (для дослідника) значеннями існує внутрішній зв'язок, саме він і дозволяє вважати ознаки, що досліджуються, складовими системи випадкових величин. Інші ситуації виникають у разі організації дослідів за схемою „контроль група – дослідна група“ або „до впливу – після впливу“. Їх ми будемо розглядати в розд. 10.

7.2. Спільний розподіл двох дискретних випадкових величин

Нехай результати n дослідів

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}, \tag{7.1}$$

у яких вимірювались значення двох дискретних кількісних ознак X і Y, зображені у вигляді записів лабораторного журналу, як показано в табл. 7.1.

Таблиця 7.1

Дослід	X	Y
1	x_1	y_1
2	x_2	y_2
…	…	…
i	x_i	y_i
…	…	…
n	x_n	y_n

Аналогічно побудові вибіркового ряду розподілу частот однієї дискретної випадкової величини (див. розд. 4.7) розподілимо значення $\{x_i,\ i = 1, \ldots, n\}$, яких набула в досліді величина X, і значення $\{y_i,\ i = 1, \ldots, n\}$, набуті в досліді величиною Y, відповідно за групами повторюваних значень $\{x_k,\ k = 1, \ldots, K\}$ і $\{y_l,\ l = 1, \ldots, L\}$. Відмінність обробки вибіркових значень двох дискретних випадкових величин від обробки вибіркових значень однієї дискретної випадкової величини обумовлена тим, що в кожному досліді значення величин X і Y формувалися парами (7.1). Отже,

у разі розподілу значень з лабораторного журналу, тобто значень $\{x_i,\ i=1,\ldots,n\}$ в групи $\{x_k,\ k=1,\ldots,K\}$ і значень $\{y_i,\ i=1,\ldots,n\}$ в групи $\{y_l,\ l=1,\ldots,L\}$, має зберігатися зв'язок між значеннями x_i і y_i, що утворюють пари (x_i, y_i), в результаті матимемо не дві одновимірні таблиці-ряди, а одну двовимірну – **таблицю спільного розподілу частот дискретних випадкових величин** X і Y, зображену у вигляді табл. 7.2.

Таблиця 7.2

	$n_{k=1}$	$n_{k=2}$	…	n_k	…	$n_{k=K}$		
	$n_{1,\circ}$	$n_{2,\circ}$	…	$n_{k,\circ}$	…	$n_{K,\circ}$		
$y_{l=L}$	$n_{1,L}$	$n_{2,L}$	…	$n_{k,L}$	…	$n_{K,L}$	$n_{\circ,L}$	$n_{l=L}$
…	…	…	…	…	…	…	…	…
y_l	$n_{1,l}$	$n_{2,l}$	…	$n_{k,l}$	…	$n_{K,l}$	$n_{\circ,l}$	n_l
…	…	…	…	…	…	…	…	…
$y_{l=2}$	$n_{1,2}$	$n_{2,2}$	…	$n_{k,2}$	…	$n_{K,2}$	$n_{\circ,2}$	$n_{l=2}$
$y_{l=1}$	$n_{1,1}$	$n_{2,1}$	…	$n_{k,1}$	…	$n_{K,1}$	$n_{\circ,1}$	$n_{l=1}$
	$x_{k=1}$	$x_{k=2}$	…	x_k	…	$x_{k=K}$		

Клітинки цієї таблиці на перетині стовпчиків із заголовками $\{x_k,\ k=1,\ldots,K\}$ (заголовки містяться в нижньому рядку) і рядків із заголовками $\{y_l,\ l=1,\ldots,L\}$ (заголовки містяться в лівому стовпчику) заповнені числами $n_{k,l}$, називаними **частотами клітинок.** Частота $n_{k,l}$ клітинки з номером стовпчика k і номером рядка l показує, скільки разів серед вибіркових пар (x_i, y_i) зустрічається збіг $x_i = x_k,\ y_i = y_l$.

Сума частот $n_{k,l}$ в стовпчику k дорівнює частоті появи значення x_k величини X без урахування появи значень дискретної випадкової величини Y. Ці частоти записані в другому зверху рядку таблиці спільного розподілу частот і позначені $n_{k,\circ}$:

$$n_{k,\circ} = \sum_{l=1}^{L} n_{k,l} \,. \tag{7.2}$$

Кружечок позначає, що індекс l зникає, оскільки він використаний у сумі (7.2). Якщо припустити, що індекс k належить до дискретної випадкової величини X, то частоти $n_{k,\circ}$ можна записати у звичній формі:

$$n_{k,\circ} = n_k \,, \tag{7.3}$$

тобто так, як у разі запису вибіркового ряду розподілу частот дискретної випадкової величини X, коли в досліді ніякі інші дискретні випадкові величини не розглядаються (див. розд. 4.7). Частоти n_k записані в першому зверху рядку табл. 7.2. Аналогічно, шляхом обчислення сум частот $n_{k,l}$ по рядках l, знаходимо частоти появи значень y_l дискретної випадкової величини Y:

$$n_{\circ,l} = n_l = \sum_{k=1}^{K} n_{k,l} \,. \tag{7.4}$$

Ці частоти записані в перших двох справа стовпчиках табл. 7.2.

Якщо з табл. 7.2 взяти рядок значень x_k і рядок відповідних їм частот n_k, а також стовпчик значень y_l і стовпчик відповідних їм частот n_l, то матимемо два вибіркові ряди розподілу частот дискретних випадкових величин X і Y, зображені відповідно в табл. 7.3 і 7.4.

Таблиця 7.3

x_k	$x_{k=1}$	$x_{k=2}$	...	x_k	...	$x_{k=K}$
n_k	$n_{k=1}$	$n_{k=2}$	...	n_k	...	$n_{k=K}$

Таблиця 7.4

y_l	$y_{l=1}$	$y_{l=2}$	...	y_l	...	$y_{l=L}$
n_l	$n_{l=1}$	$n_{l=2}$	...	n_l	...	$n_{l=L}$

Ці ряди розподілів частот можна було б безпосередньо одержати з лабораторного журналу (табл. 7.1) або, що те ж саме, з вибіркових пар $\{(x_i, y_i),\ i = 1,\ldots,n\}$, якби обробка значень $\{x_i,\ i = 1,\ldots,n\}$ і $\{y_i,\ i = 1,\ldots,n\}$ про-

водилася порізно. За умови спільної обробки пар вибіркових значень $\{(x_i, y_i),\ i = 1,\ldots,n\}$ ряди розподілів частот розташовуються в крайніх рядках (нижньому і верхньому для величини X) і крайніх стовпчиках (лівому і правому для величини Y) таблиці спільного розподілу частот величин X і Y. За цієї причини табл. 7.3 і 7.4 називаються **маргінальними**[1] **рядами розподілу частот дискретних випадкових величин** X і Y.

З наведених міркувань випливає справедливість **прямого твердження** – з таблиці спільного розподілу частот величин X і Y можна одержати маргінальні ряди розподілу частот величин X і Y.

Зворотне твердження є невірне, тобто з маргінальних рядів розподілу частот величин X і Y не можна одержати таблицю їхнього спільного розподілу частот. Інакше кажучи, якщо для даного досліду відомо, скільки разів зустрічалися значення $\{x_k,\ k = 1,\ldots,K\}$ в стовпчику X лабораторного журналу і значення $\{y_l,\ l = 1,\ldots,L\}$ в стовпчику Y лабораторного журналу, то сказати, скільки разів у цьому досліді зустрічалася пара значень (x_k, y_l), тобто скільки разів з'являвся у лабораторному журналі рядок $x_i = x_k,\ y_i = y_l$, не можна.

Для доведення скористаємося визначенням маргінальних частот, виписавши відповідні вирази (7.2), (7.3) для всіх значень n_k і відповідні вирази (7.4) для всіх значень n_l:

$$\begin{cases} n_{1,1} + n_{1,2} + \ldots + n_{1,l} + \ldots + n_{1,L} = n_{k=1} \\ n_{2,1} + n_{2,2} + \ldots + n_{2,l} + \ldots + n_{2,L} = n_{k=2} \\ \qquad\qquad\qquad \ldots \\ n_{k,1} + n_{k,2} + \ldots + n_{k,l} + \ldots + n_{k,L} = n_k \\ \qquad\qquad\qquad \ldots \qquad \ldots \\ n_{K,1} + n_{K,2} + \ldots + n_{K,l} + \ldots + n_{K,L} = n_{k=K} \end{cases}, \tag{7.5}$$

$$\begin{cases} n_{1,1} + n_{2,1} + \ldots + n_{k,1} + \ldots + n_{K,1} = n_{l=1} \\ n_{1,2} + n_{2,2} + \ldots + n_{k,2} + \ldots + n_{K,2} = n_{l=2} \\ \qquad\qquad\qquad \ldots \\ n_{1,l} + n_{2,l} + \ldots + n_{k,l} + \ldots + n_{K,l} = n_l \\ \qquad\qquad\qquad \ldots \qquad \ldots \\ n_{1,L} + n_{2,L} + \ldots + n_{k,L} + \ldots + n_{K,L} = n_{l=L} \end{cases}. \tag{7.6}$$

Вирази (7.5) і (7.6) утворюють систему $K+L$ лінійних рівнянь щодо невідомих частот $n_{k,l}$, число яких дорівнює числу клітинок у табл. 7.2, тобто $K \cdot L$. Необхідна

[1] Від лат. *margo* – край, межа.

умова існування єдиного розв'язку системи лінійних рівнянь (7.5) і (7.6) полягає в рівності числа невідомих $n_{k,l}$ числу рівнянь, тобто у виконанні рівності $K+L=K\,L$, звідки

$$K = \frac{L}{L-1} \quad \text{або} \quad L = \frac{K}{K-1}\ . \tag{7.7}$$

Таким чином, у випадку однозначного розв'язання системи рівнянь відносно частот $n_{k,l}$ за відомими маргінальними частотами n_k і n_l число K значень дискретної випадкової величини X і число L значень дискретної випадкової величини Y зв'язані між собою рівністю (7.7). Оскільки K і L є натуральними числами, то $K=L=2$. Отже, за маргінальними частотами n_k і n_l можна визначити частоти $n_{k,l}$ тільки для таблиці 2×2. У загальному випадку з одновимірних таблиць – рядів розподілу частот (табл 7.3 і 7.4) – двовимірну таблицю – таблицю частот (табл. 7.2) – одержати не можна. ■

Якщо всі частоти $n_{k,l}$, n_k і n_l (табл. 7.2) розділити на число дослідів n, то матимемо **таблицю спільного розподілу відносних частот дискретних випадкових величин** X і Y, відображену в табл. 7.5.

Таблиця 7.5

	$\frac{n_{k=1}}{n}$	$\frac{n_{k=2}}{n}$	…	$\frac{n_k}{n}$	…	$\frac{n_{k=K}}{n}$	
$y_{l=L}$	$\frac{n_{1,L}}{n}$	$\frac{n_{2,L}}{n}$	…	$\frac{n_{k,L}}{n}$	…	$\frac{n_{K,L}}{n}$	$\frac{n_{l=L}}{n}$
…	…	…	…	…	…	…	…
y_l	$\frac{n_{1,l}}{n}$	$\frac{n_{2,l}}{n}$	…	$\frac{n_{k,l}}{n}$	…	$\frac{n_{K,l}}{n}$	$\frac{n_l}{n}$
…	…	…	…	…	…	…	…
$y_{l=2}$	$\frac{n_{1,2}}{n}$	$\frac{n_{2,2}}{n}$	…	$\frac{n_{k,2}}{n}$	…	$\frac{n_{K,2}}{n}$	$\frac{n_{l=2}}{n}$
$y_{l=1}$	$\frac{n_{1,1}}{n}$	$\frac{n_{2,1}}{n}$	…	$\frac{n_{k,1}}{n}$	…	$\frac{n_{K,1}}{n}$	$\frac{n_{l=1}}{n}$
	$x_{k=1}$	$x_{k=2}$	…	x_k	…	$x_{k=K}$	

Якщо в умовах даного досліду спостерігається стабілізація відносних частот $n_{k,l}/n$, n_k/n і n_l/n, тоді ці частоти можна замінити статистичними ймовірностями $p^*_{k,l}$, p^*_k і p^*_l. В результаті маємо **таблицю спільного розподілу (статистичних ймовірностей) дискретних випадкових величин** X і Y, відображену у вигляді табл. 7.6.

Таблиця 7.6

	$p^*_{k=1}$	$p^*_{k=2}$	…	p^*_k	…	$p^*_{k=K}$	
$y_{l=L}$	$p^*_{1,L}$	$p^*_{2,L}$	…	$p^*_{k,L}$	…	$p^*_{K,L}$	$p^*_{l=L}$
…	…	…	…	…	…	…	…
y_l	$p^*_{1,l}$	$p^*_{2,l}$	…	$p^*_{k,l}$	…	$p^*_{K,l}$	p^*_l
…	…	…	…	…	…	…	…
$y_{l=2}$	$p^*_{1,2}$	$p^*_{2,2}$	…	$p^*_{k,2}$	…	$p^*_{K,2}$	$p^*_{l=2}$
$y_{l=1}$	$p^*_{1,1}$	$p^*_{2,1}$	…	$p^*_{k,1}$	…	$p^*_{K,1}$	$p^*_{l=1}$
	$x_{k=1}$	$x_{k=2}$	…	x_k	…	$x_{k=K}$	

Аналогічно розгляду таблиці спільного розподілу частот $n_{k,l}$ з таблиць спільного розподілу відносних частот $n_{k,l}/n$ і статистичних ймовірностей $p^*_{k,l}$ можна одержати маргінальні ряди розподілу відносних частот n_k/n і n_l/n і ймовірностей p^*_k і p^*_l дискретних випадкових величин X і Y. Зворотне твердження не є вірне, що доводиться так само, як і у випадку таблиці спільного розподілу частот $n_{k,l}$.

7.3. Незалежні дискретні випадкові величини

У розд. 3.16 подано визначення незалежності для двох явищ, застосовуване для випадкових подій. Скористуємося цим визначенням незалежності для двох дискретних випадкових величин X і Y. Тепер досліджувані процеси (див. розд. 3.16) такі, що жоден їз них не можна „виключити“ – випадкові величини, у даному випадку дискретні, завжди набувають в досліді які-небудь значення із множини допустимих.

Припустимо, що умови досліду такі, що дискретна випадкова величина X набуває одне і те ж значення x_k. Досягнути цього можна, залишивши в лабораторному журналі (табл. 7.1) результати тільки тих дослідів, в яких дискретна випадкова величина X набуває обраного значення x_k. Розглянемо, що відбувається зі спостережуваними в досліді значеннями дискретної випадкової величини Y. Існує кілька можливих сценаріїв для результатів такого досліду. Проаналізуємо їх докладно.

1. Дискретна випадкова величина Y у разі повторення досліду набуває одне і те ж значення:

а) якщо це значення одне і те ж за всіх можливих значень x_k, то Y – стала;

б) якщо це значення змінюється за інших значень x_k, то дискретна випадкова величина Y функціонально пов'язана з дискретною випадковою величиною X, інакше кажучи, $Y=Y(X)$.

2. Дискретна випадкова величина Y у разі повторення досліду набуває різні значення. Побудуємо для величини Y ряди розподілу частот і відносних частот у вигляді відповідних стовпчиків із заголовками x_k (табл. 7.2 і 7.5). Якщо дослід проведений достатню для стабілізації відносних частот кількість разів, то можна сподіватись, що значення $p^*_{k,l}$ (тобто стовпчик із заголовком x_k в табл. 7.6), спільно зі значеннями y_l (тобто крайній зліва стовпчик цієї ж табл. 7.6), утворять ряд розподілу величини Y. А втім це не так, оскільки за умовою досліду дискретна випадкова величина X є сталою, тобто $\{X = x_k\}$. Отже, для того щоб одержати ряд розподілу дискретної випадкової величини Y за цієї умови, потрібно обчислювати відносні частоти інакше, ніж у табл. 7.6. Оскільки число дослідів, у яких виконувалася умова $\{X = x_k\}$, дорівнює n_k, то відносні частоти для побудови ряду розподілу дискретної випадкової величини Y з урахуванням умови досліду потрібно обчислювати так:

$$\frac{n_{k,l}}{n_k} = \frac{n_{k,l}}{n}\frac{n}{n_k} = \frac{n_{k,l}/n}{n_k/n} \quad . \tag{7.8}$$

Оскільки умова стабілізації відносних частот виконується (число дослідів n досить велике), то в стовпчику k знаходяться значення $p^*_{k,l}/p^*_k$. Перш ніж уводити для них позначення, згадаємо формулу умовної ймовірності (див. розд. 3.13), з якої випливає, що стовпчик k, який складається з таких значень, являє собою умовні ймовірності того, що дискретна випадкова величина Y набуває свої значення y_l за настання події $\{X = x_k\}$. Позначимо ці умовні ймовірності так:

$$p_l^*\Big|_{\{X=x_k\}} = p_{l|k}^* = \frac{p_{k,l}^*}{p_k^*}. \qquad (7.9)$$

Зведемо умовні ряди розподілу дискретної випадкової величини Y у вигляді табл. 7.7 (позначені сірим кольором).

Таблиця 7.7

	$p_{l\|k=1}^*$	$p_{l\|k=2}^*$	…	$p_{l\|k}^*$	…	$p_{l\|k=K}^*$	
$y_{l=L}$	$p_{L\|1}^*$	$p_{L\|2}^*$	…	$p_{L\|k}^*$	…	$p_{L\|K}^*$	$p_{l=L}^*$
…	…	…	…	…	…	…	…
y_l	$p_{l\|1}^*$	$p_{l\|2}^*$	…	$p_{l\|k}^*$	…	$p_{l\|K}^*$	p_l^*
…	…	…	…	…	…	…	…
$y_{l=2}$	$p_{2\|1}^*$	$p_{2\|2}^*$	…	$p_{2\|k}^*$	…	$p_{2\|K}^*$	$p_{l=2}^*$
$y_{l=1}$	$p_{1\|1}^*$	$p_{1\|2}^*$	…	$p_{1\|k}^*$	…	$p_{1\|K}^*$	$p_{l=1}^*$
	$x_{k=1}$	$x_{k=2}$	…	x_k	…	$x_{k=K}$	

Пари, утворені першим зліва стовпчиком значень y_l дискретної випадкової величини Y і стовпчиками ймовірностей $\{p_{l|k}^*,\ \ l=1,\ldots,L\}$, $k=1,\ldots,K$, складають **умовні ряди розподілів (статистичних ймовірностей)**.

Перший справа стовпчик p_l^* – це ймовірності, обчислені за настання стабілізації відносних частот звичайним способом, тобто $p_l^* = n_l/n$. Тому даний стовпчик спільно зі стовпчиком значень y_l утворює звичайний ряд розподілу дискретної випадкової величини Y. В таблиці умовних рядів розподілу величини Y можливі дві різні ситуації.

1. Умовні ряди розподілу величини Y за різних значень величини X не відрізняються один від одного. Це означає, що експериментатор не помітить розбіжності у ймовірностях появи значення величини Y у довгих серіях дослідів (для забезпечення стабілізації відносних частот), що відрізняються значеннями величини X. Тоді є підстави вважати, що дискретна випадкова величина Y **не залежить** від дискретної випадкової величини X.

2. Умовні ряди розподілу величини Y за різних значень величини X відрізняються один від одного. Тоді є підстави вважати, що дискретна випадкова величина Y **залежить** від дискретної випадкової величини X.

Підіб'ємо підсумок наших міркувань, вдаючись до декількох визначень.

Вибіркові ряди частот, відносних частот і ряд розподілу дискретної випадкової величини Y називаються **умовними**, якщо вони одержані за настання події $\{X = x_k\}$. Умовні відносні частоти обчислюються діленням частот $n_{k,l}$ не на кількість усіх дослідів n, а на кількість дослідів n_k, у яких наставала подія $\{X = x_k\}$. Умовні ймовірності $p^*_{l|k}$ (7.5) обчислюються за формулою (3.14) (див. розд. 3.13). Якщо дискретна випадкова величина X має K значень, то в дискретної випадкової величини Y відповідно є K умовних рядів частот, відносних частот і рядів розподілів.

Дискретна випадкова величина Y називається **незалежною** від дискретної випадкової величини X (ще говорять, що дискретна випадкова величина Y не залежить від дискретної випадкової величини X), якщо умовні ряди розподілу дискретної випадкової величини Y за різних значень дискретної випадкової величини X збігаються.

Із цього визначення незалежності можна відразу ж одержати ще одне. Для цього обчислимо безумовну ймовірність p^*_l того, що дискретна випадкова величина Y набуває значення y_l. Зобразимо подію $\{Y = y_l\}$ у вигляді суми несумісних подій

$$\{Y = y_l\} = \{Y = y_l\}\Omega =$$

$$= \{Y = y_l\}(\{X = x_1\} + \{X = x_2\} + \ldots + \{X = x_k\} + \{X = x_K\}) =$$

$$= \{Y = y_l\}\{X = x_1\} + \ldots + \{Y = y_l\}\{X = x_k\} + \ldots + \{Y = y_l\}\{X = x_K\} \ . \qquad (7.10)$$

Тоді, за правилом додавання ймовірностей для декількох несумісних подій (див. розд. 3.11),

$$P(\{Y = y_l\}) = P(\{Y = y_l\}\{X = x_1\} + \ldots + \{Y = y_l\}\{X = x_K\}) =$$

$$= P(\{Y = y_l\}\{X = x_1\} + \ldots + \{Y = y_l\}\{X = x_K\}) \ . \qquad (7.11)$$

Застосовуючи для умовних ймовірностей $P(\{Y = y_l\}\{X = x_k\})$ позначення $p^*_{l|k}$, цей результат запишемо так:

$$p^*_l = p^*_{1,l} + p^*_{2,l} + \ldots + p^*_{K,l} \, . \qquad (7.12)$$

Для кожної із ймовірностей у правій частині (7.9) застосуємо правило добутку ймовірностей для двох довільних подій (див. розд. 3.15).

$$\begin{cases} \overset{*}{p}_{1,l} = \overset{*}{p}_{k=1}\ \overset{*}{p}_{l|k=1}, \\ \overset{*}{p}_{2,l} = \overset{*}{p}_{k=2}\ \overset{*}{p}_{l|k=2}, \\ \ldots \\ \overset{*}{p}_{k,l} = \overset{*}{p}_{k}\ \overset{*}{p}_{l|k}, \\ \ldots \\ \overset{*}{p}_{K,l} = \overset{*}{p}_{k=K}\ \overset{*}{p}_{l|k=K}. \end{cases} \tag{7.13}$$

Тоді

$$\overset{*}{p}_{l} = \sum_{k=1}^{K} (\overset{*}{p}_{k}\ \overset{*}{p}_{l|k}).$$

За визначенням незалежності дискретної випадкової величини Y від дискретної випадкової величини X всі значення $\overset{*}{p}_{l|k}$ однакові, тобто

$$\overset{*}{p}_{l} = \sum_{k=1}^{K} (\overset{*}{p}_{l|k'}\ \overset{*}{p}_{k}) = \overset{*}{p}_{l|k'} \underbrace{\left(\sum_{k=1}^{K} \overset{*}{p}_{k} \right)}_{=1} = \overset{*}{p}_{l|k'}, \qquad k' = 1, \ldots, K. \tag{7.14}$$

Отже, ще одне визначення незалежності можна сформулювати так: дискретна випадкова величина Y не залежить від дискретної випадкової величини X, якщо її умовний ряд розподілу за настання кожної з подій $\{X = x_k\}$ збігається з безумовним рядом розподілу. Дане визначення є аналогічне визначенню незалежності випадкових подій (див. розд. 3.16).

Можна інакше сформулювати визначення незалежності дискретної випадкової величини Y від дискретної випадкової величини X. Для цього уведемо випадкові події

$$A_k = \{X = x_k\}, \qquad B_l = \{Y = y_l\}, \qquad C_k = \{X = x_k\}\{Y = y_l\} = A_k B_l.$$

Їх дослідні ймовірності відповідно дорівнюють $\overset{*}{p}_{k}$, $\overset{*}{p}_{l}$, $\overset{*}{p}_{k,l}$ і їх можна одержати з табл. 7.7.

Будемо вважати, що дискретні випадкові величини X та Y незалежні, якщо для будь-яких $k = 1, \ldots, K$ і $l = 1, \ldots, L$ виконується така рівність:

$$P(C_{k,l}) = P(A_k)P(B_l) \tag{7.15}$$

або

$$\overset{*}{p}_{k,l} = \overset{*}{p}_k \overset{*}{p}_l . \tag{7.16}$$

З урахуванням (7.9) із рівності (7.16) випливає, що

$$\overset{*}{p}_l = \overset{*}{p}_{l|k} , \tag{7.17}$$

тобто ми одержали раніше сформульоване визначення незалежності дискретних випадкових величин X і Y.

Таким же чином можна було б розглянути умовні ряди розподілу дискретної випадкової величини X (зведені у вигляді табл. 7.8 і позначені сірим кольром), і дати визначення незалежності випадкової дискретної величини X від дискретної випадкової величини Y. Зрозуміло, що ми знову б прийшли до визначення (7.16). Отже, якщо величина Y не залежить від X, то й величина X не залежить від Y, тобто незалежність дискретних випадкових величин є взаємна.

Таблиця 7.8

	$\overset{*}{p}_{k=1}$	$\overset{*}{p}_{k=2}$	$\cdots$	$\overset{*}{p}_k$	$\cdots$	$\overset{*}{p}_{k=K}$	
$y_{l=L}$	$\overset{*}{p}_{1\|L}$	$\overset{*}{p}_{2\|L}$	$\cdots$	$\overset{*}{p}_{k\|L}$	$\cdots$	$\overset{*}{p}_{K\|L}$	$\overset{*}{p}_{k\|l=L}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_l	$\overset{*}{p}_{1\|l}$	$\overset{*}{p}_{2\|l}$	$\cdots$	$\overset{*}{p}_{k\|l}$	$\cdots$	$\overset{*}{p}_{K\|l}$	$\overset{*}{p}_{k\|l}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$y_{l=2}$	$\overset{*}{p}_{1\|l=2}$	$\overset{*}{p}_{2\|l=2}$	$\cdots$	$\overset{*}{p}_{k\|l=2}$	$\cdots$	$\overset{*}{p}_{K\|l=2}$	$\overset{*}{p}_{k\|l=2}$
$y_{l=1}$	$\overset{*}{p}_{1\|l=1}$	$\overset{*}{p}_{2\|l=1}$	$\cdots$	$\overset{*}{p}_{k\|l=1}$	$\cdots$	$\overset{*}{p}_{K\|l=1}$	$\overset{*}{p}_{k\|l=1}$
	$x_{k=1}$	$x_{k=2}$	$\cdots$	x_k	$\cdots$	$x_{k=K}$	

Із (7.16) випливає, що незалежні дискретні випадкові величини X і Y являють собою ту виняткову ситуацію, за якої з маргінальних рядів розподілу дискретних випадкових величин X і Y (табл. 7.3 і 7.4) може бути одержана таблиця спільного розподілу (табл. 7.6), тобто виконується зворотне твердження.

Зазначимо також, що наведені визначення незалежності дискретних випадкових величин справедливі не тільки для статистичних імовірностей, а й для класичних (див. розд. 3.5), тобто спосіб обчислення ймовірностей не є принциповий. Отже, у виразі (7.16) зірочку, що вказує на статистичну ймовірність, можна вилучити.

7.4. Дії над дискретними випадковими величинами

Нехай відома таблиця спільного розподілу дискретних випадкових величин X і Y.

Сумою (різницею) дискретних випадкових величин X і Y називається дискретна випадкова величина $Z=X\pm Y$, ряд розподілу якої одержимо за таким правилом:

1) утворимо прямий добуток множин значень $\{x_k,\ k=1,\ldots,K\}$ і $\{y_l,\ l=1,\ldots,L\}$, тобто множину всіляких упорядкованих пар $\{(x_k, y_l),\ k=1,\ldots K,\ l=1,\ldots,L\}$ (див. розд. 3.1);

2) для пар (x_k, y_l), що утворюють прямий добуток, обчислимо відповідні значення $x_k \pm y_l$;

3) співвіднесемо обчислені значення $x_k \pm y_l$ і їхні ймовірності $p_{k,l}$ з таблиці спільного розподілу величин X і Y;

4) зі значень $x_k \pm y_l$ і ймовірностей $p_{k,l}$ складемо ряд розподілу величини Z з урахуванням можливого збігу деяких значень $x_k \pm y_l$ і несумісності відповідних подій (тобто з використанням правила додавання ймовірностей для несумісних подій, див. розд. 3.10, 3.11).

Приклад 7.1. Спільний розподіл дискретних випадкових величин X і Y заданий табл. 7.9.

Таблиця 7.9

$y_{l=3}=1$	0,10	0,15	0,05
$y_{l=2}=0$	0,10	0,05	0,25
$y_{l=1}=-1$	0,05	0,15	0,10
Y \ X	$x_{k=1}=0$	$x_{k=2}=1$	$x_{k=3}=2$

Побудуємо ряд розподілу величини $Z=X\pm Y$ за наведеним вище правилом:

1) утворимо прямий добуток множин значень величин X і Y:

{(0, –1), (0, 0), (0, 1), (1, –1), (1, 0), (1, 1), (2, –1), (2, 0), (2, 1)};

2) обчислимо суми значень в кожній парі: {–1, 0, 1, 0, 1, 2, 1, 2, 3};

3) співіднесемо обчислені суми і їхні ймовірності з таблиці спільного розподілу: {0,05; 0,10; 0,10; 0,15; 0,05; 0,15; 0,10; 0,25; 0,05};

4) складемо попередній ряд розподілу величини Z у вигляді табл. 7.10:

Таблиця 7.10

–1	0	1	2	3
0,05	0,10	0,10	0,15	0,05
	0,15	0,05	0,25	
		0,10		

5) складемо остаточний ряд розподілу величини Z у вигляді табл. 7.11:

Таблиця 7.11

–1	0	1	2	3
0,05	0,25	0,25	0,40	0,05

Для наочності на рис. 7.1 показані ряди розподілу величин X, Y і Z (побудова рядів розподілу величин X і Y за таблицею спільного розподілу пояснюється в розд. 7.2).

Пояснимо перехід від попереднього ряду розподілу (табл. 7.10) до остаточного ряду розподілу величини Z (табл. 7.11). Наприклад, набуття значення $z = 0$ величиною Z (див. другий стовпчик у табл. 7.10, 7.11) є наслідком настання двох незалежних подій

$$\{Z = 0\} = \{X = 0\}\{Y = 0\} + \{X = -1\}\{Y = 1\}.$$

Отже, за правилом додавання ймовірностей для двох несумісних подій (див. розд. 3.10)

$$\begin{aligned} P(\{Z = 0\}) &= P(\{X = 0\}\{Y = 0\} + \{X = -1\}\{Y = 1\}) = \\ &= P(\{X = 0\}\{Y = 0\}) + P(\{X = -1\}\{Y = 1\}) = \\ &= 0{,}10 + 0{,}15 = 0{,}25. \end{aligned}$$

Так само обчислюються і ймовірності інших повторюваних значень величини Z, які дорівнюють 1 і 2. ●

Якщо дискретні випадкові величини X і Y незалежні, то у разі побудови ряду розподілу дискретної випадкової величини $Z=X\pm Y$ таблиця спільного розподілу не потрібна, оскільки ймовірності $p_{k,l}$ настання значень $x_k \pm y_l$ (див. пункт 3 правила на с. 16) можуть бути обчислені безпосередньо з рядів розподілу дискретних випадкових величин X і Y за умови застосування правила множення ймовірностей для двох незалежних подій: $p_{k,l} = p_k\, p_l$ (див. розд. 3.17).

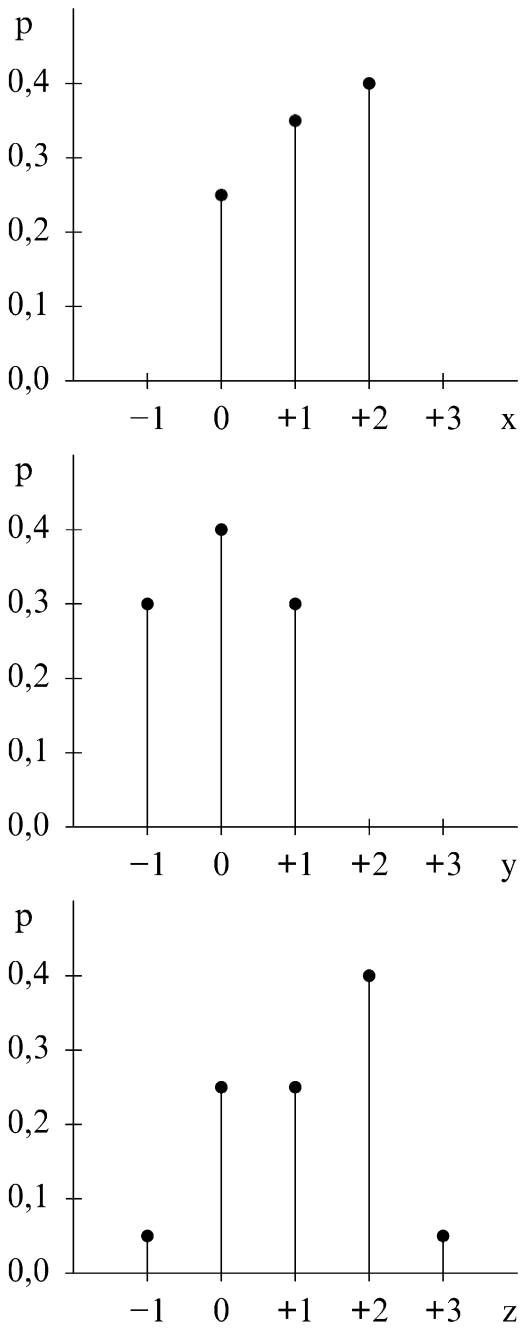


Рис. 7.1. Ряди розподілів дискретних випадкових величин X, Y (див. табл. 7.9) і Z (див. табл. 7.11)

Приклад 7.2. Незалежні дискретні випадкові величини X і Y задані рядами розподілів, зображеними в табл. 7.12 і 7.13 і на рис. 7.2.

Таблиця 7.12

$x_{k=1}=0$	$x_{k=2}=1$	$x_{k=3}=2$
0,2	0,5	0,3

Таблиця 7.13

$y_{l=1}=-1$	$y_{l=2}=0$	$y_{l=3}=1$
0,3	0,4	0,3

Побудуємо ряд розподілу дискретної випадкової величини $Z=X-Y$ за відповідним правилом (див. с. 16):

1) утворимо прямий добуток множин значень величин X і Y:

{(0, –1), (0, 0), (1, +1), (1, –1), (1, 0), (1, +1), (2, –1), (2, 0), (2, +1)};

2) обчислимо різниці значень у кожній парі: {–2, 0, –1, 2, 1, 0, 3, 2, 1};

3) співведнесемо обчислені різниці і їхні ймовірності, обчислені за допомогою рядів розподілів величин X і Y:

{0,2·0,3; 0,2·0,4; 0,2·0,3; 0,5·0,3; 0,5·0,4; 0,5·0,3; 0,3·0,3; 0,3·0,4; 0,3·0,3}=
={0,06; 0,08; 0,06; 0,15; 0,20; 0,15; 0,09; 0,12; 0,09};

4) складемо попередній ряд розподілу величини Z:

Таблиця 7.14

–2	–1	0	1	2	3
0,06	0,06	0,08	0,20	0,15	0,09
		0,15	0,09	0,12	

5) складемо остаточний ряд розподілу величини Z:

Таблиця 7.15

–2	–1	0	1	2	3
0,06	0,06	0,23	0,29	0,27	0,09

Ряд розподілу величини Z зображений також на рис. 7.2. ●

Добутком (часткою) дискретних випадкових величин X і Y називається дискретна випадкова величина $Z=X\cdot Y$, ряд розподілу якої одержується за таким правилом:

1) утворимо прямий добуток множин значень $\{x_k,\ k=1,\ldots,K\}$ і $\{y_l,\ l=1,\ldots,L\}$, тобто множину всіляких упорядкованих пар $\{(x_k, y_l),\ k=1,\ldots K,\ l=1,\ldots,L\}$ (див. розд. 3.1);

2) для пар (x_k, y_l), що утворюють прямий добуток, обчислимо відповідні значення $x_k \cdot y_l$ (x_k / y_l);

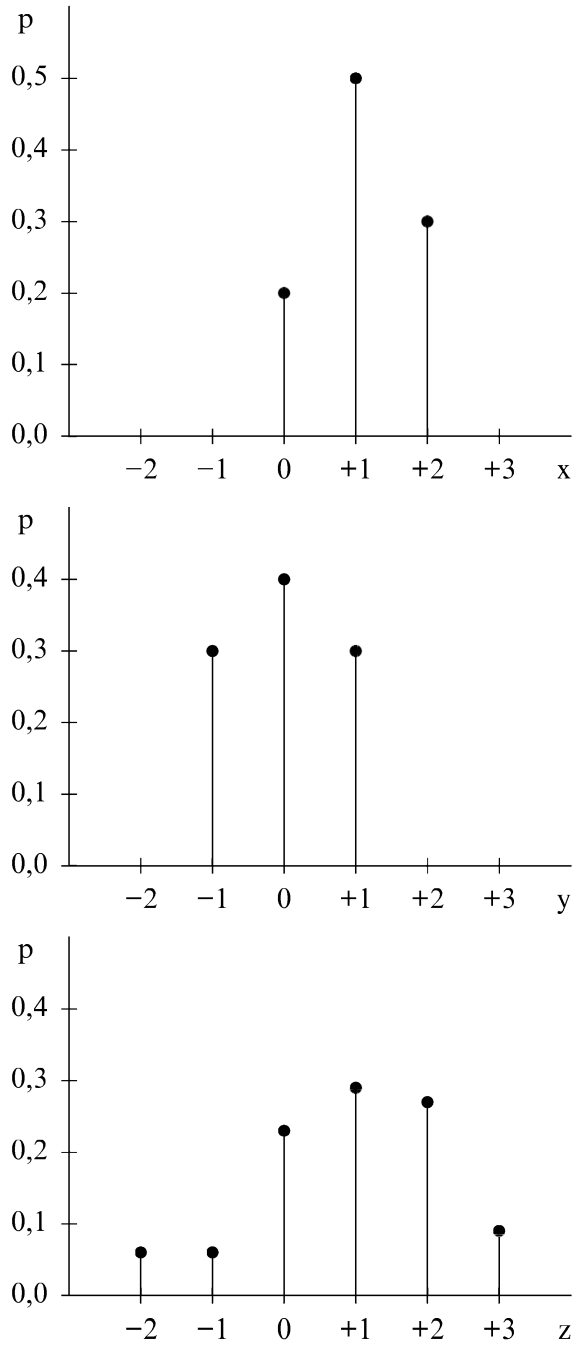


Рис. 7.2. Ряди розподілів дискретних випадкових величин X, Y (див. табл. 7.12 і 7.13) і Z (див. табл. 7.15)

3) співвіднесемо обчислені значення $x_k \cdot y_l$ (x_k / y_l) і їхні ймовірності $p_{k,l}$ з таблиці спільного розподілу випадкових величин X і Y;

4) зі значень $x_k \cdot y_l$ (x_k / y_l) і відповідних ймовірностей $p_{k,l}$ складемо ряд розподілу величини Z, з урахуванням можливого збігу деяких значень $x_k \cdot y_l$

(x_k / y_l) і несумісності відповідних подій (тобто з використанням правила додавання ймовірностей для несумісних подій, див. розд. 3.10, 3.11).

Сума (різниця) і добуток (**частка**) двох дискретних випадкових величин **узагальнюються для довільного числа дискретних випадкових величин** очевидним способом:

$$U=X+Y+Z=(X+Y)+Z\,,$$

$$V=X\cdot Y\cdot Z=(X\cdot Y)\cdot Z, \qquad \text{і т.ін.}$$

Властивості дій над дискретними випадковими величинами

1. Вибіркове середнє значень суми (різниці) дискретних випадкових величин X і Y дорівнює сумі (різниці) їх вибіркових середніх:

$$m_{X\pm Y} = m_X \pm m_Y \tag{7.18}$$

Дійсно, за визначенням вибіркового середнього дискретної випадкової величини $X\pm Y$, заданої своїм рядом розподілу відносних частот (див. розд. 4.11), і суми (різниці) $X\pm Y$ дискретних випадкових величин X і Y (див. розд. 7.4)

$$m_{X\pm Y} = \sum_{k=1}^{K}\sum_{l=1}^{L}(x_k \pm y_l)\frac{n_{k,l}}{n} = \sum_{k=1}^{K}\sum_{l=1}^{L} x_k \frac{n_{k,l}}{n} \pm \sum_{k=1}^{K}\sum_{l=1}^{L} y_l \frac{n_{k,l}}{n} =$$

$$= \sum_{k=1}^{K} x_k \left(\sum_{l=1}^{L}\frac{n_{k,l}}{n}\right) \pm \sum_{l=1}^{L} y_l \left(\sum_{k=1}^{K}\frac{n_{k,l}}{n}\right) = \sum_{k=1}^{K} x_k \frac{n_{k,l}}{n} \pm \sum_{l=1}^{L} y_l \frac{n_{k,l}}{n} = \tag{7.19}$$

$$= \frac{1}{n}\sum_{k=1}^{K} x_k n_k \pm \frac{1}{n}\sum_{l=1}^{L} y_l n_l = m_X \pm m_Y \ .$$

Зауваження 7.1. Зверніть увагу на те, що визначення незалежності у властивості 1 не використовується, отже, ця властивість справедлива для будь-яких дискретних випадкових величин X і Y. ▲

Зауваження 7.2. Властивість 1 узагальнюється для довільного числа випадкових величин. ▲

2. Математичне сподівання суми (різниці) дискретних випадкових величин X і Y дорівнює сумі (різниці) їхніх математичних сподівань:

$$M[X \pm Y] = M[X] \pm M[Y]. \tag{7.20}$$

Доведення випливає з попередньої властивості, оскільки в її формулюванні явно не використовується число дослідів n (воно справедливе для будь-якого числа n), і з визначення математичного сподівання (див. розд. 4.14 і 4.17).

3. Математичне сподівання добутку незалежних дискретних випадкових величин X і Y дорівнює добутку їхніх математичних сподівань:

$$M[XY] = M[X]M[Y]\,. \tag{7.21}$$

За визначеннями математичного сподівання дискретної випадкової величини XY, заданої своїм рядом розподілу, і добутку дискретних випадкових величин X і Y

$$M[XY] = \sum_{k=1}^{K}\sum_{l=1}^{L} x_k y_l p^*_{k,l} = \sum_{k=1}^{K}\sum_{l=1}^{L} x_k y_l p^*_k p^*_l = \sum_{k=1}^{K}\sum_{l=1}^{L} (x_k p^*_k)(y_l p^*_l) =$$

$$= \left(\sum_{k=1}^{K} x_k p^*_k\right)\left(\sum_{l=1}^{L} y_l p^*_l\right) = M[X]M[Y]. \tag{7.22}$$

Зауваження 7.3. Властивість 3 не можна сформулювати для вибіркових середніх, тому що у процесі її доведення використовується визначення незалежності дискретних випадкових величин X і Y, а останнє ґрунтується на ймовірностях, а не на відносних частотах, тобто заміна в (7.19) ймовірностей відносними частотами буде незаконна. ▲

4. Вибіркова дисперсія суми (різниці) дискретних випадкових величин X і Y дорівнює

$$s^2_{X\pm Y} = s^2_X + s^2_Y \pm 2s_{X,Y}\,, \tag{7.23}$$

де $s_{X,Y}$ – вибіркова коваріація дискретних випадкових величин X і Y, визначена нижче.

Для доведення цієї властивості введемо позначення $Z = X \pm Y$ й звернемо увагу на те, що вибіркова дисперсія величини Z, обумовлена виразом (4.7) (див. розд. 14.11), може бути обчислена як середнє значення квадратів відхилень $(z_i - m_Z)^2$ або, що те ж саме, як вибіркове середнє величини $(Z - m_Z)^2$:

$$s^2_Z = m_{(Z-m_Z)^2}\,. \tag{7.24}$$

Розгорнутий вираз для величини $(Z - m_Z)^2$ випливає безпосередньо з її визначення й властивості 1:

$$(Z - m_Z)^2 = [(X \pm Y) - m_{(X \pm Y)}]^2 =$$
$$= [(X \pm Y) - (m_X \pm m_Y)]^2 = [(X - m_X) \pm (Y - m_Y)]^2 = \qquad (7.25)$$
$$= (X - m_X)^2 \pm 2(X - m_X)(Y - m_Y) + (Y - m_Y)^2 .$$

Обчислюючи вибіркове середньє значення одержаного виразу, варто застососувати властивість 1, звідки

$$s^2_{X \pm Y} = m^2_{(X - m_X)} \pm 2\, m_{(X - m_X)(Y - m_Y)} + m^2_{(Y - m_Y)} = s^2_X + s^2_Y \pm 2\, s_{XY} \ , \qquad (7.26)$$

де $s_{X,Y}$ – вибіркова коваріація:

$$s_{X,Y} = m_{(X - m_X)(Y - m_Y)} = \frac{1}{n} \sum_{k=1}^{K} \sum_{l=1}^{L} n_{k,l}\, (x_k - m_x)(y_l - m_Y) . \qquad (7.27)$$

Зауваження 7.4. Обґрунтування властивості 4 явно не використовує число дослідів, тому воно може бути сформульовано і для дисперсії. ▲

5. Дисперсія суми (різниці) дискретних випадкових величин X і Y дорівнює

$$D[X \pm Y] = D[X] + D[Y] \pm 2K[X,Y] \ , \qquad (7.28)$$

де $K[X,Y]$ – коваріація дискретних випадкових величин X і Y, визначена нижче.

Доведення випливає з попередньої властивості, у якій вибіркові дисперсії s^2_X і s^2_Y і вибіркова коваріація s_{XY} заміняються дисперсіями $D[X]$ і $D[Y]$ і коваріацією

$$K[X,Y] = \sum_{k=1}^{K} \sum_{l=1}^{L} p^*_{k,l} (x_k - \mu_X)(y_l - \mu_Y) . \qquad (7.29)$$

Для коваріації також використовується позначення $\sigma_{X,Y}$, за допомогою якого властивість може бути записана так:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\sigma_{XY} . \qquad (7.30)$$

6. Дисперсія суми (різниці) незалежних дискретних випадкових величин X і Y дорівнює сумі їхніх дисперсій:

$$D[X \pm Y] = D[X] + D[Y] \tag{7.31}$$

або

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y . \tag{7.32}$$

Доведення ґрунтується на тому, що для незалежних дискретних випадкових величин X і Y їхня коваріація дорівнює нулю. Дійсно, якщо підставити в (7.29) умову незалежності (7.16), то

$$K[X,Y] = \sum_{k=1}^{K}\sum_{l=1}^{L} p_k^* p_l^* (x_k - \mu_X)(y_l - \mu_Y) = \sum_{k=1}^{K}\sum_{l=1}^{L} [p_k^*(x_k - \mu_X)][p_l^*(y_l - \mu_Y)] =$$

$$= \left(\sum_{k=1}^{K} p_k^*(x_k - \mu_X)\right)\left(\sum_{l=1}^{L} p_l^*(y_l - \mu_Y)\right) =$$

$$= \left(\sum_{k=1}^{K} p_k^* x_k - \sum_{k=1}^{K} p_k^* \mu_X\right)\left(\sum_{l=1}^{L} p_l^* y_l - \sum_{l=1}^{L} p_l^* \mu_Y\right) =$$

$$= \left(\mu_X - \mu_X \sum_{k=1}^{K} p_k^*\right)\left(\mu_Y - \mu_Y \sum_{l=1}^{L} p_l^*\right) = (\mu_X - \mu_X)(\mu_Y - \mu_Y) = 0 . \tag{7.30}$$

Зауваження 7.5. Властивість 6 не можна сформулювати для вибіркових дисперсій, тому що в умові незалежності (7.16) використовуються ймовірності, а не відносні частоти (див. зауваження 7.3). ▲

Підіб'ємо підсумки, зібравши одержані властивості попарно – для вибіркового середнього й вибіркової дисперсії у табл 7.16 і для математичного сподівання і дисперсії у табл. 7.17[2].

Таблиця 7.16

Властивість вибіркового середнього	Властивість вибіркової дисперсії
1. Для довільних дискретних випадкових величин X і Y $m_{X \pm Y} = m_X \pm m_Y$.	4. Для довільних дискретних випадкових величин X і Y $s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2s_{XY}$.

[2] Для зручності роботи властивості у табл. 7.16, 7.17 наведені із їхніми порядковими номерами (див. с. 21 – 24).

Таблиця 7.17

Властивості математичного сподівання	Властивості дисперсії
2. Для довільних дискретних випадкових величин X і Y $$M[X \pm Y] = M[X] \pm M[Y],$$ $$\mu_{X \pm Y} = \mu_X \pm \mu_Y.$$	5. Для довільних дискретних випадкових величин X і Y $$D[X \pm Y] = D[X] + D[Y] \pm 2K[X, Y],$$ $$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\sigma_{XY}.$$
3. Для незалежних дискретних випадкових величин X і Y $$M[XY] = M[X]M[Y],$$ $$\mu_{XY} = \mu_X \mu_Y.$$	6. Для незалежних дискретних випадкових величин X і Y $$D[X \pm Y] = D[X] + D[Y],$$ $$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y.$$

7.5. Спільний розподіл двох інтервальних і неперервних випадкових величин

Розглянемо дослід, у якому вимірюються значення неперервних випадкових величин X і Y. Ці значення у вихідному вигляді являють собою записи в лабораторному журналі (див. табл. 7.1) або n пар значень (7.1). Так само, як за аналізу вибіркових значень однієї неперервної випадкової величини (за однією кількісною ознакою) не було сенсу обробляти їх у вигляді рядів частот (див. розд. 4.8), так саме немає сенсу обробляти вибіркові значення двох неперервних випадкових величинах у вигляді таблиць частот (див. розд. 7.2). Тому будемо діяти таким чином. За алгоритмом, описаним у розд. 4.8, уведемо для неперервної випадкової величини X систему K інтервалів $[a_k, a_{k+1})$, $k = 1, \ldots, K$, а для неперервної випадкової величини Y – систему L інтервалів $[b_l, b_{l+1})$, $l = 1, \ldots, L$. Прямий добуток цих інтервалів

$$[a_k, a_{k+1}) \times [b_l, b_{l+1}) \tag{7.34}$$

визначає на площині (x, y) $K \cdot L$ прямокутників (клітинок), ліва й нижня ділянки межі яких належать їм, а права й верхня ділянки межі – ні. Надалі будемо позначати клітинки парою індексів k і l – (k, l).

Якщо кожну пару вибіркових значень (x_i, y_i), $i = 1, \ldots, n$ розглядати як декартові координати точки на площині (x, y), то всі вибіркові значення можна зобразити у вигляді n точок. Одержане зображення множини n точок, що відповідають вибір-

ковим значенням (x_i, y_i), $i = 1, \ldots, n$, називається **кореляційним полем** або **діаграмою розсіювання** (рис. 7.3).

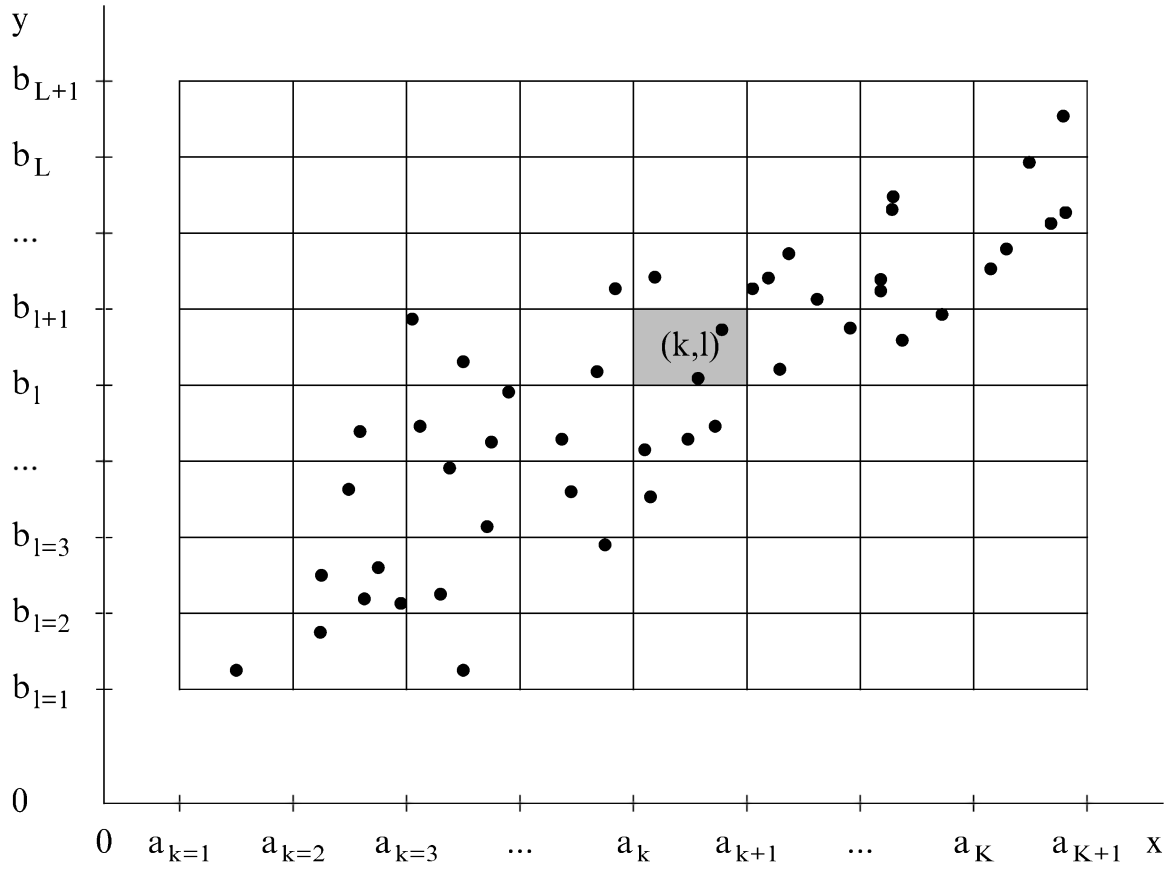


Рис. 7.3. Кореляційне поле для вибіркових даних, наведених в лабораторному журналі (див. табл. 7.1 або (7.1))

Зауваження 7.6. Зображення кореляційного поля не потребує нанесення меж клітинок (k, l). На рис. 7.3 межі показані з міркувань зручності подальшої обробки вибіркових значень. ▲

Обчислимо за кореляційним полем кількість $n_{k,l}$ (частоту) вибіркових значень у кожній клітинці (k, l) і складемо двовимірну таблицю (табл. 7.18), називану **інтервальною таблицею спільного розподілу частот** неперервних випадкових величин X і Y.

У першому знизу рядку й першому ліворуч стовпчику таблиці позначені номери k і l клітинок (k, l).

Як і для дискретних випадкових величин X і Y, для неперервних випадкових величин X і Y можна підсумовуванням частот $n_{k,l}$ за стовпчиками і рядками одержати частоти n_k й n_l, зображені в першому зверху рядку й першому праворуч стовпчику табл. 7.18.

Таблиця 7.18

	$n_{k=1}$	$n_{k=2}$	…	n_k	…	$n_{k=K}$	
l=L	$n_{1,L}$	$n_{2,L}$	…	$n_{k,L}$	…	$n_{K,L}$	$n_{l=L}$
…	…	…	…	…	…	…	…
L	$n_{1,l}$	$n_{2,l}$	…	$n_{k,l}$	…	$n_{K,l}$	n_l
…	…	…	…	…	…	…	…
l=2	$n_{1,2}$	$n_{2,2}$	…	$n_{k,2}$	…	$n_{K,2}$	$n_{l=2}$
l=1	$n_{1,1}$	$n_{2,1}$	…	$n_{k,1}$	…	$n_{K,1}$	$n_{l=1}$
	k=1	k=2	…	k	…	k=K	

Нижній і верхній рядки та лівий і правий стовчики інтервальної таблиці спільного розподілу частот утворюють інтервальні ряди розподілу частот неперервних випадкових величин X і Y, аналогічні розглянутим у розд. 4.8.

Так само, як і за розгляду таблиць спільного розподілу частот дискретних випадкових величин X і Y у розд. 7.2, можна довести, що з інтервальних рядів розподілу неперервних випадкових величин X і Y інтервальну таблицю розподілу частот одержати не можна.

У такий же спосіб, як було зроблено за обробки вибіркових значень дискретних випадкових величин X і Y, можна ввести інтервальні таблиці відносних частот $n_{k,l}/n$ і статистичних ймовірностей $\overset{*}{p}_{k,l}$ (умова стабілізації відносних частот вважається виконана).

Інтервальна таблиця ймовірностей спільного розподілу випадкових величин X і Y наведена в табл. 7.19.

Відносно інтервальних таблиць спільного розподілу частот і відносних частот (статистичних ймовірностей) за зміни обсягу вибірки n і ширини інтервалів Δa і Δb справедливі ті ж твердження, що були раніше сформульовані в розд. 4.9 для інтервальних рядів однієї випадкової величини. Тому вводиться інваріантна щодо зміни обсягу вибірки n і ширини інтервалів Δa і Δb величина

$$f^*_{k,l} = \frac{\overset{*}{p}_{k,l}}{\Delta a\,\Delta b}, \quad k = 1,\ldots,K, \quad l = 1,\cdots,L, \tag{7.35}$$

що називається **вибірковою функцією густиниі спільного розподілу неперевних випадкових величин** X і Y. Зі значень $f^*_{k,l}$ для всіх клітинок (k,l) одержимо інтервальну таблицю вибіркової функції густини спільного розподілу X і Y, наведену в табл. 7.19.

Таблиця 7.19

	$p^*_{k=1}$	$p^*_{k=2}$	…	p^*_{k}	…	$p^*_{k=K}$	
$l=L$	$p^*_{1,L}$	$p^*_{2,L}$	…	$p^*_{k,L}$	…	$p^*_{K,L}$	$n_{l=L}$
…	…	…	…	…	…	…	…
l	$p^*_{1,l}$	$p^*_{2,l}$	…	$p^*_{k,l}$	…	$p^*_{K,L}$	n_l
…	…	…	…	…	…	…	…
$l=2$	$p^*_{1,2}$	$p^*_{2,2}$	…	$p^*_{k,2}$	…	$p^*_{K,2}$	$n_{l=2}$
$l=1$	$p^*_{1,1}$	$p^*_{2,1}$	…	$p^*_{k,1}$	…	$p^*_{K,1}$	$n_{l=1}$
	$k=1$	$k=2$	…	k	…	$k=K$	

Таблиця 7.20

	$f^*_{k=1}$	$f^*_{k=2}$	…	f^*_{k}	…	$f^*_{k=K}$	
$l=L$	$f^*_{1,L}$	$f^*_{2,L}$	…	$f^*_{k,L}$	…	$f^*_{K,L}$	$n_{l=L}$
…	…	…	…	…	…	…	…
l	$f^*_{1,l}$	$f^*_{2,l}$	…	$f^*_{k,l}$	…	$f^*_{K,L}$	n_l
…	…	…	…	…	…	…	…
$l=2$	$f^*_{1,2}$	$f^*_{2,2}$	…	$f^*_{k,2}$	…	$f^*_{K,2}$	$n_{l=2}$
$l=1$	$f^*_{1,1}$	$f^*_{2,1}$	…	$f^*_{k,1}$	…	$f^*_{K,1}$	$n_{l=1}$
	$k=1$	$k=2$	…	k	…	$k=K$	

Гістограма вибіркової функції густини зображена на рис. 7.3. **Функція густини** $f(x, y)$ спільного розподілу неперервних випадкових величин X і Y є результатом граничного переходу $n \to N$, $\Delta a \to 0$ $(K \to \infty)$, $\Delta b \to 0$ $(L \to \infty)$ для вибіркової функції густини $f^*_{k,l}$. Графіком функції густини $f(x, y)$ є відповідна поверхня.

Тепер перейдемо до формулювання означень незалежності неперервних випадкових величин X і Y. Будемо вважати X і Y для зручності інтервальними, тобто такими, всі дані про які наводяться для клітинок (k, l).

Інтервальна випадкова величина *Y* не залежить від інтервальної випадкової величини *X*, якщо умовні інтервальні ряди розподілу випадкової величини Y за настання подій $\{a_k \le X < a_{k+1}\}$, $k = 1, \ldots, K$ є майже однакові.

Інтервальна випадкова величина *Y* не залежить від інтервальної випадкової величини *X*, якщо умовний інтервальний ряд випадкової величини Y за настання будь-якої події $\{a_k \le X < a_{k+1}\}$, $k = 1, \ldots, K$ майже збігається з інтервальним рядом розподілу випадкової величини Y.

Якщо ввести допоміжні випадкові події $A_k = \{a_k \le X < a_{k+1}\}$, $B_l = \{b_l \le Y < b_{l+1}\}$, $C_{k,l} = A_k B_l$ з відповідними ймовірностями p^*_k, p^*_l і $p^*_{k,l}$, то умови незалежності інтервальних випадкових величин X і Y в наведених двох означеннях можуть бути сформульовані однаково як виконання (звичайно, приблизне) такої рівності:

$$p^*_{k,l} = p^*_k \, p^*_l . \tag{7.36}$$

Звідси легко довести, що якщо інтервальна випадкова величина Y не залежить від інтервальної випадкової величини X, то й інтервальна випадкова величина X не залежить від інтервальної випадкової величини Y, тобто незалежність інтервальних випадкових величин є взаємна (див. розд. 7.3).

Якщо будь-яке з наведених вище рівносильних означень незалежності інтервальних випадкових величин X і Y справедливе за будь-якого достатньо великого числа інтервалів K і L або, що те ж саме, за будь-якої достатньо малої ширини інтервалів Δa і Δb

$$\Delta a = a_{k=2} - a_{k=1} = a_{k=3} - a_{k=2} = \ldots = a_{K+1} - a_K \ , \tag{7.37}$$

$$\Delta b = b_{l=2} - b_{l=1} = b_{l=3} - b_{l=2} = \ldots = b_{L+1} - b_L \ , \tag{7.38}$$

то маємо підставу говорити, що **інтервальні випадкові величини** X і Y **незалежні**.

7.6. Дії над інтервальними та неперервними випадковими величинами

Розглянуті в розд. 7.4 дії над дискретними випадковими величинами, заданими таблицею спільного розподілу або рядами розподілу (для незалежних величин), з незначними змінами переносяться на інтервальні величини.

Нехай відома таблиця спільного розподілу інтервальних випадкових величин X і Y.

Сумою (різницею) інтервальних випадкових величин X і Y називається інтервальна випадкова величина $Z=X\pm Y$, ряд розподілу якої одержуємо за таким правилом:

1) утворюємо прямий добуток інтервалів $[a_k, a_{k+1})\times[b_l, b_{l+1})$, $k=1,\ldots,K$, $l=1,\ldots,L$, тобто множину всіляких упорядкованих пар інтервалів $([a_k, a_{k+1})\times$ $\times[b_l, b_{l+1}))$ (див. розд. 1.3);

2) для пар інтервалів, що утворюють прямий добуток, обчислюємо пари значень $a_k \pm b_l$, $a_{k+1} \pm b_{l+1}$, що складають інтервал значень величини Z;

3) співвіднесемо обчислені інтервали значень $[a_k \pm b_l, a_{k+1} \pm b_{l+1})$ величини Z і їхні ймовірності $p^*_{k,l}$ з таблиці спільного розподілу інтервальних випадкових величин X і Y;

4) з інтервалів значень $[a_k \pm b_l, a_{k+1} \pm b_{l+1})$ і ймовірностей $p^*_{k,l}$ складемо інтервальний ряд величини $Z=X\pm Y$ з урахуванням можливого перекриття інтервалів $[a_k \pm b_l, a_{k+1} \pm b_{l+1})$.

Приклад 7.3. Інтервальні випадкові величини X і Y задані інтервальною таблицею спільного розподілу, зображеною в табл. 7.21.

Таблиця 7.21

Y \ X		k=1	k=2	k=3
		[0, 1)	[1, 2)	[2, 3)
l=2	[0, 1)	0,10	0,25	0,10
l=1	[1, 0)	0,20	0,15	0,20

Побудуємо інтервальний ряд розподілу інтервальної випадкової величини $Z=X+Y$ за наведеним вище правилом:

1) утворимо множину всіляких упорядкованих пар інтервалів
{ [0, 1)([–1, 0); [0, 1)([0, 1); [1, 2)([-1, 0); [1, 2)([0, 1); [2, 3)([-1, 0); [2, 3)([0, 1)};

2) утворимо множину інтервалів значень величини Z
{[–1, 1), [0, 2), [0, 2), [1, 3), [1, 3), [2, 4)};

3) співвіднесемо обчислені інтервали і їхні ймовірності
{0,20; 0,10; 0,15; 0,25; 0,20; 0,10};

4) позначимо на осі Z межі інтервалів і відповідні їм ймовірності (рис. 7.4).

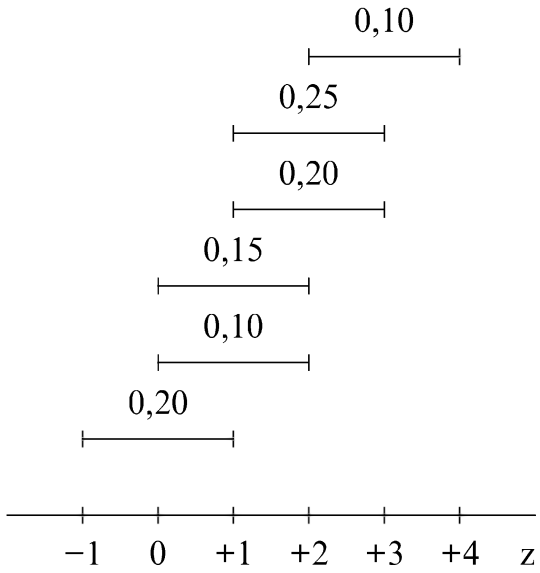


Рис. 7.4. Межі інтервалів значень з перекриттям та іхні ймовірності для величини *Z* (див. прикл. 7.3)

З урахуванням перекриття інтервалів і припущення про рівномірний розподіл усередині інтервалів уведемо нові інтервали й складемо інтервальний ряд величини *Z*, зображений у табл. 7.22.

Таблиця 7.22

[–1, 0)	[0, 1)	[1, 2)	[2, 3)	[3, 4)
0,100	0,100	0,050	0,100	0,050
	0,050	0,075	0,005	
	0,075	0,100	0,125	
		0,125		

Остаточний вигляд інтервального ряду зображений у табл. 7.23.

Таблиця 7.23

[–1, 0)	[0, 1)	[1, 2)	[2, 3)	[3, 4)
0,100	0,225	0,350	0,275	0,050

Гістограми інтервальних рядів розподілу інтервальних випадкових величин *X*, *Y* і *Z* зображені на рис. 7.5.

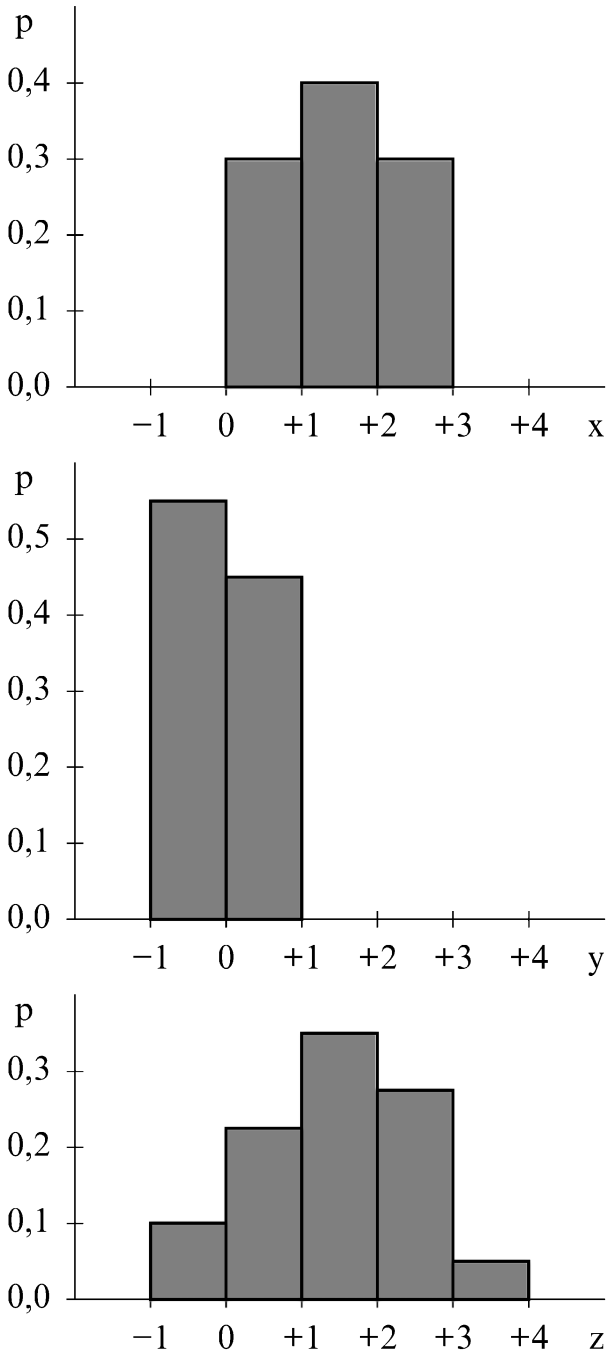


Рис. 7.5. Гістограми інтервальних випадкових величин *X*, *Y*, *Z* (див. прикл. 7.3)

Відзначимо, що інтервальна величина-результат може бути одержана не єдиним способом. Якщо укрупнити інтервали, то можливий інтервальний ряд для величини *Z* наведений в табл. 7.24, 7.25:

Таблиця 7.24

[–1, 1)	[1, 2)	[2, 4)
0,200	0,050	0,100
0,050	0,075	0,125
0,075	0,100	0,050
	0,125	

Таблиця 7.25

[–1, 1)	[1, 2)	[2, 4)
0,325	0,450	0,275

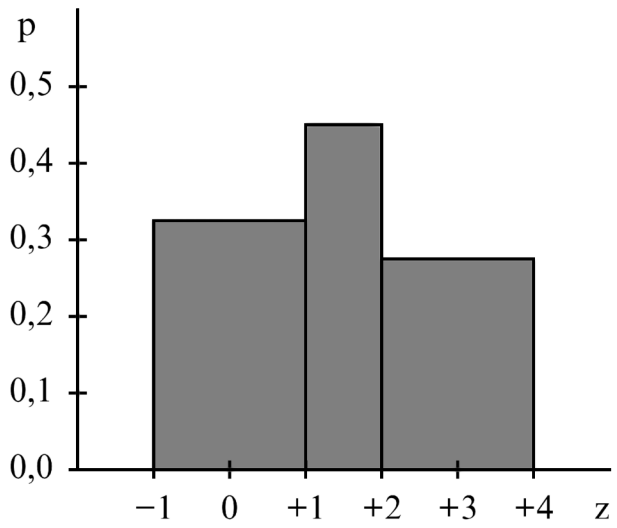


Рис. 7.6. Гістограма величини Z після укрупнення інтервалів (див. прикл. 7.3 та рис. 7.5)

Гістограма інтервального ряду розподілу інтервальної випадкової величини Z з табл. 7.25 показана на рис. 7.6. ●

Для того щоб виконати дії над неперервними випадковими величинами X і Y, потрібно за функцією густини їхнього спільного розподілу $f(x, y)$ утворити інтервальний спільний розподіл, для якого виконати необхідні дії з інтервальними величинами X і Y. Для одержаної інтервальної величини потім потрібно виконати граничний перехід $\Delta a \to 0$, $\Delta b \to 0$, у результаті якого одержимо неперервну випадкову величину.

Перейдемо до розгляду властивостей дій над інтервальними й неперервними випадковими величинами.

Розглянуті в розд. 7.5. властивості числових характеристик суми (різниці) і добутку дискретних випадкових величин X і Y можна перенести на випадок інтервальних і неперервних випадкових величин X і Y, однак для початку слід уточнити поняття вибіркових числових характеристик для інтервальних і неперервних випадкових величин, аналогічно тому, як це зроблено в розд. 4.16 і 7.4.

Для інтервальних випадкових величин X і Y під вибірковим середнім значенням суми (різниці) будемо розуміти величину

$$m_{X\pm Y} \approx \frac{1}{n}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{k,l}\,(x_k \pm y_l)\,, \tag{7.39}$$

під вибірковим середнім значенням добутку –

$$m_{XY} \approx \frac{1}{n}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{k,l}\,x_k\,y_l\,, \tag{7.40}$$

під вибірковою коваріацією –

$$s_{X,Y} \approx \frac{1}{n}\sum_{k=1}^{K}\sum_{l=1}^{L} n_{k,l}\,(x_k - m_X)(y_l - m_Y)\,. \tag{7.41}$$

Відтак відповідні властивості 1 і 4 з розд. 7.4 (див. табл. 7.16) для вибіркових числових характеристик інтервальних випадкових величин X і Y будуть виражати тільки наближені рівності (подробиці доведення можна відновити, звернувшись ще раз до розд. 4.7 і 7.4), як представлено в табл. 7.26.

Таблиця 7.26

Властивість вибіркового середнього	Властивість вибіркової дисперсії
1. Для довільних інтервальних випадкових величин X і Y $$m_{X\pm Y} \approx m_X \pm m_Y\,.$$	4. Для довільних інтервальних випадкових величин X і Y $$s^2_{X\pm Y} \approx s^2_X + s^2_Y \pm 2\,s_{X,Y}\,.$$

Для неперервних випадкових величин X і Y вирази для $M[X \pm Y]$, $M[X,Y]$ і $K[X,Y]$ випливають із (7.39) – (7.41) граничним переходом за умови $n \to N(\infty)$, $\Delta a \to 0$ $(K \to \infty)$, $\Delta b \to 0$ $(L \to \infty)$:

$$M[X,Y]=\iint\limits_D f(x,y)(x\pm y)\,dx\,dy, \tag{7.42}$$

$$M[XY]=\iint\limits_D f(x,y)\,xy\,dx\,dy, \tag{7.43}$$

$$K[X,Y]=\iint\limits_D f(x,y)(x-\mu_X)(y-\mu_Y)\,dx\,dy. \tag{7.44}$$

Для неперервних випадкових величин X і Y надалі будем використовувати такий вираз для вибіркової коваріації:

$$s_{X,Y}=\frac{1}{n}\sum_{i=1}^{n}(x_i-m_X)(y_i-m_Y)\ , \tag{7.45}$$

від якого легко перейти до (7.44).

Подробиці одержання виразів (7.42) – (7.45) легко відновити, якщо ще раз звернутися до розд. 4.16. Для числових характеристик суми (різниці) і добутку неперервних випадкових величин X і Y властивості 2, 3, 5 і 6 для дискретних випадкових величин X і Y (див. розд. 7.4, табл. 7.17) переносяться без зміни:

Таблиця 7.27

Властивості математичного сподівання	Властивості дисперсії
2. Для довільних неперервних випадкових величин X і Y $M[X\pm Y]=M[X]\pm M[Y]$, $\mu_{X\pm Y}=\mu_X\pm\mu_Y$.	5. Для довільних неперервних випадкових величин X і Y $D[X\pm Y]=D[X]+D[Y]\pm 2K[X,Y]$, $\sigma^2_{X\pm Y}=\sigma^2_X+\sigma^2_Y\pm 2\sigma_{X,Y}$.
3. Для незалежних неперервних випадкових величин X і Y $M[XY]=M[X]M[Y]$, $\mu_{XY}=\mu_X\mu_Y$.	6. Для незалежних неперервних випадкових величин X і Y $D[X\pm Y]=D[X]+D[Y]$, $\sigma^2_{X\pm Y}=\sigma^2_X+\sigma^2_Y$.

7.7. Регресія

Розглянемо кореляційне поле, побудоване за n вибірковими значеннями (x_i, y_i) неперервних випадкових величин X і Y, з показаними на ньому клітинками (k, l), $k = 1,\ldots,K$, $l = 1,\ldots,L$, (див. рис. 7.2). Так само, як і у випадку введення умовних рядів частот, відносних частот і умовного ряду розподілу неперервної випадкової величини Y за настання події $A_k = \{X = x_k\}$, будемо розглядати тільки ті вибіркові значення (x_i, y_i), які розташовані у вертикальних смугах

$$\{a_k \le x_i < a_{k+1}\}, \quad \{b_{l=1} \le y_i < b_{l=L+1}\}, \quad k = 1,\ldots,K . \tag{7.46}$$

Число таких вибіркових значень, як це випливає з табл. 7.2, дорівнює інтервальним частотам n_k. Обчислимо за цими значеннями умовні вибіркові середні значення Y:

$$m_{Y|\{a_k \le X < a_{k+1}\}} = m_{Y|k} = \frac{1}{n_k} \sum_{i:\, a_k \le X < a_{k+1}}^{n} y_i \ , \quad k = 1,\ldots,K . \tag{7.47}$$

У виразі (7.44) враховано, що осереднюються тільки ті n_k з n значень y_i, які входять до вибіркових пар значеннь (x_i, y_i) з відповідних смуг k (7.43). Визначимо на кореляційному полі K точок з координатами

$$(x_k^*, m_{Y|k}), \qquad x_k^* = \frac{1}{2}(a_k + a_{k+1}) \ , \tag{7.48}$$

де перша координата задає положення середини смуги k уздовж осі x, а друга – положення умовного вибіркового середнього значення Y в смузі k уздовж осі y. З'єднуючи ці точки відрізками прямих, одержимо суцільну ламану (рис. 7.7).

Ламана, що з'єднує умовні середні значення $m_{Y|k}$ в смугах k, називається **лінією вибіркової регресії випадкової величини *Y* по випадковій величині *X* (регресії інтервальної випадкової величини *Y* по інтервальній випадковій величині *X*)**.

Аналогічно можна розглянути горизонтальні смуги

$$\{a_{k=1} \le x_i < a_{k=K+1}\}, \quad \{b_l \le y_i < b_{l+1}\}, \quad l = 1,\ldots,L , \tag{7.49}$$

у кожній з яких обчислити умовні вибіркові середні значення інтервальної випадкової величини X за настання відповідних подій $B_l = \{b_l \le Y < b_{l+1}\}$:

$$m_{X|\{b_l \le Y < b_{l+1}\}} = m_{X|l} = \frac{1}{n_k} \sum_{i:\, b_l \le Y < b_{l+1}}^{n} x_i \ , \quad l = 1, \ldots, L \, . \tag{7.50}$$

Позначимо на кореляційному полі L точок з координатами

$$(m_{X|l}, y_l^*), \qquad y_l^* = \frac{1}{2}(b_l + b_{l+1}) \tag{7.51}$$

і, з'єднавши їх відрізками прямих, одержимо пунктирну ламану (рис. 7.7).

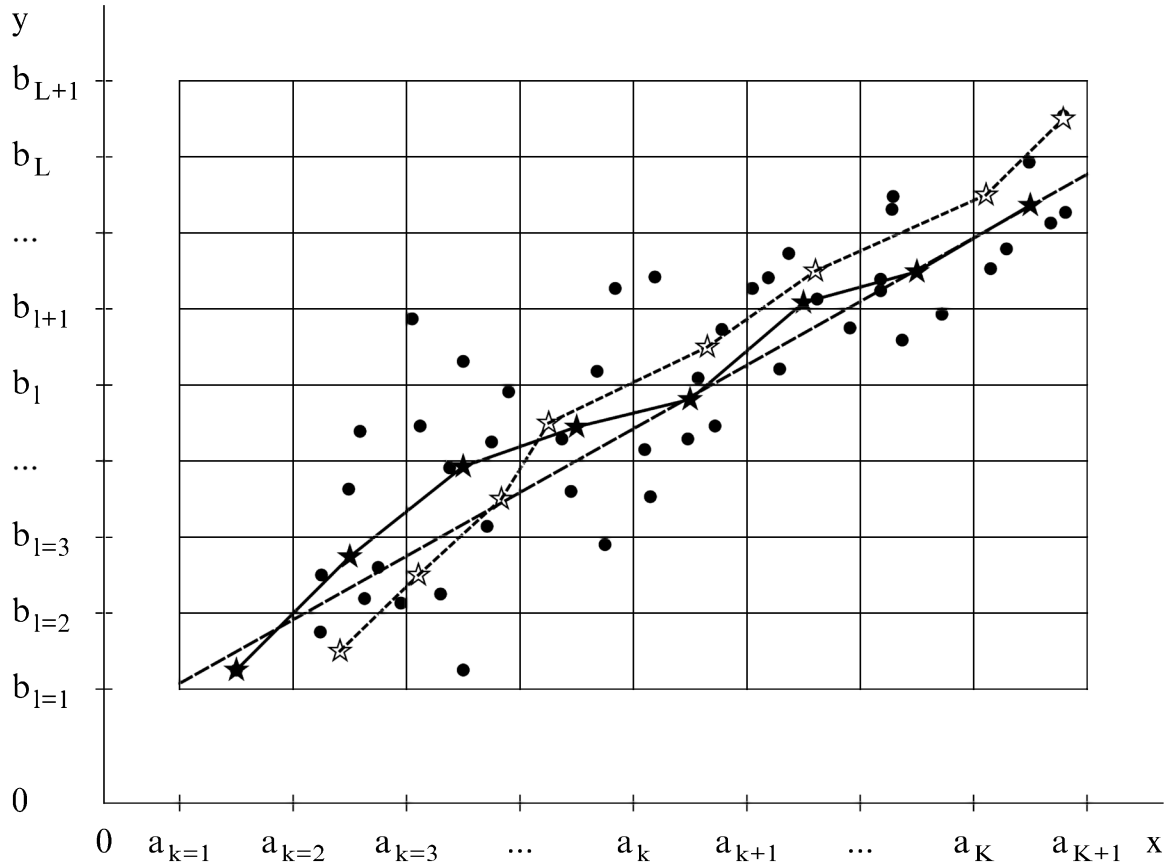


Рис. 7.7. Лінії регресії на кореляційному полі

Ламана, що з'єднує умовні середні значення $m_{X|l}$ в смугах l, називається **(вибірковою) інтервальною регресією випадкової величини *X* по випадковій величині *Y* (вибірковою регресією інтервальної випадкової величини *X* по інтервальній випадковій величині *Y*).**

За допомогою граничних переходів $\Delta a \to 0$, $\Delta b \to 0$, $n \to N(\infty)$ можна з вибіркових ліній регресії одержати (генеральні) **лінії регресії**. Розглянемо доклад-

но організацію граничних переходів на прикладі регресії Y по X. Спочатку зафіксуємо число вертикальних смуг K шириною Δa й будемо збільшувати обсяг вибірки n. Після досягнення стабілізації відносних частот у кожній зі смуг k замість умовних вибіркових середніх значень (7.44) будемо обчислювати умовні математичні сподівання

$$M\,[Y \mid \{a_k \le X < a_{k+1}\}] = \mu_{Y|\{a_k \le X < a_{k+1}\}}\,. \tag{7.52}$$

Їх можна позначити на кореляційному полі в кожній смузі у вигляді K точок $(x_k^*, \mu_{Y|\{a_k \le X < a_{k+1}\}})$ і з'єднати ламаною. Відтак перейдемо до границі за $\Delta a \to 0$ $(K \to \infty)$, продовжуючи збільшувати обсяг вибірки n. У результаті матимемо гладку лінію регресії неперервної випадкової величини Y по неперервній випадковій величині X:

$$M\,[Y \mid \{X = x\}] = \mu_{Y|\{X=x\}}\,, \tag{7.53}$$

зображену на рис. 7.7 суцільною лінією.

Так само можна одержати лінію регресії неперервної випадкової величини X по неперервній випадковій величині Y:

$$M\,[X \mid \{Y = y\}] = \mu_{X|\{Y=y\}}\,, \tag{7.54}$$

зображену на тому ж рис. 7.7 пунктирною лінією.

Визначені вище вибіркові й генеральні лінії регресії Y по X і X по Y будемо називати **лініями регресії-1**. Лінії регресії не дають повного опису системи неперервних випадкових величин X і Y, вони тільки показують, яке середнє значення набуває одна величина, якщо друга бере або значення з певного інтервалу (для вибіркової регресії), або цілком певне значення (для генеральної регресії). Відхилення (розкид, розсіювання) значень однієї випадкової величини щодо її середнього значення (вибіркового або генерального) характеризується умовною дисперсією. Для вибіркової регресії Y по X умовна вибіркова дисперсія неперервної випадкової величини Y має вигляд

$$s^2_{Y|\{a_k \le X < a_{k+1}\}} = s^2_{Y|k} = \frac{1}{n_k} \sum_{i:\, a_k \le X < a_{k+1}}^{n} (y_i - m_{Y|k})^2\,, \tag{7.55}$$

а для вибіркової регресії X по Y умовна вибіркова дисперсія безперервної випадкової величини X –

$$s^2_{X|\{b_l \le Y < b_{l+1}\}} = s^2_{X|l} = \frac{1}{n_l} \sum_{i:\, b_l \le Y < b_{l+1}}^{n} (x_i - m_{X|l})^2 . \qquad (7.56)$$

За описаних вище граничних переходів можна одержати (генеральні) умовні дисперсії

$$D[Y \mid \{X = x\}] = \sigma^2_{Y|\{X=x\}}, \qquad (7.57)$$

$$D[X \mid \{Y = y\}] = \sigma^2_{X|\{Y=y\}}. \qquad (7.58)$$

Зауваження 7.7. Термін “регресія” зустрічається також у біології й означає “спрощення”. Стосовно аналізу системи неперервних випадкових величин X і Y спрощення, що досягається за допомогою вибіркової регресії, полягає в тому, що замість n пар вибіркових значень, що здебільшого важко піддаються осмисленню, залишається K пар значень координат ламаної для регресії Y по X або L пар значень координат ламаної для регресії X по Y. У випадку (генеральної) регресії спрощення полягає в тому, що замість функції двох змінних – функції густини $f(x, y)$ – залишається одна функція $M[Y \mid \{X = x\}] = \varphi(x)$ для регресії Y по X або одна функція $M[X \mid \{Y = y\}] = \psi(y)$ для регресії X по Y. Тобто перехід до регресії супроводжується спрощенням як на рівні повноти опису зв’язку між двома кількісними ознаками X і Y в генеральній або вибірковій сукупності, так і на рівні збереженої для цього інформації. ▲

Подальше спрощення опису вибіркової або генеральної сукупності досягається вже на рівні регресії – переходом від регресії-1 до регресії-2. У регресії-2 замість ламаної лінії, що з'єднує значення умовних вибіркових середніх у смугах, підбирають гладкі лінії, що описуються простими функціями і залежать від декількох параметрів. Ці лінії не обов'язково проходять через всі точки ламаної, але за рахунок вибору функцій і значень їхніх параметрів лінії можуть досить близько проходити біля точок. Ще одне спрощення полягає в тому, що фактично підбір функцій і значень їхніх параметрів відбувається не після обчислення умовних вибіркових середніх значень у смугах, а безпосередньо за вихідними вибірковими значеннями (x_i, y_i).

Розглянемо цей алгоритм для вибіркової регресії Y по X.

Нехай вибіркові дані зображені у вигляді кореляційного поля на рис. 7.8. Вигляд кореляційного поля підказує, що прийнятним наближенням для вибіркової регресії Y по X буде пряма

$$\overline{y} = b_0 + b_1 x , \qquad (7.59)$$

рівняння якої має два невідомі параметри b_0 й b_1. Замість умовного вибіркового середнього значення в лівій частині записана змінна y (значення неперервної випадкової величини Y) з рискою зверху, що вказує за традицією на середнє значення. Тобто про опис залежності умовного вибіркового середнього значення неперервної випадкової величини Y від значень неперервної випадкової величини X, як у регресії-1, мова вже не йде, але ідея опису в середньому неперервної випадкової величини Y зберігається.

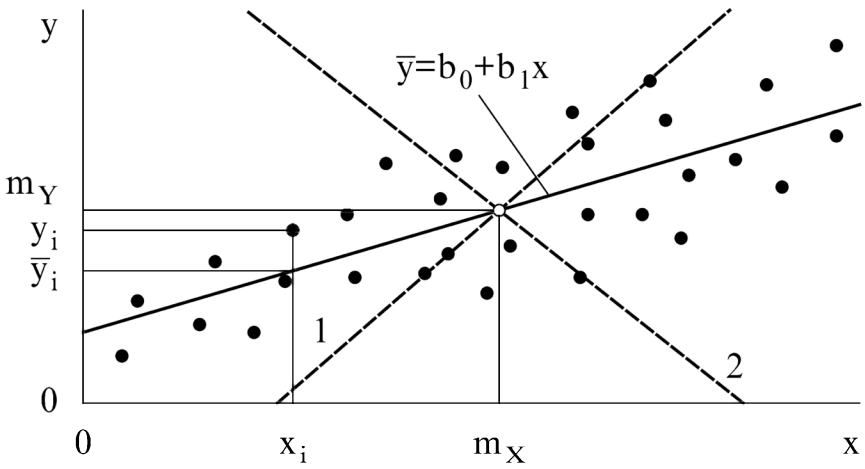


Рис. 7.8. Лінійні вибіркові регресії – „справжня“ (суцільна) та визначені за „простим правилом“ (пунктирні 1, 2)

Перейдемо до опису алгоритму обчислення параметрів регресії-2 за обраного виду функції. Досить відоме „просте правило“ побудови лінії (7.59) полягає в тому, що лінійку потрібно прикласти так, щоб по обидва боки проведеної лінії розташувалось приблизно однакове число точок. Легко переконатися, що це „просте правило“ не визначає однозначно положення лінії регресії (не визначає однозначно значень параметрів b_0 і b_1). На рис. 7.8 пунктиром проведені лінії, які задовольняють дане правило, але лініями регресіїї не є.

Перейдемо до формулювання строгого критерію для визначення значень коефіцієнтів b_0 і b_1. Підставляючи вибіркові значення x_i у рівняння лінійної регресії (7.59), ми одержимо значення $\bar{y}_i$. Різницю між вибірковими значеннями y_i та обчисленими значеннями $\bar{y}_i$ будемо називати залишками Δy_i (рис. 7.8):

$$\Delta y_i = y_i - \bar{y}_i. \tag{7.60}$$

Можна спробувати провести пряму (7.59) або, що те ж саме, визначити значення b_0 й b_1 так, щоб сума залишків Δy_i була мінімальною. Для цього замість значень $\bar{y}_i$ підставимо їхні вирази з (7.59) в (7.60) і обчислимо суму за всіма вибірковими значеннями:

$$H(b_0, b_1) = \sum_{i=1}^{n} \Delta y_i = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]. \qquad (7.61)$$

Сума (7.61) позначена як функція $H(b_0, b_1)$ двох змінних b_0 і b_1, оскільки у разі зміни значень b_0 і b_1 змінюється і значення суми (7.61). Спробуємо знайти мінімум функції двох змінних $H(b_0, b_1)$ за допомогою необхідної умови экстремуму

$$\begin{cases} \dfrac{\partial H}{\partial b_0} = 0, \\ \dfrac{\partial H}{\partial b_1} = 0. \end{cases} \qquad (7.62)$$

Значення частинних похідних функції $H(b_0, b_1)$ обчислимо безпосередньо з (7.62):

$$\begin{cases} \dfrac{\partial H}{\partial b_0} = -\sum_{i=1}^{n} (-1) = -n, \\ \dfrac{\partial H}{\partial b_1} = -\sum_{i=1}^{n} x_i. \end{cases} \qquad (7.63)$$

Легко переконатися, що оскільки вирази (7.63) не залежать від b_0 і b_1, то розраховувати на визначення значень b_0 і b_1, мінімізуючих функцію $H(b_0, b_1)$, не доводиться. Отже, зрозуміло, що мінімізація суми залишків виявляється невідповідним критерієм для визначення b_0 й b_1, втім цей результат потребує пояснення.

Розглянемо на площині (x, y) дві точки $N_1(x_1, y_1)$ й $N_2(x_2, y_2)$, за якими пряма (7.59) визначається однозначно, як показано на рис. 7.9. Легко також переконатися, що ця пряма перетворює функцію $H(b_0, b_1)$, тобто суму залишків, на нуль. Але будь-яка пряма $N_1'N_2'$, що проходить через точку N_0 – середину відрізка N_1N_2 (на рис. 7.9 ця пряма показана пунктиром), також перетворює функцію $H(b_0, b_1)$ на нуль.

Відповідно до умов, які докладно обговорювались у розд. 4.16 у разі введення дисперсії, доцільно мінімізувати не суму залишків, а суму їхніх квадратів. Тому цей метод визначення значень b_0 й b_1 називають **методом найменших квадратів**. Це означає, що значення b_0 й b_1 потрібно знаходити з умови мінімуму функції, тобто

$$G(b_0, b_1) = \sum_{i=1}^{n} \Delta y_i^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2. \qquad (7.64)$$

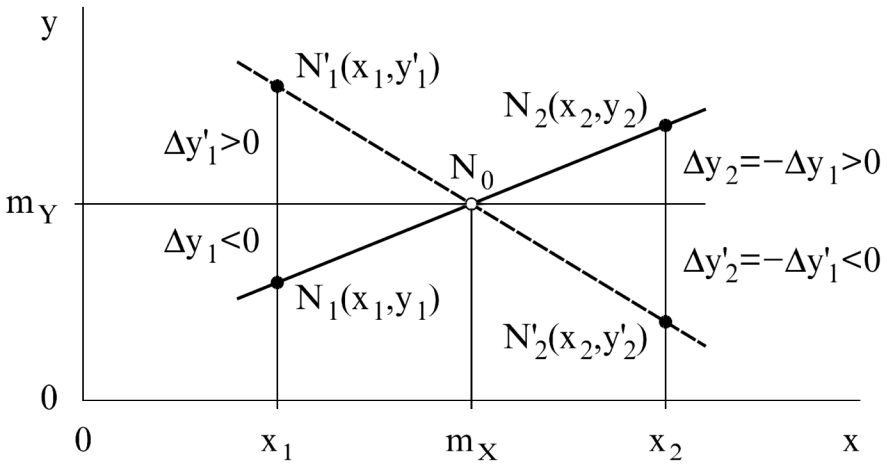


Рис. 7.9. Неоднозначність визначення прямої лінії регресії за умови мінімуму суми залишків $H(b_0, b_1)$ (7.58)

Необхідні умови мінімуму функції (7.61) приводять у цьому випадку до таких рівнянь для знаходження значень b_0 і b_1:

$$\begin{cases} \dfrac{\partial G}{\partial b_0} = -\sum\limits_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] = 0\,, \\ \dfrac{\partial G}{\partial b_1} = -2\sum\limits_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]\, x_i = 0\,. \end{cases} \tag{7.65}$$

У рівняннях (7.65) коефіцієнти за невідомих b_0 і b_1 можуть бути явно виділені, що приводить до лінійної системи двох рівнянь для двох невідомих:

$$\begin{cases} n\, b_0 + \left(\sum\limits_{i=1}^{n} x_i\right) b_1 = \sum\limits_{i=1}^{n} y_i \\ \left(\sum\limits_{i=1}^{n} x_i\right) b_0 + \left(\sum\limits_{i=1}^{n} x_i^2\right) b_1 = \sum\limits_{i=1}^{n} y_i x_i \end{cases}. \tag{7.66}$$

Розв’язок системи (7.66), одержаний методом виключення невідомих, має такий вигляд:

$$\begin{cases} b_0 = \dfrac{\dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i^2\right)\left(\sum\limits_{i=1}^{n} y_i\right) - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} x_i y_i\right)}{\sum\limits_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i\right)}, \\ b_1 = \dfrac{\sum\limits_{i=1}^{n} x_i y_i - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{\sum\limits_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} x_i\right)}. \end{cases} \tag{7.67}$$

У розд. 7.8 буде зазначена компактна форма запису розв'язку (7.67).

Під (генеральною) регресією-2 розуміється та ж функція, що використовується й для вибіркової регресії-2, але параметри якої знаходяться не за вибірковим значенням, а за всією генеральною сукупністю. Параметри генеральної регресії-2 будемо позначати буквою β з індексами. Наприклад, для генеральної лінійної регресії-2

$$\overline{y} = \beta_0 + \beta_1 x. \tag{7.68}$$

До задачі визначення параметрів у регресії-2 прилучається інша задача, що виникає за іншої організації дослідів. Мова йде про експерименти, у яких значення однієї випадкової кількісної ознаки Y вимірюється за відомих значень іншої невипадкової кількісної ознаки X. Ознакою X може бути, наприклад, час, за різних значень якого вимірюються значення випадкової ознаки Y. Подібна ситуація виникає, скажімо, у випадку калібрування приладів, коли значення ознаки X є відомими і відповідають спеціально підготовленим зразкам, а випадкові значення ознаки Y вимірюються калібрувальним приладом. Набір даних у таких дослідах звичайно сформований так. Для декількох значень X накопичуються серії значень Y, за якими потрібно підібрати залежності вигляду

$$\overline{y} = f(x; \textit{параметри}). \tag{7.69}$$

Такі задачі називаються задачами вирівнювання даних, але розв'язуються вони так само, як і задачі визначення параметрів для вибіркової регресії-2, за допомогою методу найменших квадратів. Надалі будемо розглядати тільки задачі регресії-2, але тип регресії вказувати не будемо.

7.8. Кореляція

Знову розглянемо кореляційне поле, побудоване за вибірковими значеннями, (x_i, y_i), $i = 1, \ldots, n$, і проаналізуємо його окремі випадки. Основна задача даного розділу – ввести кількісний показник, "що вимірює" ступінь залежності неперервних випадкових величин X і Y. Відправною точкою для введення такого показника вибираються властивості 5 і 6 для дисперсії $D[X \pm Y]$ суми (різниці) неперервних випадкових величин X і Y (див. розд. 7.6, табл. 7.12). Порівняння цих властивостей дозволяє припустити, що коваріація $K[X, Y]$ може бути випробувана як міра залежності, що перетворюється на нуль для незалежних неперервних випадкових величин X і Y.

Спочатку розглянемо випадок лінійної залежності неперервних випадкових величин X і Y

$$Y = \beta_0 + \beta_1 X, \tag{7.70}$$

причому закон розподілу неперервної випадкової величини X вважатимо довільним. Очевидно, що рівняння вибіркової й генеральної регресії-1 і регресії-2 мають вигляд лінійної функції з тими ж параметрами $\beta_0 = b_0$ і $\beta_1 = b_1$, що й в (7.59) і (7.68). Кореляційне поле складається із точок, розташованих на лінії регресії (рис. 7.10).

Обчислимо вибіркову коваріацію $s_{X,Y}$ за вибірковими значеннями (x_i, y_i), $i = 1, \ldots, n$, причому потрібно враховувати, що виконуються рівності

$$y_i = \beta_0 + \beta_1 x_i. \tag{7.71}$$

Із (7.71) випливає, що й для вибіркових середніх значень m_X і m_Y неперервних випадкових величин X і Y справедливий подібний зв'язок. Дійсно, знаходячи суми лівих і правих частин виразів (7.68) для всіх вибіркових значень, одержуємо

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i, \tag{7.72}$$

звідки, після ділення на n, маємо

$$m_Y = \beta_0 + \beta_1 m_X. \tag{7.73}$$

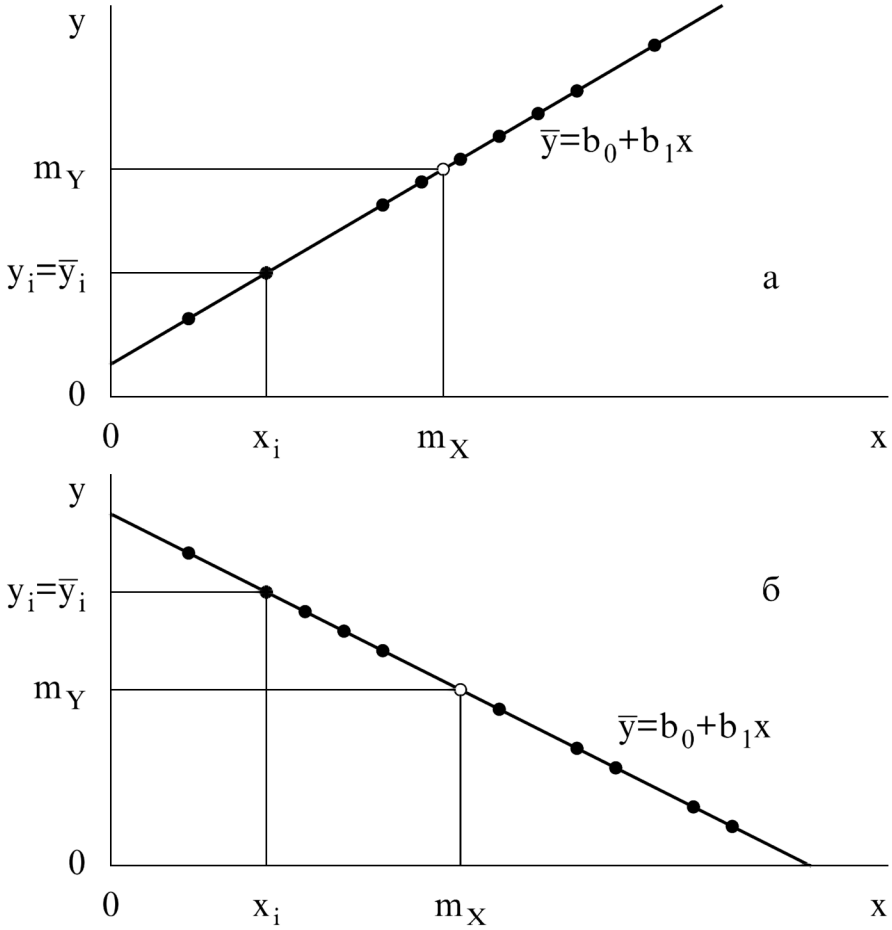


Рис. 7.10. Лінійна регресія між випадковими неперевними лінійно залежними величинами X і Y: а – додатна ($b_1 > 0$); б – від'ємна ($b_1 < 0$)

За визначенням вибіркової коваріації (7.45)

$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 m_X) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)\,\beta_1(x_i - m_X) = \beta_1\,\frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)^2\,, \qquad (7.74)$$

тобто

$$s_{X,Y} = \beta_1\, s_X^2\ . \qquad (7.75)$$

Якщо з виразу для вибіркової коваріації (7.45) вилучати не значення y_i й m_Y, а x_i й m_X

$$x_i = \frac{y_i - \beta_0}{\beta_1}, \qquad m_X = \frac{m_Y - \beta_0}{\beta_1}, \qquad \beta_1 \neq 0, \tag{7.76}$$

то

$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0}{\beta_1} - \frac{m_Y - \beta_0}{\beta_1}\right)(y_i - m_Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\beta_1}(y_i - m_Y)^2 = \tag{7.77}$$

$$= \frac{1}{\beta_1}\frac{1}{n}\sum_{i=1}^{n}(y_i - m_Y)^2 ,$$

тобто

$$s_{X,Y} = \frac{1}{\beta_1} s_Y^2 . \tag{7.78}$$

Знак вибіркової коваріації визначається знаком β_1, тому

$$s_{X,Y} = \operatorname{sgn}(\beta_1)\sqrt{s_X^2}\sqrt{s_Y^2} = \operatorname{sgn}(\beta_1)\, s_X s_Y . \tag{7.7}$$

Сформулюємо висновки, що випливають із (7.79).

По-перше, коваріація лінійно залежних неперервних випадкових величин з точністю до знака дорівнює добутку їх вибіркових середньоквадратичних відхилень.

По-друге, знак коваріації може бути визначений за нахилом лінії регресії (див. рис. 7.10).

Якщо врахувати, що за рахунок вибору одиниць виміру неперервних випадкових величин X і Y величину коваріації можна зробити як завгодно великою і як завгодно малою, то перехід до величини

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y} \tag{7.80}$$

означає нормування коваріації, тобто значення нової величини, називаної **вибірковим коефіцієнтом кореляції**, у випадку лінійно залежних неперервних випадкових

величин X і Y завжди дорівнює ±1. Крім того, $[s_{X,Y}]=[X][Y]$, і якщо, приміром, $[X]=г/л$, а $[Y]=кг$, то розмірність $s_{X,Y}$ є фізично беззмістовною (див. розд. 4.11). Величина $r_{X,Y}$ позбавлена цього недоліку, оскільки вибірковий коефіцієнт кореляції безрозмірний.

Можна довести, що за довільних значень (x_i, y_i), $i=1,\ldots,n$ завжди виконується умова

$$-1 \le r_{X,Y} \le +1. \tag{7.81}$$

Звернемо також увагу на те, яким чином формується знак вибіркової коваріації $s_{X,Y}$, що визначає знак вибіркового коефіцієнта кореляції $r_{X,Y}$. Значення $s_{X,Y}$ є середнє арифметичне добутку відхилень $\Delta x_i = x_i - m_X$ і $\Delta y_i = y_i - m_Y$ вибіркових значень від своїх середніх значень. За $\beta_1 > 0$ більшість вибіркових значень (x_i, y_i) розташовується або в I, або в III чвертях, утворених прямими $x = m_X$ й $y = m_Y$. За $\beta_1 < 0$ більшість вибіркових значень (x_i, y_i) розташована або в II, або в IV чвертях. Знаки відхилень Δx_i, Δy_i і знаки їхніх добутків у різних чвертях за $\beta_1 > 0$ й $\beta_1 < 0$ зображені на рис. 7.11.

Кореляцію (не обов'язково у випадку лінійної залежності між неперервними випадковими величинами X і Y) з додатним коефіцієнтом β_1 будемо називати додатною, а з від'ємним коефіцієнтом β_1 – від'ємною (рис. 7.10).

Продовжимо розгляд окремих випадків кореляційного поля. Нехай неперервні випадкові величини X і Y є незалежні, причому кожна з них має симетричну функцію густини. За досить великого обсягу вибірки можна вважати, що симетрія буде спостерігатись і для гістограм частот, відносних частот і вибіркових функцій густини. Якщо в досліді виконується умова стабілізації відносних інтервальних частот, то будемо розглядати інтервальні випадкові величини X і Y як такі, що задовольняють умову незалежності. Виберемо для визначеності число інтервалів K=5 і L=7 і розглянемо інтервальні ряди розподілу, зображені на рис. 7.12.

Із умови незалежності інтервальних випадкових величин X і Y безпосередньо випливає, що інтервальні ряди розподілу X і Y є маргінальні для інтервальної таблиці спільного розподілу, причому вона може бути відновлена за маргінальними рядами X і Y. На рис. 7.12 клітинки інтервальної таблиці заповнені відповідними значеннями $\overset{*}{p}_{k,l} = \overset{*}{p}_k \overset{*}{p}_l$, за якими проведено ізолінії рівної густини розташування точок кореляційного поля за досить великої їхньої кількості.

Загальний висновок розгляду кореляційного поля для незалежних неперервних випадкових величин X і Y полягає в тому, що їхнє кореляційне поле має овальну форму. Сплюснутість овалу визначається співвідношенням дисперсій величин X і Y.

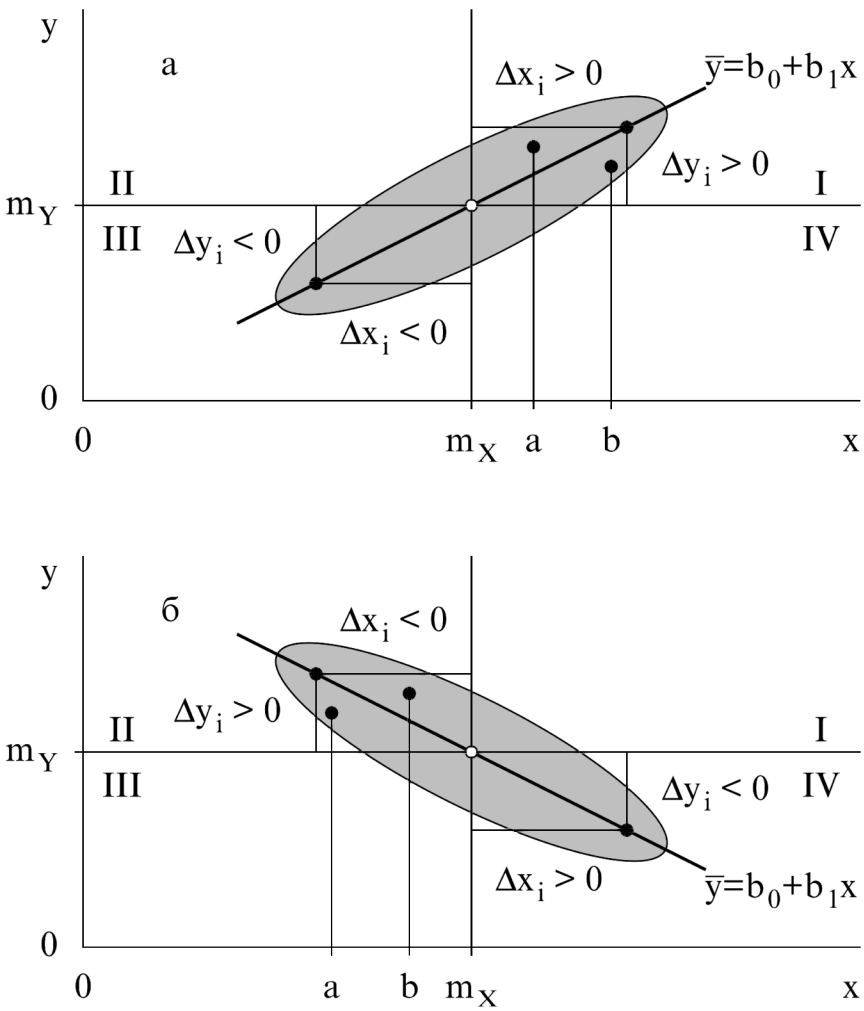


Рис. 7.11. Розташування вибіркових значень на кореляційному полі залежно від знака кореляції: а – додатна ($b_1 > 0$); б – від'ємна ($b_1 < 0$)

Легко також переконатися, що вибіркові лінії регресії-1 і регресії-2 для незалежних величин повинні бути горизонтальними у разі регресії Y по X і вертикальними у разі регресії X по Y. Це означає, що за зміни значень однієї величини значення іншої в середньому не змінюються. Цей висновок складає досить надійне первинне правило, що дозволяє судити про незалежність неперервних випадкових величин X і Y. Обґрунтування цього правила проведемо в розд. 10.5.

Простий розгляд знаків відхилень Δx_i, Δy_i у кожній із чвертей на кореляційному полі з урахуванням приблизно рівного числа точок, що зображують вибіркові значення (x_i, y_i), $i = 1, \ldots, n$, приводить до висновку, що в цьому випадку

$$s_{X,Y} \approx 0 \quad \Rightarrow \quad r_{X,Y} \approx 0 . \tag{7.82}$$

Ліва частина наближеної рівності (7.79) є математичний вираз правила, що дозволяє робити попередній висновок про незалежність лінійно зв'язаних випадкових величин.

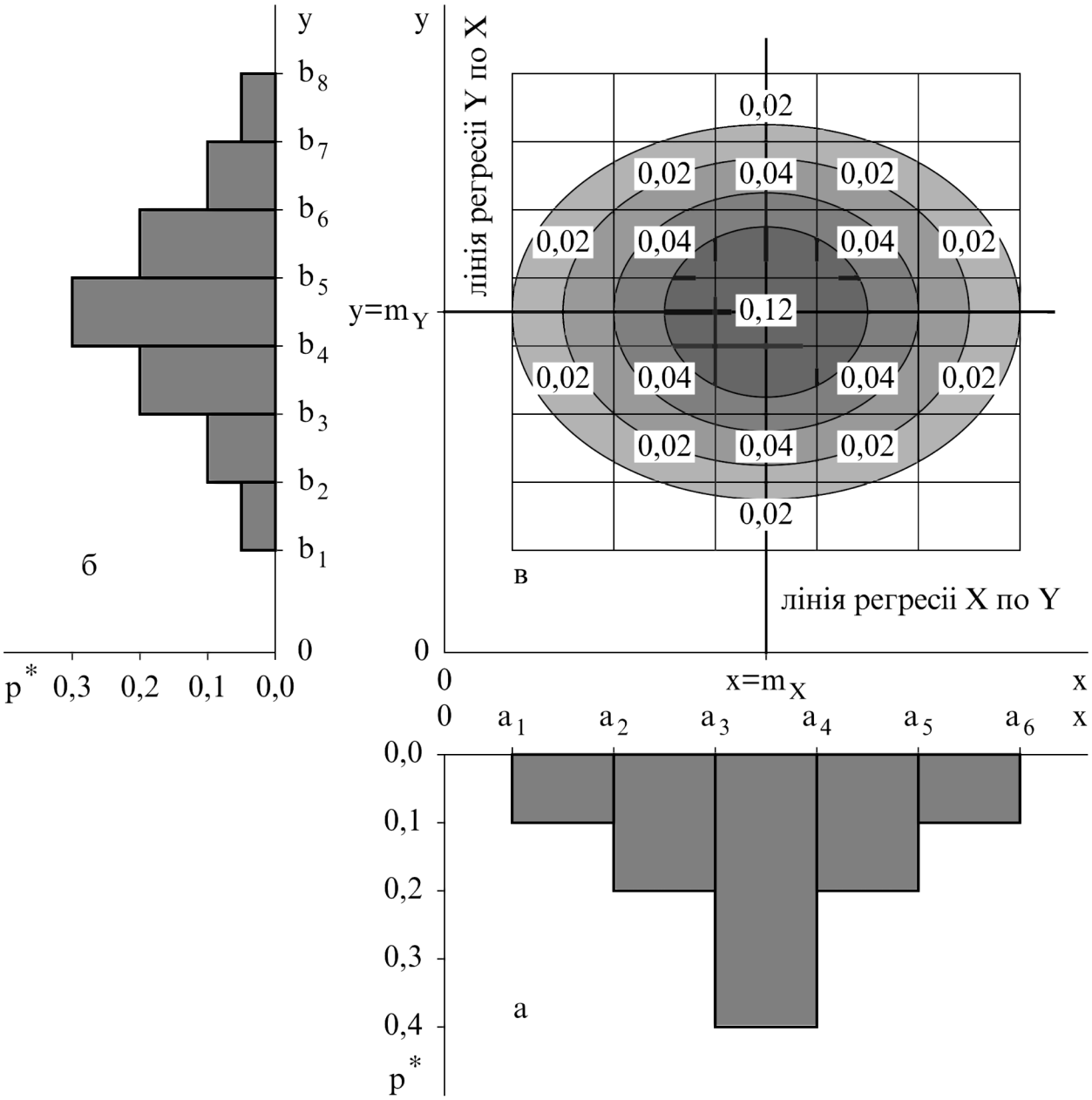


Рис. 7.12.Ілюстрація до первинного правила визначення незалежності лінійно зв'язаних неперевних випадкових величин X і Y із симетричними розподілами: а, б – ряди розподілу інтервальних величин, що віповідають величинам X і Y; в – кореляційне поле для інтервальних величин з лініями рівної густини точок

Однак було б помилкою вважати, що близьке до нуля значення вибіркового коефіцієнта кореляції (навіть за досить великого обсягу вибірки) вказує на незалежність неперервних випадкових величин X і Y.

Приклад 7.4. Кореляційне поле (рис. 7.13) для неперервних випадкових величин X і Y із квадратичним функціональним зв'язком

$$Y = X^2, \tag{7.83}$$

у випадку довільного (!) симетричного закону розподілу X вказує на те, що

$$\rho_{X,Y} = 0, \qquad r_{X,Y} \approx 0. \tag{7.84}$$

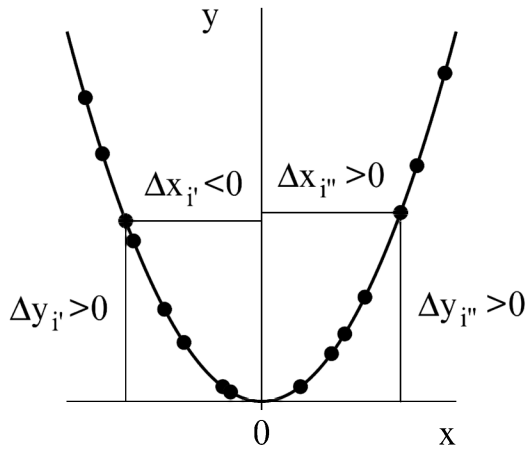


Рис. 7.13. Кореляційне поле та лінія регресії для квадратично залежних випадкових величин (див. прикл. 7.4 та (7.80))

Даний приклад указує також на те, що коефіцієнт кореляції погано пристосований для виміру незалежності нелінійно зв'язаних випакових величин. ●

Неперервні випадкові величини X і Y, для яких $\rho_{X,Y} = 0$, називаються **некорельованими**. Отже, із некорельованості неперервних випадкових величин X і Y не випливає їхня незалежність. Є, однак, один винятковий випадок.

Якщо неперервні випадкові величини X і Y **розподілені за нормальним законом** із параметрами μ_X, σ_X і μ_Y, σ_Y, **то із некорельованості** X і Y **випливає їх незалежність**.

Розглянемо далі випадок, що є проміжний між лінійною залежністю неперервних випадкових величин X і Y і їхньою незалежністю. Цей випадок сконструюємо штучно в такий спосіб. Нехай лінія регресії буде пряма, але зв'язок між Y і X не є лінійний функціональний, і є розсіювання точок кореляційного поля відносно лінії регресії. Кореляційне поле являє собою смужку, усередині якої розташована лінія регресії (рис. 7.11).

Додатна кореляція в цьому випадку відрізняється від додатної кореляції у випадку лінійного функціонального зв'язку X і Y (див. рис. 7.10) тим, що за збільшення значень X збільшуються не всі значення Y. Наприклад, на рис. 7.11 точками показані значення ознаки Y за $X=a$ і $X=b$. Видно, що більшому значенню X відповідає менше значення Y. Отже, у випадку додатної кореляції збільшенню значень X відповідає, як правило, збільшення значень Y, однак є й виняток із цієї тенденції. Знак кутового коефіцієнта β_1 лінійної регресії (7.76) визначає знак кореляційного зв'язку надійніше.

Зкореляційного поля (рис. 7.11) видно також, як формується знак вибіркової коваріації $s_{X,Y}$, а отже, і знак вибіркового коефіцієнта кореляції $r_{X,Y}$. Більша частина точок кореляційного поля за $\beta_1 > 0$ знаходиться в I і III чвертях, де $\Delta x_i \Delta y_i > 0$, а за $\beta_1 < 0$ – в II і IV чвертях, де $\Delta x_i \Delta y_i < 0$. Однак із зростанням розсіювання значень неперервної випадкової величини Y відносно її середнього значення за сталого значення неперервної випадкової величини X (тобто за збільшення умовної вибіркової дисперсії Y) число точок в II і IV чвертях за $\beta_1 > 0$ й в I і III чвертях за $\beta_1 < 0$ збільшується, а отже, зменшується й значення $s_{X,Y}$, а з ним і $r_{X,Y}$. У граничному випадку, коли умовна дисперсія неперервної випадкової величини Y зрівняється з її дисперсією, матимемо випадок незалежності величин X і Y, за якого лінія регресії горизонтальна, тобто її кутовий коефіцієнт дорівнює нулю, $\beta_1 = 0$. Дані висновки у випадку включення до розгляду усіх можливих пар значень (x_i, y_i) із генеральної сукупності можуть бути перенесені на коваріацію $K[X,Y] = \sigma_{X,Y}$ і коефіцієнт кореляції $\rho_{X,Y}$.

Точне формулювання цього результату таке. Зв'язок між дисперсією σ_Y^2, умовною дисперсією $\sigma_{Y|X=x}^2$ й коефіцієнтом кореляції $\rho_{X,Y}$ має такий вигляд:

$$\frac{\sigma_Y^2 - \sigma_{Y|X=x}^2}{\sigma_Y^2} = \rho_{X,Y}^2 , \qquad (7.85)$$

при цьому вважається, що $\sigma_{Y|X=x}^2 = const$, тобто розсіювання по Y відносно умовного математичного сподівання $\mu_{Y|X=x}$ те ж саме за всіх значень $X=x$. Отже, квад-

рат коефіцієнта кореляції між неперервними випадковими величинами X і Y дорівнює відносному зменшенню дисперсії Y у випадку переходу від максимального інтервалу зміни X (розмаху значень величини X) до фіксованого значення X (у разі набуття неперервною випадковою величиною X певного значення x).

Коефіцієнт кореляції $\rho_{X,Y}$ і є кількісною мірою зв'язку між неперервними випадковими величинами X і Y, для яких підходить лінійна регресія $\overline{y} = \beta_0 + \beta_1 x$. При цьому граничним значенням для оцінки зв'язку між X і Y є значення $\rho^2_{X,Y} = 0{,}5$, тобто $\rho_{X,Y} \approx 0{,}7$. У випадку, коли $\rho_{X,Y} > 0{,}7$, випадкова величина Y залежить від випадкової величини X більше, ніж на 50% (розсіювання відносно $\mu_{Y|X=x}$ визначається іншими причинами), а коли $\rho_{X,Y} < 0{,}7$, випадкова величина Y залежить від випадкової величини X менше, ніж на 50%.

Підіб'ємо підсумки і сформулюємо стисло властивості коефіцієнта кореляціїї, з яких властивсті 1–3 вище були обгрунтовані і проілюстровані прикладами, властивості 4, 5 наведені нижче без доведень.

Властивості коефіцієнта кореляції

1. Коефіцієнт кореляції дозволяє вимірювати величину й знак лінійного (не функціонального) зв'язку між неперервними випадковими величинами X і Y.

2. Для незалежних неперервних випадкових величин X і Y $\rho_{X,Y} = 0$ (X і Y некорельовані).

3. Із умови $\rho_{X,Y} = 0$ (некорельованості неперервних випадкових величин X і Y) не випливає, що X і Y незалежні (у тому випадку, коли лінія регресії Y по X або X по Y не є пряма, тобто саме тоді, коли коефіцієнт кореляції й не повинен застосовуватися для оцінки зв'язку між X і Y). Із цієї властивості випливає, що перед обчисленням значення $r_{X,Y}$ слід подивитися на кореляційне поле.

4. Для нормально розподілених неперервних випадкових величин X і Y із некорельованості X і Y випливає незалежність X і Y.

5. Із рівності $\left|\rho_{X,Y}\right| = 1$ випливає лінійний функціональний зв'язок між неперервними випадковими величинами X і Y і навпаки, для неперервних випадкових величин X і Y, зв'язаних лінійним функціональним зв'язком, коефіцієнт кореляції за абсолютною величиною дорівнює одиниці.

За допомогою вибіркового коефіцієнта кореляції можна також записати в стислій формі розв'язок (7.67) системи рівнянь (7.66) для коефіцієнтів лінійної регресіїї:

$$\begin{cases} b_1 = r_{X,Y} \dfrac{\hat{s}_Y}{\hat{s}_X}, \\ b_0 = m_Y - b_1 m_X. \end{cases} \tag{7.86}$$

8. Неперервні випадкові величини, похідні від нормально розподіленої випадкової величини

8.1. Ще одне визначення нормального розподілу

Наведене в розд. 5.3 визначення нормального розподілу є формальне і зовсім не пояснює ні назви, ні уваги, що йому приділяється під час статистичної обробки даних. Розглянемо інший підхід до введення нормального розподілу, заснований на одному з варіантів центральної граничної теореми. Формулювання її таке.

Нехай $U, V, W, \ldots$ – незалежні випадкові величини (дискретні або неперервні) з довільними законами розподілу, математичними сподіваннями

$$\mu_U = M[U], \quad \mu_V = M[V], \quad \mu_W = M[W], \ \ldots \tag{8.1}$$

й порівняними дисперсіями (тобто ці дисперсії є величинами одного порядку):

$$\sigma_U^2 = D[U] \sim \sigma_V^2 = D[V] \sim \sigma_W^2 = D[W] \sim \ \ldots, \tag{8.2}$$

а X – випадкова величина, що є сумою випадкових величин $U, V, W, \ldots$ (див. розд. 7.3, 7.7):

$$X = U + V + W + \ldots \ . \tag{8.3}$$

Із властивості 4 математичного сподівання суми довільних випадкових величин (див. розд. 7.4, табл. 7.25 і розд. 7.5, табл. 7.28) випливає, що

$$\mu = M[X] = M[U] + M[V] + M[W] + \ldots = \mu_U + \mu_V + \mu_W + \ldots \ , \tag{8.4}$$

а з властивості 3 дисперсії суми незалежних випадкових величин (див. розд. 7.4, табл. 7.25 і розд. 7.5, табл. 7.28) –

$$\sigma^2 = D[X] = D[U] + D[V] + D[W] + \ldots = \sigma_U^2 + \sigma_V^2 + \sigma_W^2 + \ldots \ . \tag{8.5}$$

Центральна гранична теорема стверджує, що за великої кількості випадкових величин-доданків в сумі (8.3) розподіл стандартизованої випадкової велични X як завгодно мало відрізняється від стандартизованого нормального розподілу (див. розд. 6.1), тобто

$$\frac{X - \mu}{\sigma} \sim \tilde{z} \equiv N(0,1) \, . \tag{8.6}$$

Зауваження 8.1. Наближення (8.6) з достаньою для практики точністю спостерігається вже за невеликої кількості доданків $U, V, W, \ldots$ в (8.3). ▲

Зауваження 8.2. Якщо $U, V, W, \ldots$ – неперевні випадкові величини, то за умов центральної граничної теореми можна наближати неперевну випадкову величину X (8.3) розподілом $N(\mu, \sigma)$, тобто сума неперевних випадкових величин $U, V, W, \ldots$ вже за невеликої їхньої кількості практично наближається нормальним розподілом

$$X = U + V + W + \ldots \sim N(\mu, \sigma) \tag{8.7}$$

з відповідними параметрами μ (8.4) і σ^2 (8.5). ▲

Зауваження 8.3. Якщо $U, V, W, \ldots$ – дискретні випадкові величини, то для їхньої суми X (8.3) неможливе наближення (8.6) нормальним розподілом, оскільки сума (8.3) є дискретна випадкова величина, а нормально розподілена випадкова величина – неперервна. Але це не стосується їхньої стандартизованої суми за великої кількості доданків $U, V, W, \ldots$. Також зазначимо, що з центральної граничної теореми випливає, що сума X (8.3) дискретних випадкових величин є дискретна випадкова величина з симетричною функцією густини $f(x)$. ▲

Приклад 8.1. Розглянемо суму незалежних дискретних випадкових величин U і V з рядами розподілу, зображеними в табл. 8.1. і 8.2 і на рис. 8.1.

Таблиця 8.1

U	–1	0	+1
	1/3	1/3	1/3

Таблиця 8.2

V	–1	0	+1
	1/3	1/3	1/3

Відповідно до визначення суми двох дискретних випадкових величин (див. розд. 7.6)

1) утворимо прямий добуток множин значень U і V:

{(–1, –1); (–1, 0); (–1, +1); (0, –1); (0, 0); (0, +1); (+1, –1); (+1, 0); (+1, +1)};

2) обчислимо за правилом добутку для незалежних випадкових подій імовірності пар у прямому добутку:

1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3; 1/3·1/3;

3) обчислимо суму значень у кожній парі: {–2; –1; 0; –1; 0; +1; 0; +1; +2};

4) складемо з одержаних значень і їхніх ймовірностей ряд розподілу дискретної випадкової величини $U+V$ у вигляді табл. 8.3, показаний також на рис. 8.1.

Таблиця 8.3

$Y=U+V$	–2	–1	0	+1	+2
	1/9	2/9	3/9	2/9	1/9

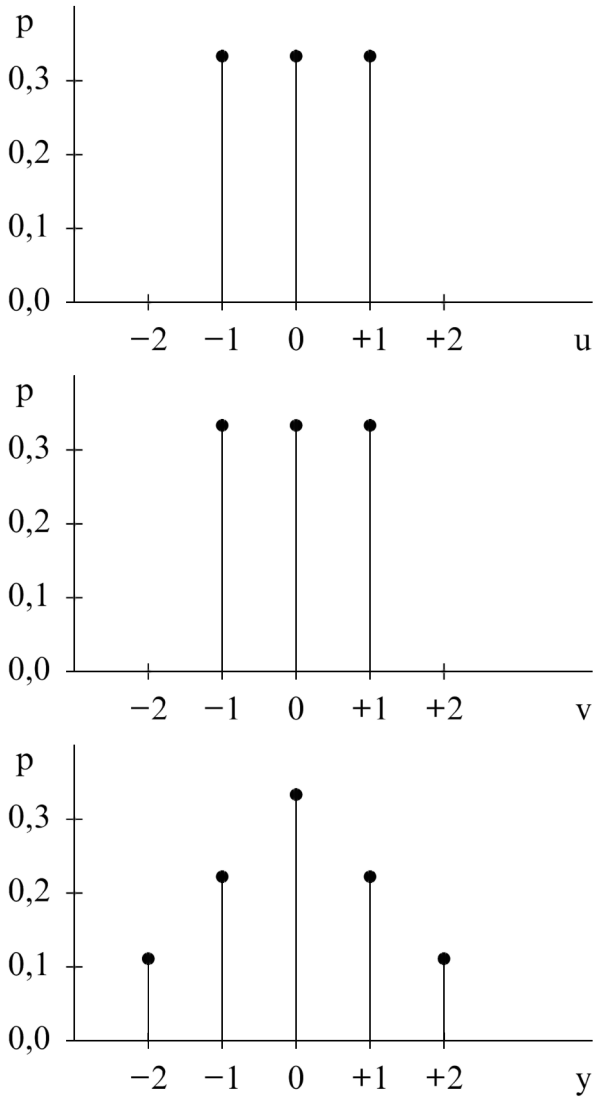


Рис. 8.1. Ряди розподілу дискретних випадкових величин U, V і $Y = U + V$ із прикл. 8.1 (див. табл. 8.1, 8.2 і 8.3)

У даному прикладі симетричний розподіл суми Y дискретних випадкових величин U і V з симетричними розподілами, звичайно, не є несподіванкою. ●

Приклад 8.2. Розглянемо дискретні випадкові величини *U*, *V* і *W*, ряди яких наведені в табл. 8.4, 8.5 і 8.6 і показані на рис. 8.2.

Таблиця 8.4

U	–1	0	+1	+2
	4/10	3/10	2/10	1/10

Таблиця 8.5

V	–1	0	+1
	1/3	1/3	1/3

Таблиця 8.6

W	–2	–1	0	+1
	1/10	2/10	3/10	4/10

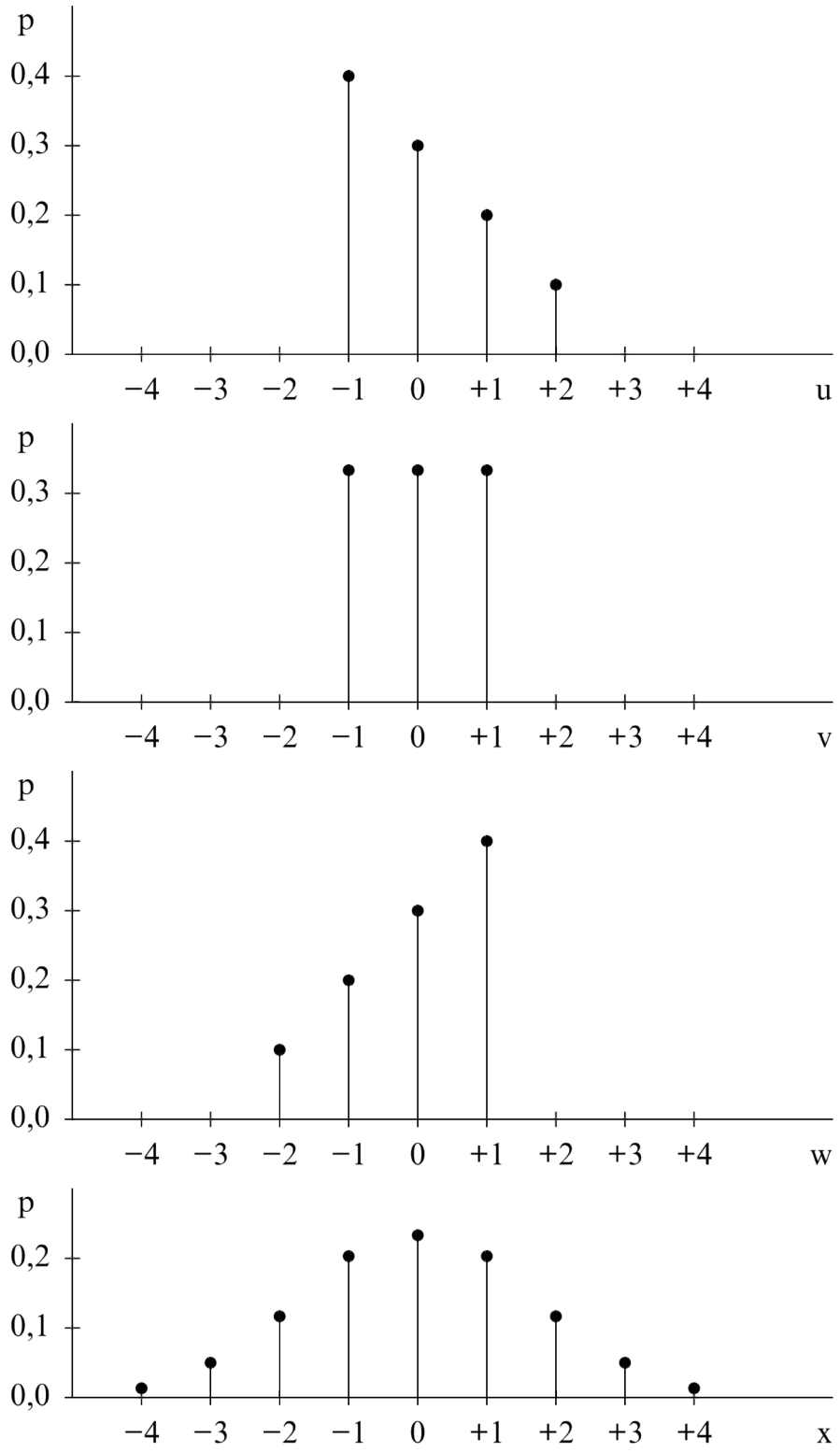


Рис. 8.2.Ряди розподілу дискретних випадкових величин U, V, W і $X = U + V + W$ із прикл. 8.2 (див. табл. 8.4, 8.5, 8.6 і 8.8)

Утворимо прямий добуток множин значень дискретних випадкових величин U, V і W, причому під кожним зі значень випишемо їхні ймовірністі віповідно до табл. 8.4, 8.5 і 8.6:

(u=–1;	v=–1;	w=–2);	(u=–1;	v=–1;	w=–1);	(u=–1;	v=–1;	w=0);
4/10	1/3	4/10	4/10	1/3	2/10	4/10	1/3	3/10
(u=–1;	v=–1;	W=+1);	(u=–1;	v=0;	w=–2);	(u=–1;	v=0;	w=–1);
4/10	1/3	4/10	4/10	1/3	4/10	4/10	1/3	2/10
(u=–1;	v=0;	W=0);	(u=–1;	v=0;	w=+1);	(u=–1;	v=+1;	w=–2);
4/10	1/3	3/10	4/10	1/3	4/10	4/10	1/3	1/10
(u=–1;	v=+1;	w=–1);	(u=–1;	v=+1;	w=0);	(u=–1;	v=+1;	w=+1);
4/10	1/3	2/10	4/10	1/3	3/10	4/10	1/3	4/10
(u=0;	v=–1;	W=–2);	(u=0;	v=–1;	w=–1);	(u=0;	v=–1;	w=0);
3/10	1/3	1/10	3/10	1/3	2/10	3/10	1/3	3/10
(u=0;	v=–1;	W=+1);	(u=0;	v=0;	w=–2);	(u=0;	v=0;	w=–1);
3/10	1/3	4/10	3/10	1/3	1/10	3/10	1/3	2/10
(u=0;	v=0;	W=0);	(u=0;	v=0;	w=+1);	(u=0;	v=+1;	w=–2);
3/10	1/3	3/10	3/10	1/3	4/10	3/10	1/3	1/10
(u=0;	v=+1;	w=–1);	(u=0;	v=+1;	w=0);	(u=0;	v=+1;	w=+1);
3/10	1/3	2/10	3/10	1/3	3/10	3/10	1/3	4/10
(u=+1;	v=–1;	w=–2);	(u=+1;	v=–1;	w=–1);	(u=+1;	v=–1;	w=0);
2/10	1/3	1/10	2/10	1/3	2/10	2/10	1/3	3/10
(u=+1;	v=–1;	w=+1);	(u=+1;	v=0;	w=–2);	(u=+1;	v=0;	w=–1);
2/10	1/3	4/10	2/10	1/3	1/10	2/10	1/3	2/10
(u=+1;	v=0;	w=0);	(u=+1;	v=0;	w=+1);	(u=+1;	v=+1;	w=–2);
2/10	1/3	3/10	2/10	1/3	4/10	2/10	1/3	1/10
(u=+1;	v=+1;	w=–1);	(u=+1;	v=+1;	w=0);	(u=+1;	v=+1;	w=+1);
2/10	1/3	2/10	2/10	1/3	3/10	2/10	1/3	4/10
(u=+2;	v=–1;	w=–2);	(u=+2;	v=–1;	w=–1);	(u=+2;	v=–1;	w=0);
1/10	1/3	1/10	1/10	1/3	2/10	1/10	1/3	3/10
(u=+2;	v=–1;	w=+1);	(u=+2;	v=0;	w=–2);	(u=+2;	v=0;	w=–1);
1/10	1/3	4/10	1/10	1/3	1/10	1/10	1/3	2/10
(u=+2;	v=0;	w=0);	(u=+2;	v=0;	w=+1);	(u=+2;	v=+1;	w=–2);
1/10	1/3	3/10	1/10	1/3	4/10	1/10	1/3	1/10
(u=+2;	v=+1;	w=–1);	(u=+2;	v=+1;	w=0);	(u=+2;	v=+1;	w=+1)}
1/10	1/3	2/10	1/10	1/3	3/10	1/10	1/3	4/10

Обчислимо суму значень у кожній трійці та ймовірність кожної трійки за правилом добутку для незалежних подій. Результати зведемо в табл. 8.7, перший і останній рядки якої утворять ряд розподілу дискретної випадкової величини $Y = U + V + W$. Для зручності ряд розподілу суми $Y = U + V + W$ зображений також у вигляді окремої табл. 8.8. і рис. 8.2.

Таблиця 8.7

–4	–3	–2	–1	0	+1	+2	+3	+4	
4/10·1/3·1/10	4/10·1/3·2/10	4/10·1/3·3/10	4/10·1/3·4/10	4/10·1/3·4/10	4/10·1/3·4/10	3/10·1/3·4/10	2/10·1/3·4/10	1/10·1/3·4/10	1
	4/10·1/3·1/10	4/10·1/3·2/10	4/10·1/3·3/10	4/10·1/3·3/10	3/10·1/3·4/10	2/10·1/3·4/10	1/10·1/3·4/10		2
	3/10·1/3·1/10	4/10·1/3·1/10	4/10·1/3·2/10	3/10·1/3·4/10	3/10·1/3·3/10	2/10·1/3·3/10	1/10·1/3·3/10		3
		3/10·1/3·2/10	3/10·1/3·3/10	3/10·1/3·3/10	3/10·1/3·2/10	1/10·1/3·4/10			4
		3/10·1/3·1/10	3/10·1/3·2/10	3/10·1/3·2/10	2/10·1/3·3/10	1/10·1/3·3/10			5
		2/10·1/3·1/10	3/10·1/3·1/10	2/10·1/3·3/10	2/10·1/3·2/10	1/3·2/10·1/10			6
			2/10·1/3·2/10	2/10·1/3·2/10	2/10·1/3·3/10				7
			2/10·1/3·1/10	2/10·1/3·1/10	1/10·1/3·2/10				8
			1/10·1/3·4/10	1/10·1/3·2/10	1/10·1/3·1/10				9
				1/10·1/3·1/10					10
$\frac{4}{300}$	$\frac{15}{300}$	$\frac{35}{300}$	$\frac{61}{300}$	$\frac{70}{300}$	$\frac{61}{300}$	$\frac{35}{300}$	$\frac{15}{300}$	$\frac{4}{300}$	$\frac{300}{300}$
0,01(3)	0,05	0,11(6)	0,20(3)	0,2(3)3	0,20(3)	0,11(6)	0,05	0,01(3)	1

Таблиця 8.8

$X=U+V+W$	–4	–3	–2	–1	0	+1	+2	+3	+4
	0,01(3)	0,05	0,11(6)	0,20(3)	0,2(3)	0,20(3)	0,11(6)	0,05	0,01(3)

Симетричність розподілу суми X доданків U, V, W з несиметричними розподілами не є попередньо очевидна, проте гарантується центральною граничною теоремою, оскільки її умови виконані. ●

Утворення з інтервальних випадкових величин суми із симетричним законом розподілу наведено у розд. 6.8 (див. прикл. 6.12).

Тепер стає зрозумілою причина симетричності розподілів багатьох фізіологічних і морфологічних кількісних показників у живих організмах і близькості їх до нормального закону. Кожний із таких показників складається під впливом декількох причин (факторів), кожна з яких чинить хоч і невеликий, але порівняно з іншими певний вплив на кінцевий результат.

8.2. Розподіл Пірсона $\tilde{\chi}^2$

Нехай $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_\nu$ – попарно незалежні неперервні випадкові величини (див. розд. 7.3), розподілені за законом $N(0,1)$ (див. розд. 6.1), тоді за визначенням неперервна випадкова величина

$$\tilde{\chi}_\nu^2 = \tilde{z}_1^2 + \tilde{z}_2^2 + \ldots + \tilde{z}_\nu^2 \tag{8.8}$$

має розподіл Пірсона $\tilde{\chi}_\nu^2$ з параметром ν („хі“-квадрат із числом степенів вільності „ню“).

Основні числові характеристики – математичне сподівання і дисперсія неперервної випадкової величини, розподіленої за законом Пірсона $\tilde{\chi}_\nu^2$, виражаються безпосередньо через параметр ν:

$$M[\tilde{\chi}_\nu^2] = \nu, \tag{8.9}$$

$$D[\tilde{\chi}_\nu^2] = 2\nu. \tag{8.10}$$

Графіки функцій густини $f(\chi_\nu^2)$ і розподілу $F(\chi_\nu^2)$ наведені відповідно на рис. 8.3 і 8.4. За малих значень параметра ν функція густини $f(\chi_\nu^2)$ має довгий правий хвіст, а у разі збільшення значень параметра ν ($\nu > 100$) асимптотично стає майже симетричною і наближається нормальним розподілом з математичним сподіванням (8.9) і дисперсією (8.10).

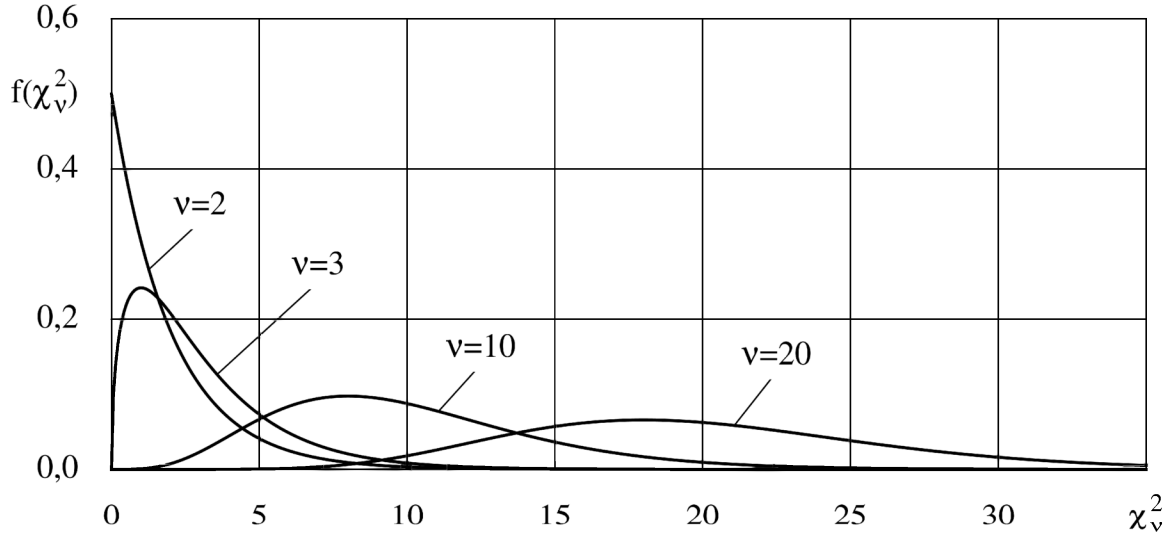


Рис. 8.3. Функція густини $f(\chi_\nu^2)$ Пірсона (8.8) залежно від значень степенів вільності ν

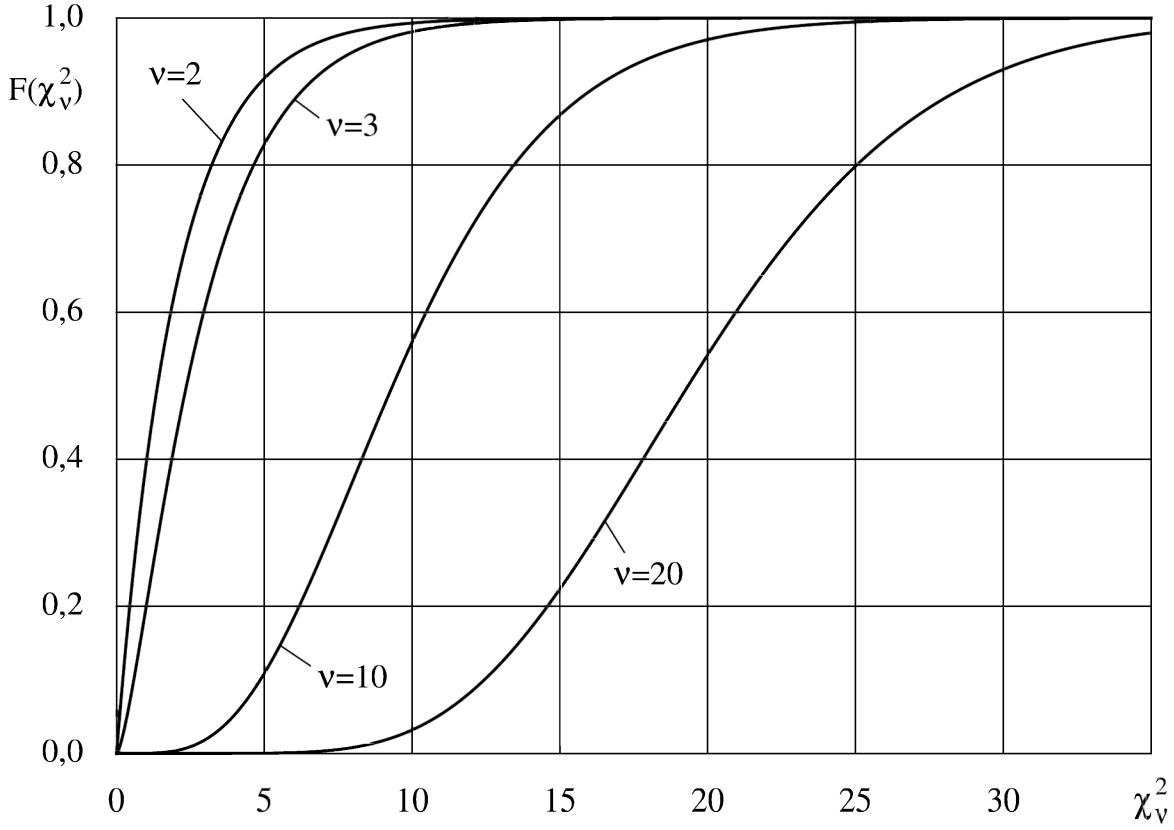


Рис. 8.4. Функція розподілу $F(\chi_\nu^2)$ Пірсона (8.8) залежно від значень степенів вільності ν

Застосування випадкової величини $\tilde{\chi}^2$, наприклад, за „ручної“ перевірки статистичних гіпотез (див. розд. 10), значно спрощується у разі використання таблиці квантилів $\chi^2_{\nu,q}$ (див. розд. 4.17), наведеної в табл. Д. 1 додатку.

8.3. Розподіл Фішера *F*

Нехай $\tilde{\chi}^2_{\nu_1}$ і $\tilde{\chi}^2_{\nu_2}$ – випадкові величини, розподілені за законом Пірсона $\tilde{\chi}^2_{\nu}$ із числом степенів вільності відповідно ν_1 й ν_2 (див. розд. 8.2), тоді за визначенням неперервна випадкова величина

$$\tilde{F}_{\nu_1,\nu_2} = \frac{\tilde{\chi}^2_{\nu_1} / \nu_1}{\tilde{\chi}^2_{\nu_2} / \nu_2} \tag{8.11}$$

має розподіл Фішера F_{ν_1,ν_2} з параметрами ν_1 й ν_2 (розподіл Фішера із числом степенів вільності чисельника ν_1 й знаменника ν_2).

Основні числові характеристики – математичне сподівання й дисперсія неперервної випадкової величини, розподіленої за законом Фішера F_{ν_1,ν_2}, виражаються безпосередньо через параметри ν_1 й ν_2:

$$M[F_{\nu_1,\nu_2}] = \frac{\nu_1}{\nu_2 - 2}, \tag{8.12}$$

$$D[F_{\nu_1,\nu_2}] = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}. \tag{8.13}$$

Графіки функцій густини $f(F_{\nu_1,\nu_2})$ і розподілу $F(F_{\nu_1,\nu_2})$ наведені відповідно на рис. 8.5 і 8.6. За малих значень параметрів ν_1 і ν_2 функція густини має довгий правий хвіст, а за збільшення значень ν_1 і ν_2 асимптотично стає майже симетричною й наближається нормальним розподілом з математичним сподіванням (8.12) і дисперсією (8.13).

Застосування випадкової величини F_{ν_1,ν_2}, наприклад, за „ручної“ перевірки статистичних гіпотез (див. розд. 10), значно спрощується за використання таблиці квантилів $F_{\nu_1,\nu_2,q}$ (див. розд. 4.17), наведеної в табл. Д. 2 додатку.

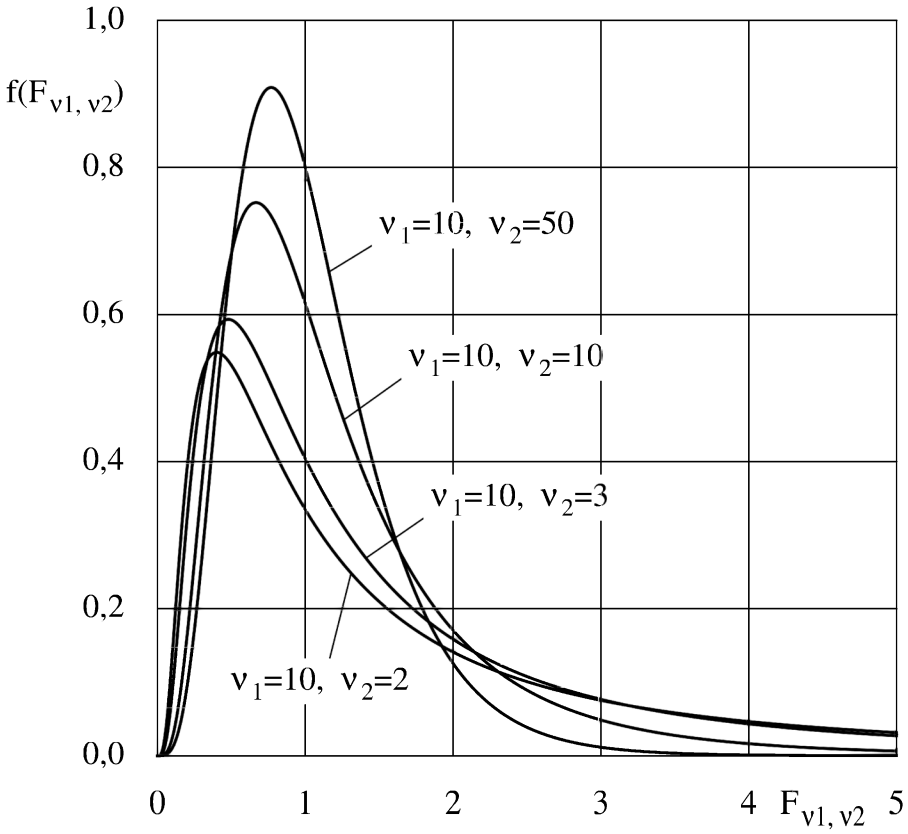


Рис. 8.5. Функція густини $f(F_{\nu_1,\nu_2})$ Фішера (8.11) залежно від значень степенів вільності ν_1 і ν_2

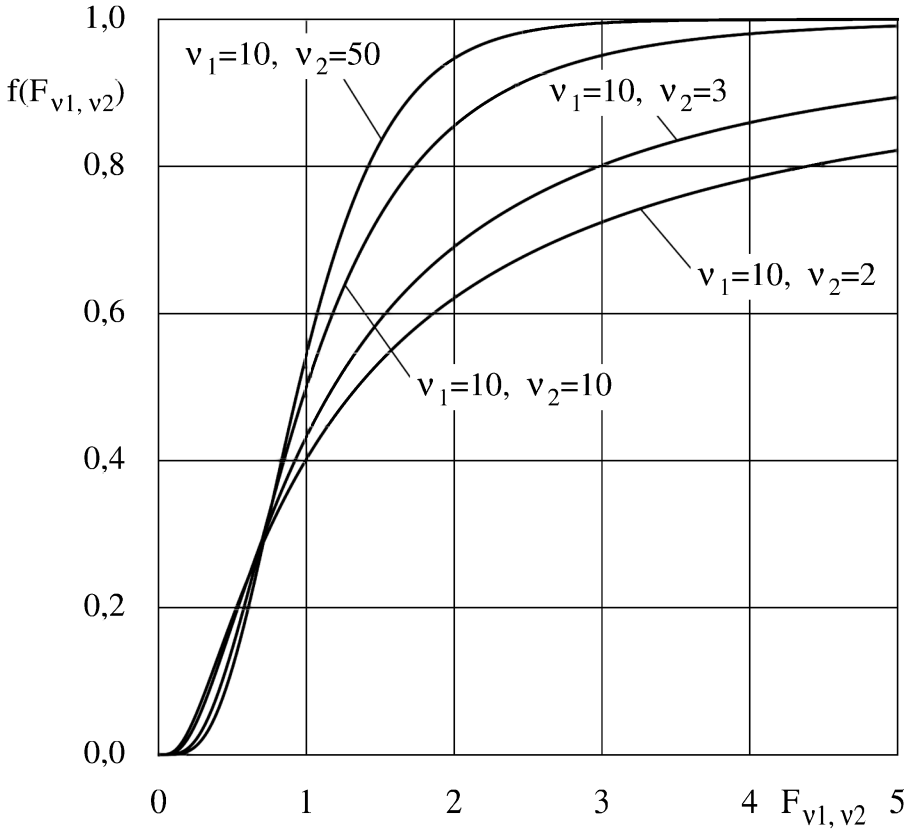


Рис. 8.6. Функція розподілу $F(F_{\nu_1,\nu_2})$ Фішера (8.11) залежно від значень степенів вільності ν_1 і ν_2

Зауваження 8.4. Необхідно розрізняти позначення функції розподілу $F(x)$ довільної випадкової величини X (див. розд. 4.8) і випадкової величини, розподіленої за законом Фішера F_{ν_1, ν_2}, останню, в свою чергу, слід відрізняти від значень, яких вона набуває. ▲

8.4. Розподіл Стьюдента *t*

Нехай $\tilde{z}$ – випадкова величина, розподілена за законом $N(0,1)$, $\tilde{\chi}_\nu^2$ – випадкова величина, розподілена за законом Пірсона із числом степенів вільності ν, тоді за визначенням випадкова величина

$$\tilde{t}_\nu = \frac{\tilde{z}}{\sqrt{\frac{\tilde{\chi}_\nu^2}{\nu}}} \tag{8.13}$$

має розподіл Стьюдента з параметром ν (розподіл Стьюдента із числом степенів вільності ν).

Математичне сподівання випадкової величини, разподіленої за законом Стьюдента $\tilde{t}_\nu$, є стала, а дисперсія виражається безпосередньо через параметр ν:

$$M[\tilde{t}_\nu] = 0 \ , \tag{8.15}$$

$$D[\tilde{t}_\nu] = \frac{\nu}{\nu - 2} \ . \tag{8.16}$$

Графіки функцій густини $f(t_\nu)$ і розподілу $F(t_\nu)$ наведені відповідно на рис. 8.7 і 8.8. За малих значень параметра ν функція $f(t_\nu)$ має довгий правий "хвіст", а за збільшення значень ν ($\nu > 30$) наближається стандартизованим нормальним розподілом.

Застосування випадкової величини $\tilde{t}_\nu$, наприклад, за „ручної" перевірки статистичних гіпотез (див. розд. 10), значно спрощується за використання таблиці квантилів $t_{\nu, q}$ (див. розд. 4.17), наведеної в табл. Д. 3 додатку.

8.5. Вибіркові розподіли

Розглянемо для неперервних випадкових величин X і Y, розподілених за нормальним законом, тобто

$$X \sim N(\mu_X, \sigma_X) \ , \tag{8.17}$$

$$Y \sim N(\mu_Y, \sigma_Y) \ , \tag{8.18}$$

різні модельні ситуації, зведені в табл. 8.9.

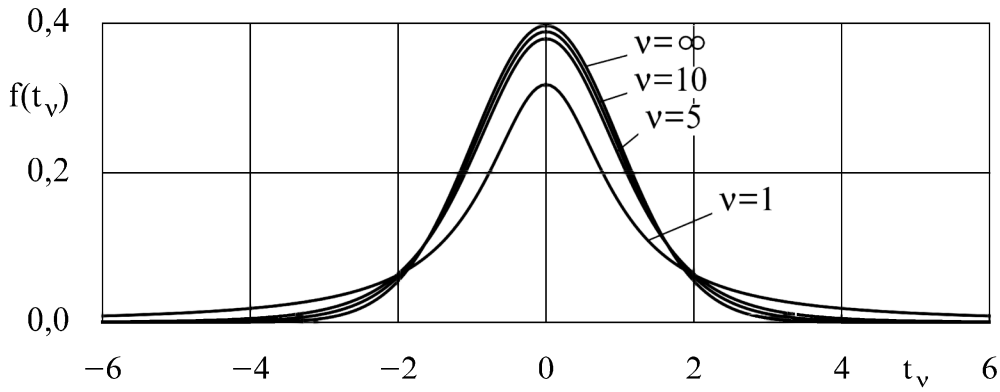


Рис. 8.7. Функція густини $f(t_\nu)$ Стьюдента (8.13) залежно від значень степенів вільності ν

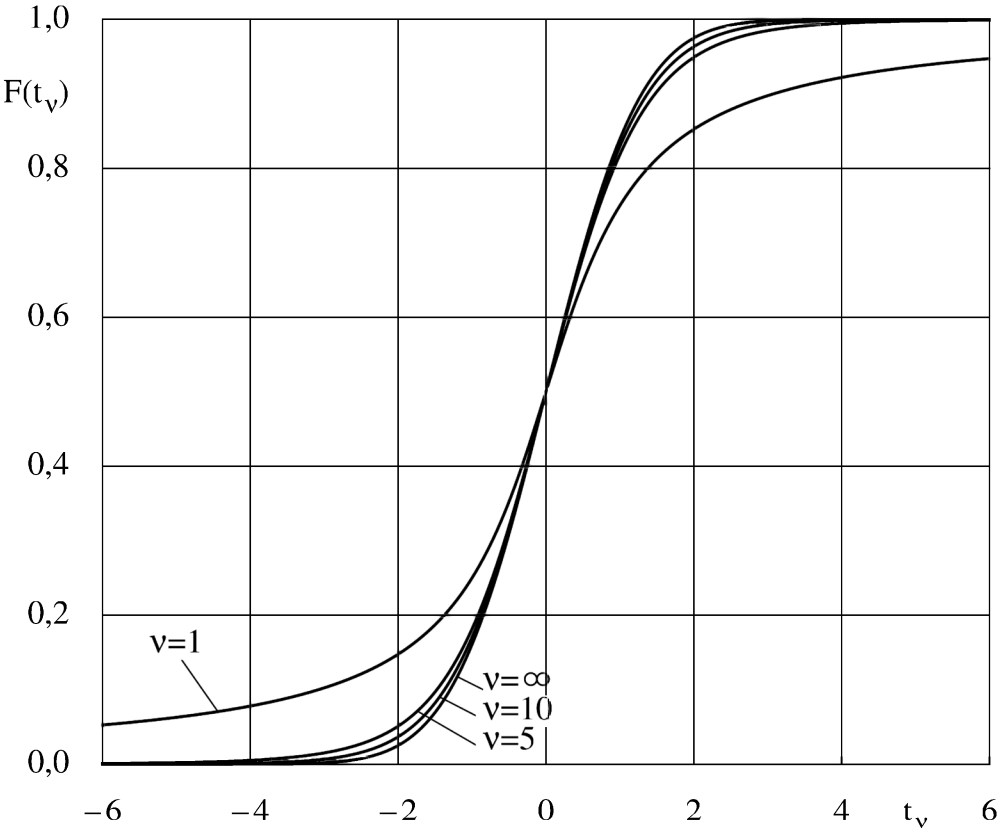


Рис. 8.8. Функція розподілу $F(t_\nu)$ Стьюдента (8.13) залежно від значень степенів вільності ν

Таблиця 8.9

Модельна ситуація	Опис модельної ситуації
1	Неперервні випадкові величини X і Y визначені на об'єктах різних генеральних сукупностей, з яких витягають вибірки обсягу n_X й n_Y відповідно. За вибірками обчислюються точкові незсунені оцінки m_X, $\hat{s}_X$, $\hat{s}_X^2$ і m_Y, $\hat{s}_Y$, $\hat{s}_Y^2$ параметрів μ_X, σ_X, σ_X^2 і μ_Y, σ_Y, σ_Y^2.
2	Неперервні випадкові величини X і Y визначені на об'єктах однієї генеральної сукупності, з якої витягається вибірка обсягу n. За вибіркою обчислюються: 1) точкові незсунені оцінки m_X, $\hat{s}_X$, $\hat{s}_X^2$ і m_Y, $\hat{s}_Y$, $\hat{s}_Y^2$ параметрів μ_X, σ_X, σ_X^2 і μ_Y, σ_Y, σ_Y^2; 2) точкові незсунені оцінки $\hat{s}_{X,Y}$, $r_{X,Y}$ параметрів $\sigma_{X,Y}$ і $\rho_{X,Y}$ зв'язку величин X і Y; 3) точкові незсунені оцінки b_0 й b_1 параметрів лінійної регресії (6.20) величини Y по X.

Вибіркові числові характеристики неперервних випадкових величин X і Y та їхнього спільного розподілу (табл. 8.8), тобто точкові незсунені оцінки відповідних параметрів неперервних випадкових величин X і Y та їхнього спільного розподілу, задовольняють універсальні закономірності (табл. 8.10).

Таблиця 8.10

№	Модельна ситуація (табл. 8.9)	Комбінація числових характеристик випадкових величин X і Y та їхнього спільного розподілу	Стандартний розподіл (див. розд. 5.3, 8.1-8.4)
1	1	$\dfrac{X-\mu_X}{\sigma_X}=\dfrac{\tilde{x}-\mu_X}{\sigma_X}$	$\tilde{z}$
2	1	$\dfrac{Y-\mu_Y}{\sigma_Y}=\dfrac{\tilde{y}-\mu_Y}{\sigma_Y}$	$\tilde{z}$
3	1	$\dfrac{\tilde{m}_X-\mu_X}{\sigma_X/\sqrt{n_X}}$	$\tilde{z}$

Продовження табл. 8.10

4	1	$\dfrac{\widetilde{m}_Y - \mu_Y}{\sigma_Y / \sqrt{n_Y}}$	$\widetilde{z}$
5	1	$\dfrac{X - \mu_X}{\widehat{s}_X} = \dfrac{\widetilde{x} - \mu_X}{\widehat{s}_X}$	$\widetilde{t}_{n_X - 1}$
6	1	$\dfrac{Y - \mu_Y}{\widehat{s}_Y} = \dfrac{\widetilde{y} - \mu_Y}{\widehat{s}_Y}$	$\widetilde{t}_{n_Y - 1}$
7	1	$\dfrac{(n_X - 1)\,\widetilde{\widehat{s}}_X^2}{\sigma_X^2}$	$\widetilde{\chi}^2_{n_X - 1}$
8	1	$\dfrac{(n_Y - 1)\,\widetilde{\widehat{s}}_Y^2}{\sigma_Y^2}$	$\widetilde{\chi}^2_{n_Y - 1}$
9	1	$\dfrac{\widetilde{\widehat{s}}_X^2}{\widetilde{\widehat{s}}_Y^2}$	$F_{n_X - 1, n_Y - 1}$
10	2	$\widetilde{y} = \dfrac{1}{2} ln \dfrac{1 + \widetilde{r}_{X,Y}}{1 - \widetilde{r}_{X,Y}}$	$N(\mu_{\widetilde{y}}, \sigma_{\widetilde{y}}), \quad n > 10$ $\mu_{\widetilde{y}} = M[\widetilde{y}] =$ $= \dfrac{1}{2} \ln \dfrac{1 + \rho_{X,Y}}{1 - \rho_{X,Y}} + \dfrac{\rho_{X,Y}}{2n - 2},$ $\sigma^2_{\widetilde{y}} = D[\widetilde{y}] \approx \dfrac{1}{n - 3}$
11	2	$\sqrt{n - 2}\,\dfrac{\widetilde{r}_{X,Y}}{\sqrt{1 - \widetilde{r}^2_{X,Y}}}$	$\widetilde{t}_{n-2}$

Закінчення табл. 8.10

12	2	$\dfrac{\tilde{b}_0 - \beta_0}{\tilde{s}_m(b_0)}$, $s_m^2(b_0) = \dfrac{s_d^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - m_X)^2}$, $s_d^2 = \dfrac{1}{n-2} \sum_{i=1}^{n} d_i^2$	$\tilde{t}_{n-2}$
13	2	$\dfrac{\tilde{b}_1 - \beta_1}{\tilde{s}_m(b_0)}$, $s_m^2(b_1) = \dfrac{s_d^2}{\sum_{i=1}^{n} (x_i - m_X)^2}$	$\tilde{t}_{n-2}$

Наведені в табл. 8.10 відповідності розподілів комбінацій числових характеристик неперервних випадкових величин X і Y та їхнього спільного розподілу стандартним розподілам $\tilde{z}$, $\tilde{t}_\nu$, F_{ν_1,ν_2} і $\tilde{\chi}_\nu^2$ називаються **вибірковими розподілами**.

Відносно вибіркових розподілів можливі, як мінімум, два протилежні погляди, які проілюструємо на прикладі вибіркового розподілу 7.

Нехай з генеральної сукупності, на якій визначена неперервна випадкова величина $X \sim N(\mu_X, \sigma_X)$, витягнута вибірка обсягу n_X, за якою обчислена статистика $\hat{s}_X^2$, а через неї – величина $(n_X - 1)\hat{s}_X^2 / \sigma_X^2$. Будемо продовжувати витягати вибірки того ж обсягу n_X й обчислювати величини $(n_X - 1)\hat{s}_X^2 / \sigma_X^2$. За великої кількості значень $(n_X - 1)\hat{s}_X^2 / \sigma_X^2$ можна побудувати гістограму вибіркової функції густини випадкової величини $(n_X - 1)\tilde{\hat{s}}_X^2 / \sigma_X^2$. Відповідно до вибіркового розподілу 7 (табл. 8.10) границя цієї гістограми буде збігатися з графіком функції

густини $f(\chi_\nu^2)$, $\nu = n_X - 1$ (див. розд. 8.1). Це складає суть першого погляду на вибіркові розподіли, наведені в табл. 8.10. Результат не буде залежати ні від того, на яких об'єктах визначена неперервна випадкова величина X, ні від її розмірності, а тільки від обсягу n_X множини вибірок. Крім того, всі вибірки мають бути організовані однаково, з найкращим наближенням до механізму випадкового вибору. Одну вибірку, за якою обчислюється тільки одне значення деякої числової характеристики, будемо називати укрупненим дослідом. Отже, для дослідного підтвердження вибіркового розподілу 7 (табл. 8.10) потрібно провести багато укрупнених дослідів. Цю ситуацію пояснює схема табл. 8.11.

Таблиця 8.11

	Вибірки обсягу n_X – стовпчики $(\nu = n_X - 1)$					
	x_1	x_1	…	x_1	…	
	x_2	x_2	…	x_2	…	
	⋮	⋮	…	⋮	…	
	x_i	x_i	…	x_i	…	
	⋮	⋮	…	⋮	…	
	x_n	x_n	…	x_n	…	
Значення, обчислені за укрупненими дослідами →	$\frac{\nu\,\hat{s}_X^2}{\sigma_X^2}$	$\frac{\nu\,\hat{s}_X^2}{\sigma_X^2}$	…	$\frac{\nu\,\hat{s}_X^2}{\sigma_X^2}$	…	Границя гістограми значень – функція розподілу $f(\chi_\nu^2)$

Другий погляд на вибіркові розподіли полягає в тому, що можна обійтися без попереднього введення випадкової величини з розподілом $\tilde{\chi}_\nu^2$, як це зроблено в розд. 8.2, визначенням же функції густини $f(\chi_\nu^2)$ можна вважати граничне положення гістограми випадкової величини $(n_X - 1)\tilde{\hat{s}}_X^2 / \sigma_X^2$, побудованої за великою кількістю значень $(n_X - 1)\hat{s}_X^2 / \sigma_X^2$, обчислених за вибірками обсягів n_X – укрупненими дослідами. У цьому випадку можна уникнути труднощів, пов'язаних з осмисленням розділів 8.1-8.3.

Зауваження 8.5. Поглибимо розгляд можливих варіантів дослідного спостереження випадкових величин, вдавшись до їхнього позначення (див. розд. 4.2).

Нехай на об'єктах генеральної сукупності практично необмеженого обсягу N визначена випадкова величина $X \sim N(\mu_X, \sigma_X)$. Із генеральної сукупності витягнемо вибірку $\{x_1, x_2, \ldots, x_n\} = \{x_i, \quad i = 1, \ldots, n\}$ обсягу n. За вибірковими значеннями можуть бути обчислені вибіркові числові характеристики m_X і $\hat{s}_X^2$ і визначені вибіркова функція густини $\{f_k^*, \quad k = 1, \ldots, K\}$ і вибіркова функція розподілу $\{F_k^*, \quad k = 1, \ldots, K\}$ (див. розд. 4.8, 4.10). Якщо збільшувати обсяг вибірки, то у граничному випадку $n \to N$ вибіркові функція густини і функція розподілу і вибіркові числові характеристики будуть наближатися до (генеральних) функції густини $f(x; \mu_X, \sigma_X)$ і функції розподілу $F(x; \mu_X, \sigma_X)$ і генеральних числових характеристик-параметрів μ_X і σ_X^2. Існує чимало можливостей для витягнення репрезентативних вибірок (див. розд. 4.5), за збільшення обсягу яких буде спостерігатися збіг наближених і точкових описів випадкової величини (див. розд. 4.11). Такі вибірки і відповідні їм граничні наближення до випадкової величини X показані в табл. 8.12 як вибірки-стовпчики і випадкові величини-стовпчики.

Таблиця 8.12

	Вибірки-стовпчики					Випадкові величини-рядки
Вибірки-рядки	x_1	x_1	...	x_1	...	$\tilde{x}_1 \sim N(\mu_X, \sigma_X)$
	x_2	x_2	...	x_2	...	$\tilde{x}_2 \sim N(\mu_X, \sigma_X)$
	$\vdots$	$\vdots$	...	$\vdots$	...	...
	x_i	x_i	...	x_i	...	$\tilde{x}_i \sim N(\mu_X, \sigma_X)$
	$\vdots$	$\vdots$	...	$\vdots$	...	...
	x_n	x_n	...	x_n	...	$x_n \sim N(\mu_X, \sigma_X)$
Випадкові величини-рядки	$X \sim N(\mu_X, \sigma_X)$	$X \sim N(\mu_X, \sigma_X)$	$\vdots$	$X \sim N(\mu_X, \sigma_X)$	$\vdots$	

Проте збіг наближених і повних описів випадкової величини можна спостерігати і в інший спосіб. Будемо вважати, що з декількох можливих вибірок обирані перші вибіркові значенння x_1. Оскільки ці значення позначаються однаково, то здається, що й самі значення збігаються, тобто це одне і те ж число. Але це не так. У кожній із декількох можливих вибірок перші вибіркові значення, напевно, відрізняються. Нагадаємо, що варіантів обирання першого вибіркового значення x_1 стільки ж, скільки й об'єктів генеральної сукупності, тобто N (див. розд. 2), а це досить велике число. Множина значень x_1 з різних вибірок-стовпчиків, у свою чергу, утворює вибірку-рядок (див. табл. 8.12). Отже, за першими вибірковими значеннями x_1 також можуть бути обчислені вибіркові числові характеристики m_X і $\hat{s}_X^2$ і визначені вибіркові функції густини $\{f_k^*, \quad k=1,\ldots,K\}$ і вибіркові функції розподілу $\{F_k^*, \quad k=1,\ldots,K\}$. Якщо збільшувати кількість обираних значень x_1, тобто збільшувати обсяг вибірки-рядка, то можна наблизити вибірковими функціями густини і розподілу генеральні функції густини $f(x;\mu_X,\sigma_X)$ і розподілу $F(x;\mu_X,\sigma_X)$, а вибірковими числовими характеристиками m_X і $\hat{s}_X^2$ – генеральні числові характеристики μ_X і σ_X^2. Зрозуміло, що можна утворити вибірки-рядки з будь-яких вибіркових значень і спостерігати збіг наближених і точних описів випадкової величини X (див. табл. 8.12). Спостереження випадкових величин під таким кутом зору буде використано у подальшому викладенні, зокрема під час обґрунтування деяких властивостей точкових оцінок параметрів (див. розд. 9.2, 9.3). ▲

8.6. Взаємозв'язки між розподілами дискретних і неперервних випадкових величин

1. Якщо в схемі досліду Бернуллі (див. розд. 5.1.) виконуються такі умови: $n \to \infty$ і $p \to 0$, причому $np = \lambda = const > 0$, то біномний розподіл наближається розподілом Пуассона (див. розд. 5.2), тобто

$$b(m;n,p) = C_n^m p^m q^{n-m} \sim P_\lambda(m) = \frac{\lambda^m}{m!} e^{-\lambda}. \tag{8.19}$$

2. Якщо дискретна випадкова величина X розподілена за біномним законом з параметрами n і p (див. розд. 5.1), то за виконання умови $n \to \infty$ стандартизований біномний розподіл наближається стандартизованим нормальним розподілом, тобто[1]

$$\frac{X-\mu}{\sigma} = \frac{X-np}{\sqrt{npq}} \sim N(0,1), \tag{8.20}$$

[1] Цей результат має назву теореми Муавра-Лапласа.

причому за значень n, таких що $\sqrt{npq} > 9$, справедлива наближена рівність, що дозволяє у процесі роботи з біномним розподілом застосовувати таблиці функції Лапласа (див. розд. 6.3):

$$P(\{x_1 \le X \le x_2\}) \approx F(z_2) - F(z_1) = \Phi(z_2) - \Phi(z_1) \ , \tag{8.21}$$

$$z_1 = \frac{x_1 - np - 0{,}5}{\sqrt{npq}}, \qquad z_2 = \frac{x_2 - np + 0{,}5}{\sqrt{npq}} \ , \tag{8.22}$$

де добавки –0,5 у чисельнику z_1 і +0,5 у чисельнику z_2 називаються поправками на неперервність.

3. Якщо дискретна випадкова величина X розподілена за законом Пуассона з параметром λ (див. розд. 6.2), то за виконання умови $\lambda \to \infty$ розподіл Пуассона наближається нормальним розподілом

$$P_\lambda(m) = \frac{\lambda^m}{m!} e^{-1} \sim N(\lambda, \sqrt{\lambda}) \ , \tag{8.23}$$

причому за значень $\lambda > 9$ справедлива наближена рівність, що дозволяє, працюючи з розподілом Пуассона, застосовувати таблиці функції Лапласа (див. розд. 6.3):

$$P(\{x_1 \le X \le x_2\}) \approx F(z_2) - F(z_1) = \Phi(z_2) - \Phi(z_1), \tag{8.24}$$

$$z_1 = \frac{x_1 - \lambda - 0{,}5}{\sqrt{\lambda}}, \qquad z_1 = \frac{x_2 - \lambda + 0{,}5}{\sqrt{\lambda}}, \tag{8.25}$$

де добавки –0,5 у чисельнику z_1 і +0,5 у чисельнику z_2 називаються поправками на неперервність.

4. Якщо дискретна випадкова величина X розподілена за біномним законом з параметрами n і p, то вона пов'язана з розподілом Фішера (див. розд. 8.2) співвідношеннями

$$P(\{X \le m\}) = 1 - P\left(\left\{\frac{(m+1)\,F_{\nu_1,\nu_2}}{(n-m)+(m+1)\,F_{\nu_1,\nu_2}} \le p\right\}\right), \tag{8.26}$$

$$P(\{X \ge m\}) = 1 - P\left(\left\{\frac{m}{m+(n-m+1)\,F_{\nu_3,\nu_4}} \le p\right\}\right), \tag{8.27}$$

де

$$\nu_1 = 2(m+1), \quad \nu_2 = 2(n-m), \quad \nu_3 = 2(n-m+1), \quad \nu_4 = 2m.$$

5. Якщо дискретна випадкова величина X розподілена за законом Пуассона з параметром λ, то вона пов'язана з розподілом Пірсона (див. розд. 8.1) співвідношеннями

$$P(\{X \le m\}) = 1 - P(\{\chi^2_{\nu_1} \le 2\lambda\}), \tag{8.28}$$

$$P(\{X \ge m\}) = 1 - P(\{\chi^2_{\nu_1} \ge 2\lambda\}), \tag{8.29}$$

де

$$\nu_1 = 2(m+1), \qquad \nu_2 = 2m.$$

6. Якщо неперервна випадкова величина X розподілена за законом $\tilde{\chi}^2_\nu$, то за виконання умови $\nu \to \infty$ розподіл X наближається нормальним розподілом

$$X \sim N(\nu, \sqrt{2\nu}), \tag{8.30}$$

причому ця апроксимація може застосовуватися за значень $\nu > 100$. Більш точна оцінка має місце для перетвореної величини

$$\sqrt{2\tilde{\chi}^2_\nu} \sim N(\sqrt{2\nu-1}, 1) \tag{8.31}$$

або

$$\left(\sqrt{2\tilde{\chi}^2_\nu} - \sqrt{2\nu-1}\right) \sim N(0,1). \tag{8.32}$$

7. Якщо неперервна випадкова величина X розподілена за законом $\widetilde{t}_{\nu}$, то у разі виконання умови $\nu \to \infty$ розподіл X наближається стандартизованим нормальним розподілом

$$X \sim N(0,1), \tag{8.24}$$

причому ця апроксимація може застосовуватися за значень $\nu > 30$.

Основні взаємозв'язки між різними розподілами показані на рис. 8.9.

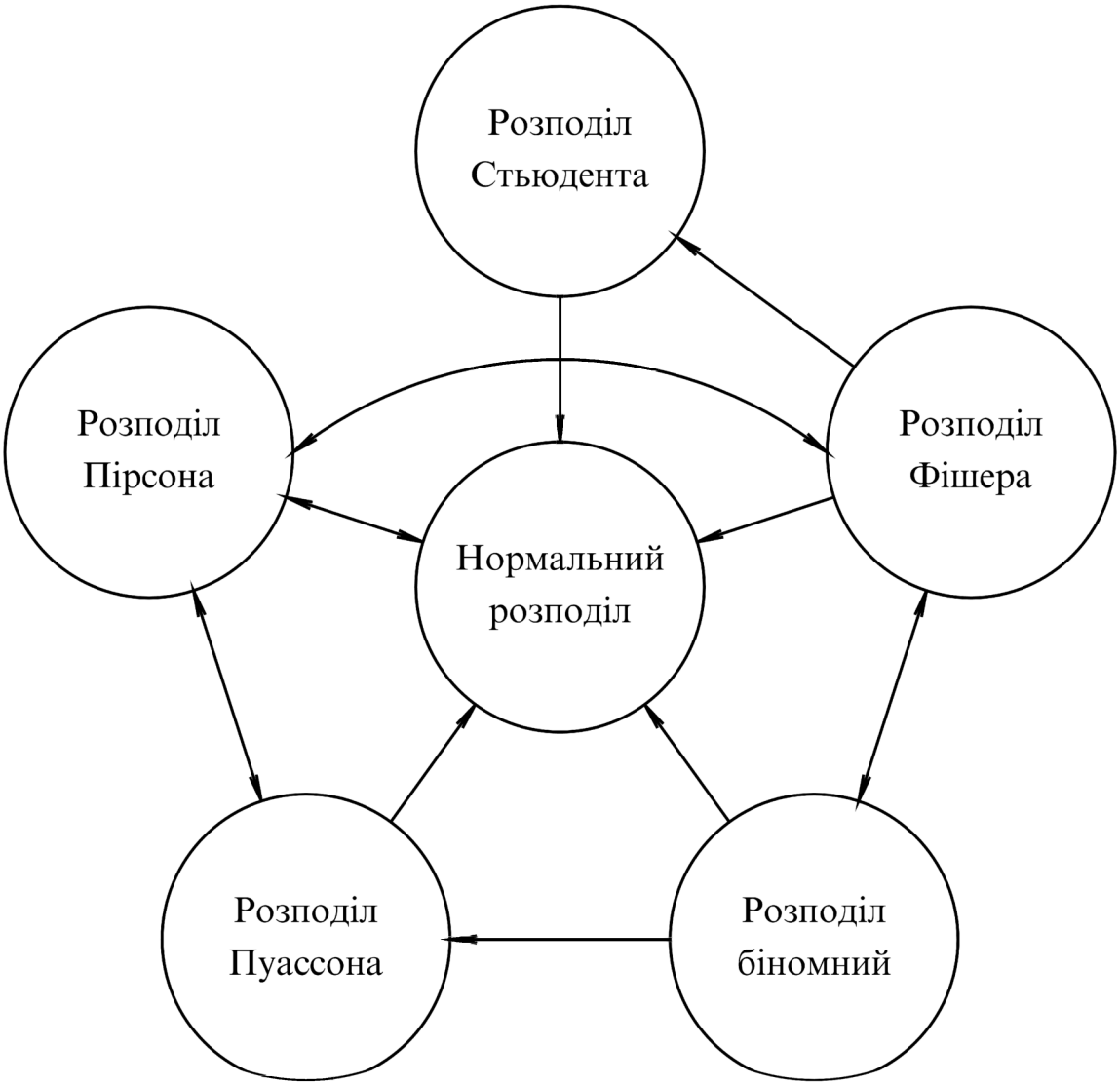


Рис. 8.9. Взаємні зв’язки та граничні переходи між стандартними розподілами дискретних і неперевних випадкових величин

9. Оцінювання параметрів випадкових величин

9.1. Точкове оцінювання параметрів неперервної випадкової величини

Розглянемо генеральну сукупність, що вивчається за певною числовою ознакою – неперервною випадковою величиною X. Нехай γ – деякий параметр величини X. Якщо для вивчення даної генеральної сукупності застосовується вибірковий метод, то за кожною з вибірок обсягу n може бути обчислена точкова оцінка (наближене значення) параметра γ – статистика g (див. розд. 4.14). Оскільки склад вибірки випадковий, то для даної вибірки статистика g візьме заздалегідь невідоме числове значення, а на множині будь-яких вибірок статистика g є випадкова величина (див. розд. 8.5):

$$\tilde{g} = \tilde{g}\left[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i, \ldots, \tilde{x}_n\right] \tag{9.1}$$

Опис будь-якої випадкової величини (див. розд. 4.16) може бути заданим законом розподілу (повним описом) і (або) числовими характеристиками (неповним описом), отже, і точкову оцінку параметра, як випадкову величину, можна так само описувати своїм законом розподілу і (або) своїми числовими характеристиками (наприклад, функцією густини $f(g)$, математичним сподіванням $\mu_{\tilde{g}} = M\left[\tilde{g}\right]$ і дисперсією $\sigma^2_{\tilde{g}} = D\left[\tilde{g}\right]$).

Основна задача точкового оцінювання – вибір способу обчислення за вибіркою значення статистики $\tilde{g}$ (9.1), що дозволяє одержати придатне наближення оцінюваного параметра γ. Отже, в основній задачі точкового оцінювання мова йде про неповний опис досліджуваної випадкової величини.

Основними властивостями точкових оцінок є **незсуненість** і **обґрунтованість**.

Оцінка $\tilde{g}$ параметра γ називається **незсуненою**, якщо її математичне сподівання дорівнює оцінюваному параметру γ, тобто

$$M\left[\tilde{g}\right] = \gamma. \tag{9.2}$$

Якщо оцінка $\tilde{g}$ параметра γ зсунена, то вона може або завищувати значення γ (тобто $M\left[\tilde{g}\right] > \gamma$), або занижувати його (тобто $M\left[\tilde{g}\right] < \gamma$). В обох випадках це призводить до систематичних (одного знака) помилок в оцінці параметра γ. Властивість незсуненості гарантує відсутність систематичних помилок в оцінці параметра.

Оцінка $\tilde{g}$ параметра γ називається **обґрунтованою**, якщо за досить великого числа n незалежних вибіркових значень $\{x_i,\ i = 1, \ldots, n\}$ невідхилення значення статистики $\tilde{g}$ від значення параметра γ на задану, як завгодно малу, величину ε,

є практично вірогідна подія (див. розд. 3.3). Строге формулювання цієї властивості полягає в тому, що ймовірність того, що різниця між $\widetilde{g}$ і γ за абсолютною величиною менша як завгодно малого додатного числа ε, близька до одиниці:

$$\lim_{n\to\infty} P\left(\left\{\left|\widetilde{g}-\gamma\right|<\varepsilon\right\}\right)=1. \tag{9.3}$$

Отже, у випадку використання обгрунтованих оцінок виправдується збільшення обсягу вибірки n, тому що можливість значної помилки в оцінці невідомого параметра стає все менш ймовірною.

9.2. Властивості вибіркового середнього значення як точкової оцінки математичного сподівання

1. Вибіркове середнє значення m, обчислене за n незалежними вибірковими значеннями випадкової величини X з математичним сподіванням μ, **є незсунена оцінка** параметра μ.

Для доведення візьмемо n незалежних вибіркових значень $\{x_i,\ i=1,\ldots,n\}$ випадкової величини X. Оскільки X має математичне сподівання μ, це означає, що $M[\widetilde{x}_i]=\mu,\ i=1,\ldots,n$. Із визначень вибіркового середнього значення й математичного сподівання (див. розд. 4.11 і 4.13) і властивості математичного сподівання суми випадкових величин (див. розд. 7.4 і 7.6) випливає, що

$$M[\widetilde{m}]=M\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i\right]=\frac{1}{n}\sum_{i=1}^{n}M[\widetilde{x}_i]=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu. \ \blacksquare \tag{9.4}$$

2. Вибіркове середнє значення m, обчислене за n незалежними вибірковими значеннями випадкової величини X з математичним сподіванням μ і дисперсією σ^2, є **обгрунтована оцінка** параметра μ.

9.3. Властивості вибіркової дисперсії як точкової оцінки дисперсії

1. Вибіркова дисперсія s^2, обчислена за n незалежними вибірковими значеннями випадкової величини X з математичним сподіванням μ і дисперсією σ^2 є **зсунена оцінка** параметра σ^2 (**зсунена вибіркова дисперсія**).

Дійсно, за умовою

$$M[\widetilde{x}_1]=M[\widetilde{x}_2]=\ldots=M[\widetilde{x}_i]=\ldots=M[\widetilde{x}_n]=\mu, \tag{9.5}$$

$$D[\widetilde{x}_1]=D[\widetilde{x}_1]=\ldots=D[\widetilde{x}_i]=\ldots=D[\widetilde{x}_n]=\sigma^2. \tag{9.6}$$

Перетворимо вираз для вибіркової дисперсії (4.7):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i \pm \mu - m)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \mu) - (m - \mu)\right]^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n}(m - \mu)\sum_{i=1}^{n}(x_i - \mu) + \frac{n}{n}(m - \mu)^2. \qquad (9.7)$$

Беручи до уваги, що з визначення вибіркового середнього значення (4.5) випливає, що $m \cdot n = \sum_{i=1}^{n} x_i$, спростимо другий доданок $\frac{2}{n}(m - \mu)\sum_{i=1}^{n}(x_i - \mu)$ у виразі (9.7):

$$\frac{2 \cdot (m - \mu)\sum_{i=1}^{n}(x_i - \mu)}{n} = \frac{2 \cdot (m - \mu) \cdot \left(\sum_{i=1}^{n} x_i - n \cdot \mu\right)}{n} =$$

$$= \frac{2 \cdot (m - \mu) \cdot (n \cdot m - n \cdot \mu)}{n} = 2 \cdot (m - \mu)^2. \qquad (9.8)$$

Одержимо такий вираз для вибіркової дисперсії:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - 2 \cdot (m - \mu)^2 + (m - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (m - \mu)^2. \qquad (9.9)$$

Відтак розглянемо математичне сподівання вибіркової дисперсії (9.9):

$$M[\tilde{s}^2] = M\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mu)^2 - (m - \mu)^2\right] = M\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \mu)^2\right] - M[(\tilde{m} - \mu)^2] =$$

$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2, \qquad (9.10)$$

тобто s^2 є зсунена оцінка дисперсії. ■

Скориставшись властивістю 3 математичного сподівання (див. розд. 4.15), легко визначити незсунену оцінку для дисперсії. Для цього досить помножити s^2 на дріб $n/(n-1)$, що називається поправкою **Бесселя**[1]:

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 . \qquad (9.11)$$

Дійсно,

$$M[\tilde{\hat{s}}^2] = M\left[\frac{n}{n-1}\tilde{\hat{s}}^2\right] = \frac{n}{n-1} M[\tilde{\hat{s}}^2] = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2 . \qquad (9.12)$$

Оцінка $\hat{s}^2$, обчислена за формулою (9.11), називається **незсуненою вибірковою дисперсією (**у розд. 4.16 вираз для неї наводився без пояснення). Розбіжність між зсуненою вибірковою дисперсією s^2 (4.7) і незсуненою вибірковою дисперсією $\hat{s}^2$ (9.11) помітна за малих обсягів вибірки. У разі n>50 практично немає різниці між оцінками s^2 й $\hat{s}^2$. Щоб уникнути сумнівів щодо суттєвої різниці між ними, краще завжди використовувати незсунену вибіркову дисперсію $\hat{s}^2$.

Величина, що дорівнює додатному квадратному кореню з незсуненої вибіркової дисперсії $\hat{s}^2$ (9.11), називається **незсуненим середньоквадратичним (середнім квадратичним) відхиленням**

$$\hat{s} = \sqrt{\hat{s}^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2} . \qquad (9.13)$$

Іншими словами, це є незсунена оцінка середньоквадратичного відхилення σ. Раніше уведена величина вибіркового середньоквадратичного відхилення s (4.8) є зсунена оцінка σ.

2. Вибіркові дисперсії s^2 й $\hat{s}^2$, обчислені за n незалежними вибірковими значеннями випадкової величини X з математичним сподіванням μ і дисперсією σ^2, є **обгрунтованими оцінками** параметра σ^2.

Зауваження 9.1. Вибіркова коваріація $s_{X,Y}$ (7.23) є **зсуненою** оцінкою коваріації $\sigma_{X,Y}$ (7.13). Незсунена оцінка для коваріації $\sigma_{X,Y}$ обчислюється за формулою

[1] Бессель (Bessel) Фрідріх Вільгельм (1784–1846), німецький астроном і геодезист, член Берлінської АН (1812), почесний член Петербурзької АН (1814), засновник теорії обліку інструментальних та інших помилок у процесі астрономічних спостережень.

$$\hat{s}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y). \qquad (9.14)$$

Для обчислення вибіркового коефіцієнта кореляції $r_{X,Y}$ можна використовувати два набори величин – або зсунені вибіркові середньоквадратичні відхилення s_X й s_Y і коваріацію $s_{X,Y}$, або незсунені вибіркові середньоквадратичні відхилення $\hat{s}_X$ й $\hat{s}_Y$ і коваріацію $\hat{s}_{X,Y}$:

$$r_{X,Y} = \frac{s_{X,Y}}{s_X\, s_Y} = \frac{\hat{s}_{X,Y}}{\hat{s}_X\, \hat{s}_Y}. \qquad (9.15)$$

У цьому випадку точкова оцінка $r_{X,Y}$ коефіцієнта кореляції $\rho_{X,Y}$ буде незсуненою. Отже, результат обчислення $r_{X,Y}$ однаковий, як за використання зсунених, так і незсунених наборів величин s_X, s_Y, $s_{X,Y}$ і $\hat{s}_X$, $\hat{s}_Y$, $\hat{s}_{X,Y}$, тому „дашок“ над позначенням вибіркового коефіцієнта кореляції не ставиться. ▲

9.4. Інтервальне оцінювання параметрів неперервної випадкової величини

Збіг точкової оцінки $\tilde{g}$ із значенням параметра γ – малоймовірна подія. Якщо відзначити значення статистики $\tilde{g}$ на числовій осі, то можна припустити, що значення оцінюваного нею параметра γ за великого обсягу вибірки знаходиться десь поруч (рис. 9.1), але де, праворуч чи ліворуч, далеко чи близько від γ? Одержати точні відповіді на ці питання за допомогою вибіркового методу не можна. Однак у деяких задачах необхідно лише знати з певним ступенем надійності β, у якому діапазоні буде перебувати значення оцінюваного параметра.

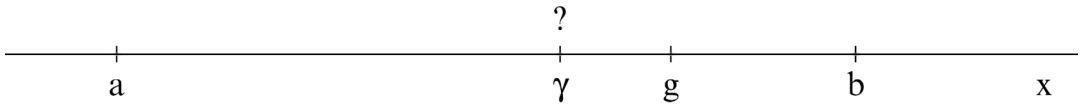


Рис. 9.1. Надійний інтервал $I_\beta(a, b)$ невідомого параметра γ містить як точкову оцінку g параметра, так, з імовірністю β, і параметр γ

Надійним інтервалом $I_\beta(a, b)$ невідомого числового параметра γ називається такий інтервал, про який можна сказати з **надійною ймовірністю** β (або **надійністю**), що він містить невідомий параметр, тобто

$$S = \{\gamma \in I_\beta(a, b)\} = \{a < \gamma < b\}, \qquad P(S) = \beta. \qquad (9.16)$$

Ймовірність протилежної події $\overline{S} = \{\gamma \notin I_\beta\}$ називається **рівнем значущості** й позначається α=1– β:

$$P(\{\gamma \le a\} + \{\gamma \ge b\}) = \alpha . \qquad (9.17)$$

Усталене значення надійності β=0,95, йому відповідає значення рівня значущості α=0,05. Докладно про вибір значень α і β говориться в розд. 9.3.

9.5. Інтервальне оцінювання математичного сподівання нормально розподіленої випадкової величини

Розглянемо побудову надійного інтервалу для математичного сподівання μ нормально розподіленої неперервної випадкової величини X.

Спочатку розберемо ситуацію, коли **середнє квадратичне відхилення σ** величини $X \sim N(\mu, \sigma)$ **відоме**. Відповідно до даних табл. 8.11 (розподіл 1) $(\widetilde{m} - \mu)/\sigma\sqrt{n} \sim \widetilde{z}$, тобто величина $(\widetilde{m} - \mu)/\sigma\sqrt{n}$ розподілена нормально з параметрами μ=0, σ=1. Оскільки розподіл величини $\widetilde{z}$ відомий, то можна визначити діапазон її найбільш і найменш імовірних значень. Подальше викладення ґрунтується на розглянутому в розд. 6.4 розв'язку оберненої задачі. Дійсно, найменш імовірні значення – це ті, які далеко відхиляються від математичного сподівання μ=0 ліворуч (убік –∞) або праворуч (убік +∞). Якщо малоймовірній події S, що полягає в значному відхиленні значень величини $\widetilde{z}$ від значення μ=0, „виділити“ ймовірність α, то ймовірність 1–α, що залишиться від вірогідної події, буде нарівно розподілена між відхиленням ліворуч і праворуч. Це легко зробити за допомогою квантилів $z_{\alpha/2}$ і $z_{1-\alpha/2}$ (див. розд. 4.17 і 6.4). Отже, діапазон найбільш імовірних значень неперервної випадкової величини $\widetilde{z}$, що складають подію $\overline{S}$ (рис. 9.1 і 9.2), такий:

$$\overline{S} = \{z_{\alpha/2} \le \widetilde{z} < z_{1-\alpha/2}\}, \qquad P(\overline{S}) = 1 - \alpha = \beta, \qquad (9.18)$$

а діапазон найменш імовірних значень, що складають подію S, набуде вигляду

$$S = \{\widetilde{z} < z_{\alpha/2}\} + \{\widetilde{z} \ge z_{1-\alpha/2}\}, \qquad P(S) = \alpha . \qquad (9.19)$$

Підставивши у вираз для $\overline{S}$ (9.18) замість $\widetilde{z}$ величину $(\widetilde{m} - \mu)/\sigma\sqrt{n} \sim \widetilde{z}$, одержимо подвійну нерівність:

$$z_{\alpha/2} < \frac{\widetilde{m} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}, \qquad (9.20)$$

яку можна розв'язати відносно невідомого параметра μ:

$$m + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < m + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}. \qquad (9.21)$$

Якщо в останньому виразі скористатися симетричністю функції густини $f(z)$ (див. розд. 6.3), тобто тим, що $z_{\alpha/2} = z_{1-\alpha/2}$, то матимемо надійний інтервал для математичного сподівання величини $X \sim N(\mu, \sigma)$:

$$m - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < \mu < m + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}. \qquad (9.22)$$

Надійний інтервал (9.22) зображений на рис. 9.3.

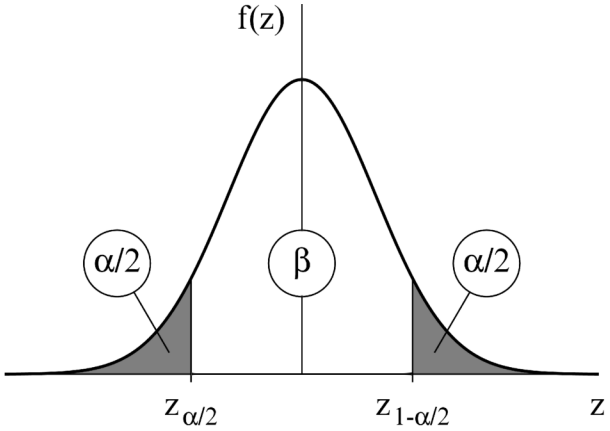


Рис. 9.2. Діапазон значень, яких випадкова величина $N(0,1)$ набуває з імовірністю $\beta = 1 - \alpha$

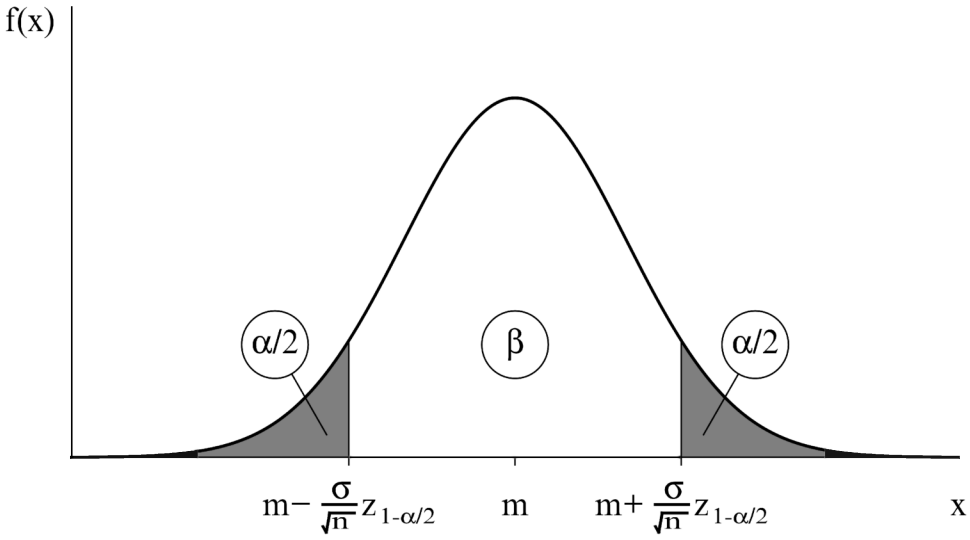


Рис. 9.3. Надійний інтервал для математичного сподівання μ випадкової величини $N(\mu, \sigma)$ за відомого значення σ

Якщо **середнє квадратичне відхилення** σ величини $X \sim N(\mu, \sigma)$ **невідоме**, то скористаємося тим, що $(\widetilde{m} - \mu)/\widehat{s}\sqrt{n} \sim \widetilde{t}_{n-1}$, тобто величина $(\widetilde{m} - \mu)/\widehat{s}\sqrt{n}$ розподілена за законом Стьюдента із числом степенів вільності $\nu = n - 1$ (див. табл. 8.11, розподіл 2). Оскільки розподіл неперервної випадкової величини $\widetilde{t}_\nu$, як і величини $\widetilde{z}$, симетричний (рис. 9.4), то надійний інтервал для математичного сподівання неперервної випадкової величини X можна відразу ж виписати за аналогією з раніше одержаним виразом (9.22):

$$m - \frac{\widehat{s}}{\sqrt{n}} t_{n-1,1-\alpha/2} < \mu < m + \frac{\widehat{s}}{\sqrt{n}} t_{n-1,1-\alpha/2}. \qquad (9.23)$$

Надійний інтервал (9.23) зображений на рис. 9.5.

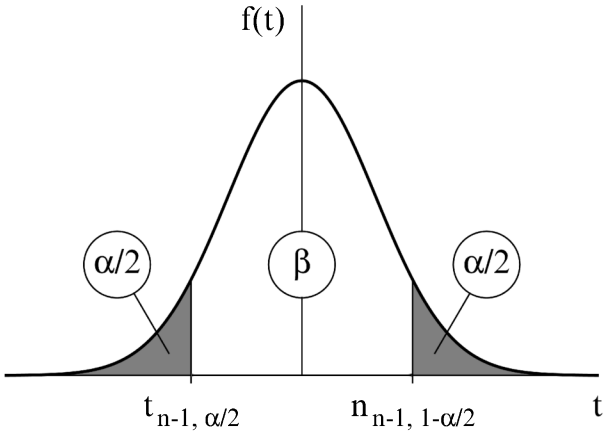


Рис. 9.4. Діапазон значень, яких випадкова величина $\widetilde{t}_{n-1}$ набуває з імовірністю $\beta = 1 - \alpha$

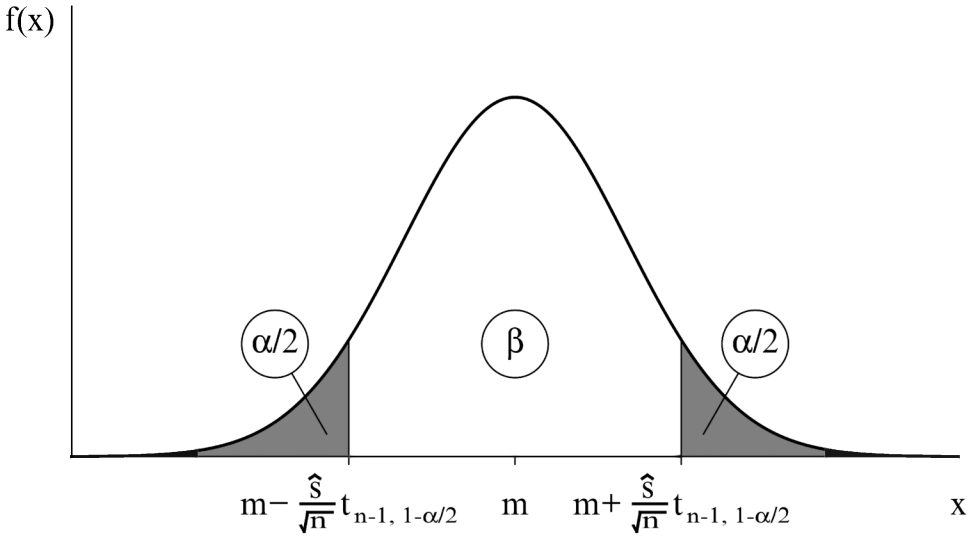


Рис. 9.5. Надійний інтервал для математичного сподівання μ випадкової величини $N(\mu, \sigma)$ за невідомого значення σ

Величина $\hat{s}/\sqrt{n}$ у виразі (9.23) називається **стандартним відхиленням** (**помилкою середнього**) і позначається s_m. Результати обробки експериментів відносно визначення невідомого значення математичного сподівання (генерального середнього значення) μ часто зображуються у вигляді $m \pm s_m$, відтак за відомого обсягу вибірки n та заданої величини рівня значущості α (або надійної ймовірності β) можна обчислити надійний інтервал за виразом (9.22) або (9.23).

Побудова надійного інтервалу (9.22) або (9.23) для математичного сподівання **не чутлива** до помірних відхилень розподілу неперервної випадкової величини X від нормального закону.

9.6. Інтервальне оцінювання дисперсії нормально розподіленої випадкової величини

Відповідно до даних табл. 8.11 $(n-1)\tilde{\hat{s}}^2/\sigma^2 \sim \tilde{\chi}^2_{\nu=n-1}$ (вибірковий розподіл 4), тобто величина $(n-1)\tilde{\hat{s}}^2/\sigma^2$ розподілена за законом Пірсона із числом степенів вільності $\nu = n-1$, тому можна обчислити діапазон її найбільш і найменш імовірних значень, виходячи зі змісту незсуненої вибіркової дисперсії $\hat{s}^2$. Якщо у вибірці виявляються значення, що далеко відстоять від вибіркового середнього значення (це один варіант малоймовірної події S), то значення незсуненої вибіркової дисперсії $\hat{s}^2$ буде велике. Якщо вибірка виявляється складеною із приблизно однакових значень (це інший варіант малоймовірної події S), то величина $\hat{s}^2$ буде близька до нуля. Отже, малоймовірна подія S буде полягати в набутті незсуненою вибірковою дисперсією $\hat{s}^2$ значень, близьких до 0 або ∞. Ймовірність α малоймовірної події S розподілимо нарівно між її варіантами (рис. 9.6):

$$P(\{\tilde{\chi}^2_\nu < \tilde{\chi}^2_{\nu,\,\alpha/2}\}) = \alpha/2, \qquad P(\{\tilde{\chi}^2_\nu > \tilde{\chi}^2_{\nu,\,1-\alpha/2}\}) = \alpha/2, \tag{9.24}$$

звідки

$$\chi^2_{n-1,\,\alpha/2} < \frac{(n-1)\tilde{\hat{s}}^2}{\sigma^2} < \chi^2_{n-1,\,1-\alpha/2}. \tag{9.25}$$

Розподіл неперервної випадкової величини $\tilde{\chi}^2_\nu$ є несиметричний, тому межі надійного інтервалу симетричні тільки за ймовірностями – площами, що відтинаються на “хвостах” розподілу квантилями $\chi^2_{\nu,\alpha/2}$, $\chi^2_{\nu,1-\alpha/2}$ (рис. 9.6). Розв’язавши нерівність (9.25) відносно невідомого параметра σ^2 величини $X \sim N(\mu,\sigma)$, одержимо шуканий надійний інтервал:

$$\frac{\hat{s}^2(n-1)}{\chi^2_{n-1,\,1-\alpha/2}} < \sigma^2 < \frac{\hat{s}^2(n-1)}{\chi^2_{n-1,\,\alpha/2}}. \qquad (9.26)$$

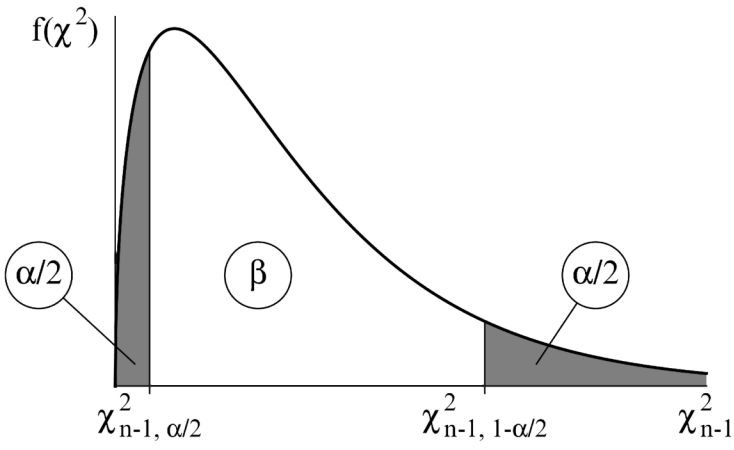


Рис. 9.6. Діапазон значень, яких випадкова величина $\tilde{\chi}^2_{n-1}$ набуває з імовірністю $\beta = 1 - \alpha$

Побудова надійного інтервалу (9.26) для дисперсії σ^2 є **дуже чутлива** до відхилень розподілу неперервної випадкової величини X від нормального закону.

Зауваження 9.2. Порівнявши формули (9.22), (9.23) і (9.26) для надійних інтервалів математичного сподівання μ і дисперсії σ^2, можна помітити, що ширина інтервалів залежить від значень α і n, вибір яких належить до компетенції дослідника. Не складно здогадатися, що зі збільшенням обсягу вибірки n ширина надійних інтервалів I_β зменшується, а зі збільшенням надійної ймовірності β (зменшенням рівня значущості α=1–β) збільшується. Дійсно, чим більший обсяг вибірки, тим з більшою ймовірністю точкова оцінка виявиться ближчою до значення параметра (це випливає з визначення параметра, див. розд. 4.13, за умови, що вибірка репрезентативна). Тому за β=const з ростом обсягу вибірки n з'являється можливість зменшувати довжину інтервалу, що включає невідоме значення параметра. Якщо ж n=const, то для того щоб збільшити ймовірність влучення невідомого значення параметра в надійний інтервал I_β, треба збільшувати довжину цього інтервалу. Таким чином, збільшення надійної ймовірності β приводить до збільшення довжини I_β. Однак безмежно збільшувати надійну ймовірність немає ніякого сенсу, тому що матимемо настільки великий надійний інтервал, що його практична цінність буде зведена нанівець. ▲

9.7. Інтервальна оцінка коефіцієнта кореляції нормально розподілених випадкових величин

Для неперервних випадкових величин $X \sim N(\mu_X, \sigma_X)$ і $Y \sim N(\mu_Y, \sigma_Y)$ статистику $\tilde{r}_{X,Y}$ за допомогою перетворення Фішера (див. табл. 8.11 і рис. 9.7)

$$\widetilde{y} = \frac{1}{2}\ln\frac{1+\widetilde{r}_{X,Y}}{1-\widetilde{r}_{X,Y}} \tag{9.27}$$

можна перевести в нормально розподілену величину $\widetilde{y} \sim N(\mu_{\widetilde{y}}, \sigma_{\widetilde{y}})$ з математичним сподіванням

$$\mu_{\widetilde{y}} = \frac{1}{2}\ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} + \frac{\rho_{X,Y}}{2n-2} \tag{9.28}$$

і дисперсією

$$\sigma^2_{\widetilde{y}} = 1/(n-3), \tag{9.29}$$

що залежить тільки від обсягу вибірки.

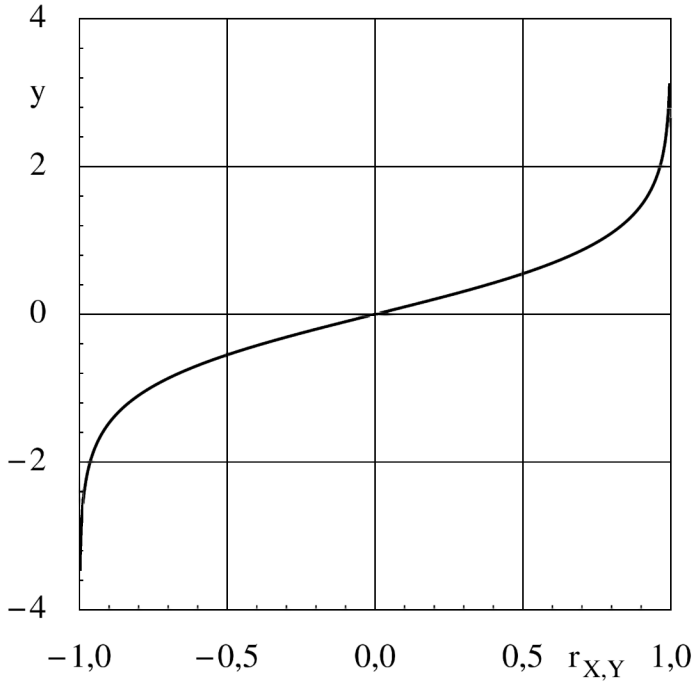


Рис. 9.7. Графічне зображення перетворення Фішера (9.27)

Поправковий член $\rho_{X,Y}/(2n-2)$ у виразі (9.28) для $\mu_{\widetilde{y}}$ є малий у порівнянні з середньоквадратичним відхиленням $\sigma_{\widetilde{y}}$, і його зазвичай нехтують, тобто

$$\mu_{\widetilde{y}} = \frac{1}{2}\ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} \ . \tag{9.30}$$

У результаті маємо ситуацію, коли потрібно побудувати надійний інтервал для математичного сподівання $\mu_{\tilde{y}}$ (9.30) величини $\tilde{y} \sim N(\mu_{\tilde{y}}, \sigma_{\tilde{y}})$, середньоквадратичне відхилення якої відоме:

$$\sigma_{\tilde{y}} = \frac{1}{\sqrt{n-3}}, \tag{9.31}$$

оскільки $\sigma_{\tilde{y}}$ залежить тільки від обсягу вибірки *n*. Ця ситуація докладно з'ясована на початку розд. 9.5.

Для побудови надійного інтервалу для математичного сподівання $\mu_{\tilde{y}}$ величини $\tilde{y}$ в подвійній нерівності (9.22) треба замінити математичне сподівання μ виразом (9.30), а середньоквадратичне відхилення σ − виразом (9.31):

$$\frac{1}{2}\ln\frac{1+r_{X,Y}}{1-r_{X,Y}} + \frac{z_{\alpha/2}}{\sqrt{n-3}} < \frac{1}{2}\ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} < \frac{1}{2}\ln\frac{1+r_{X,Y}}{1-r_{X,Y}} + \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \tag{9.32}$$

У подвійній нерівності (9.32) ліва й права частини відомі, тому для простоти подальших перетворень позначимо їх так:

$$y_{\alpha/2} = \frac{1}{2}\ln\frac{1+r_{X,Y}}{1-r_{X,Y}} + \frac{z_{\alpha/2}}{\sqrt{n-3}} = \frac{1}{2}\ln\frac{1+r_{X,Y}}{1-r_{X,Y}} - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}, \tag{9.33}$$

$$y_{1-\alpha/2} = \frac{1}{2}\ln\frac{1+r_{X,Y}}{1-r_{X,Y}} + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}. \tag{9.34}$$

У результаті одержуємо просту подвійну нерівність щодо невідомого параметра $\rho_{X,Y}$:

$$y_{\alpha/2} < \frac{1}{2}\ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} < y_{1-\alpha/2}, \tag{9.35}$$

яку, скориставшись властивостями натурального логарифма й показникової функції, перепишемо в такий спосіб:

$$2\ln e^{y_{\alpha/2}} < \ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} < 2\ln e^{y_{1-\alpha/2}}, \tag{9.36}$$

$$\ln e^{2y_{\alpha/2}} < \ln \frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} < \ln e^{2y_{1-\alpha/2}} \quad . \tag{9.37}$$

Оскільки функція логарифм є монотонна (зростаюча), то останню нерівність можна переписати відносно аргументів натурального логарифма:

$$e^{2y_{\alpha/2}} < \frac{1+\rho_{X,Y}}{1-\rho_{X,Y}} < e^{2y_{1-\alpha/2}} \quad . \tag{9.38}$$

Тепер розглянемо праву частину подвійної нерівності (9.38), розв'язавши її відносно $\rho_{X,Y}$:

$$1+\rho_{X,Y} < e^{2y_{1-\alpha/2}}(1-\rho_{X,Y}) \ , \tag{9.39}$$

$$(e^{2y_{1-\alpha/2}}+1)\rho_{X,Y} < e^{2y_{1-\alpha/2}}-1 \ , \tag{9.40}$$

$$\rho_{X,Y} < \frac{e^{2y_{1-\alpha/2}}-1}{e^{2y_{1-\alpha/2}}+1} = \frac{e^{y_{1-\alpha/2}}(e^{y_{1-\alpha/2}}-e^{-y_{1-\alpha/2}})}{e^{y_{1-\alpha/2}}(e^{y_{1-\alpha/2}}+e^{-y_{1-\alpha/2}})} = \text{th}\, y_{1-\alpha/2} \ . \tag{9.41}$$

В однобічній нерівності (9.41) використовується функція, що називається гіперболічним тангенсом

$$\text{th}\, y = \frac{e^{y}-e^{-y}}{e^{y}+e^{-y}} , \tag{9.42}$$

графік якої зображений на рис. 9.8.

Аналогічно із правої частини подвійної нерівності (9.38) маємо нерівність

$$\text{th}\, y_{\alpha/2} < \rho_{X,Y} \ . \tag{9.43}$$

Поєднуючи однобічні нерівності (9.41) і (9.43), остаточно одержуємо подвійну нерівність

$$\text{th}\, y_{\alpha/2} < \rho_{X,Y} < \text{th}\, y_{1-\alpha/2} , \tag{9.44}$$

яка є шуканий надійний інтервал для коефіцієнта кореляції випадкових величин $X \sim N(\mu_X, \sigma_X)$ і $Y \sim N(\mu_Y, \sigma_Y)$. Межі надійного інтервалу визначені відповідно виразами (9.33) і (9.34).

Зауваження 9.3. Варто звернути увагу на те, що через нелінійність функції гіперболічний тангенс $\operatorname{th} y$ (рис. 9.8) надійний інтервал для коефіцієнта кореляції $\rho_{X,Y}$ є несиметричний щодо вибіркового коефіцієнта кореляції $r_{X,Y}$. ▲

9.8. Інтервальна оцінка параметрів лінійної регресії

Побудова надійних інтервалів для параметрів β_0 і β_1 лінійної регресії (див. розд. 7.6) ґрунтується на статистиках (див. табл. 8.11, вибіркові розподіли 12, 13)

$$\frac{\tilde{b}_0 - \beta_0}{\tilde{S}_m(\tilde{b}_0)} \sim \tilde{t}_{n-1}, \tag{9.45}$$

$$\frac{\tilde{b}_1 - \beta_1}{\tilde{S}_m(\tilde{b}_1)} \sim \tilde{t}_{n-1}, \tag{9.46}$$

розподілених за законом Стьюдента $\tilde{t}_\nu$ із числом степенів вільності $\nu = n-2$, де $s_m(b_0)$, $s_m(b_1)$ – **стандартні помилки** b_0 й b_1:

$$s_m(b_0) = \sqrt{\frac{\hat{s}_d^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - m_X)^2}} = \sqrt{\frac{\hat{s}_d^2 ((n-1)\hat{s}_X^2 + n m_X^2)}{n(n-1)\hat{s}_X^2}}, \tag{9.47}$$

$$s_m(b_1) = \sqrt{\frac{\hat{s}_d^2}{\sum_{i=1}^{n} (x_i - m_X)^2}} = \sqrt{\frac{\hat{s}_d^2}{(n-1)\hat{s}_X^2}}. \tag{9.48}$$

Отже, для рівня значущості α інтервальні оцінки параметрів β_0 і β_1 можна побудувати в такий спосіб (доцільно порівняти одержані вирази з інтервальною оцінкою для математичного сподівання (9.25), (9.26)):

$$b_0 - s_m(b_0) \cdot t_{n-2,1-\alpha/2} < \beta_0 < b_0 + s_m(b_0) \cdot t_{n-2,1-\alpha/2}, \tag{9.49}$$

$$b_1 - s_m(b_1) \cdot t_{n-2,1-\alpha/2} < \beta_1 < b_1 + s_m(b_1) \cdot t_{n-2,1-\alpha/2}. \tag{9.50}$$

9.9. Інтервальне оцінювання параметрів дискретної випадкової величини

Усі одержані в розд. 9.5 – 9.8 інтервальні оцінки параметрів застосовуються тільки для неперервних випадкових величин. Якщо скористатися зв'язками між неперервними випадковими величинами й дискретними випадковими величинами (див. розд. 8.6), то на основі інтервальних оцінок для математичного сподівання й дисперсії неперервної випадкової величини можна одержати вираз для інтервальних оцінок відповідних параметрів дискретної випадкової величини (про коефіцієнт кореляції й параметри лінійної регресії мова, зрозуміло, не йде, тому що вони не визначалися для дискретної випадкової величини).

Приклад 9.1 [Афіфі, 1982]. В 1971 – 1974 рр. в окрузі Лос-Анджелес проводилося обстеження підлітків у віці 12 – 14 років для виявлення випадків захворювання на сколіоз (скривлення хребта). Усього було обстежено 3492 дітей, і сколіоз був виявлений у 474 з них. За результатами попередніх обстежень частка підлітків, у яких було це захворювання, перебувала в межах 0,02 − 0,1 і передбачалося, що на момент обстеження ця частка зросте не менше ніж на 10%. Необхідно з'ясувати, чи підтверджують результати останнього обстеження це припущення.

За умовою ми маємо справу з біноміальним розподілом (див. розд. 5.1). Уведемо подію A={виявлення сколіозу у підлітка}, число появ якої в серії n=3492 дослідів дорівнює m=474.

Якби параметр p біноміального розподілу був відомий, то за виразом (8.14) можна було б для заданої надійної ймовірності β (заданого рівня значущості α) оцінити діапазон $[m_1, m_2]$ можливих появ події A у серії n=3492 дослідів. Оскільки параметр p не відомий, то з апроксимації (8.14) маємо подвійну нерівність

$$-z_{1-\alpha/2} < \frac{m - np}{\sqrt{np(1-p)}} < z_{1-\alpha/2}, \tag{9.51}$$

яку потрібно розглядати не як надійний інтервал для числа m появ події A у серії n дослідів (значення m за результатами серії n=3492 дослідів відомо), а як нерівність для знаходження інтервалу значень параметра p, за яких у серії n дослідів можливо m появ події A.

Після піднесення обох частин нерівності (9.51) до квадрата одержимо односторонню нерівність ($-z_{\alpha/2} = z_{1-\alpha/2}$, $z^2_{\alpha/2} = z^2_{1-\alpha/2}$):

$$\frac{(m - np)^2}{np(1-p)} < z^2_{1-\alpha/2}, \tag{9.52}$$

яку перетворимо в такий спосіб:

$$m^2 - 2mnp + n^2p^2 < npz^2_{1-\alpha/2} - np^2 z^2_{1-\alpha/2}, \tag{9.53}$$

$$(n^2 + n z_{1-\alpha/2}^2)\, p^2 - (2mn + n z_{1-\alpha/2}^2)\, p + m^2 < 0\,, \tag{9.54}$$

$$p^2 - \frac{2mn + n z_{1-\alpha/2}^2}{n^2 + n z_{1-\alpha/2}^2}\, p + \frac{m^2}{n^2 + n z_{1-\alpha/2}^2} < 0\,. \tag{9.55}$$

Для того щоб із квадратної нерівності (9.55) знайти межі надійного інтервалу для параметра p, необхідно розв'язати квадратне рівняння відносно p:

$$p^2 - \frac{2\dfrac{m}{n} + \dfrac{1}{n} z_{1-\alpha/2}^2}{1 + \dfrac{1}{n} z_{1-\alpha/2}^2}\, p + \frac{\dfrac{m^2}{n^2}}{1 + \dfrac{1}{n} z_{1-\alpha/2}^2} = 0\,. \tag{9.56}$$

Розв'язком квадратного рівняння (9.46) будуть такі значення:

$$p_{1,2} = \frac{\left(\dfrac{m}{n} + \dfrac{1}{n}\dfrac{z_{1-\alpha/2}^2}{2}\right) \pm \dfrac{1}{n} z_{1-\alpha/2}^2 \sqrt{\dfrac{1}{4} + \dfrac{m}{z_{1-\alpha/2}^2}\left(1 - \dfrac{m}{n}\right)}}{1 + \dfrac{1}{n} z_{1-\alpha/2}^2}\,. \tag{9.57}$$

За малого значення відношення m/n вираз (9.57) можна спростити:

$$p_{1,2} \approx \frac{m}{n} \pm \frac{\sqrt{m}}{n} z_{1-\alpha/2}^2\,. \tag{9.58}$$

Наближені значення коренів p_1 і p_2 (9.58) квадратного рівняння (9.56) визначають такий надійний інтервал параметра p:

$$\frac{m}{n} - \frac{\sqrt{m}}{n} z_{1-\alpha/2}^2 < p < \frac{m}{n} + \frac{\sqrt{m}}{n} z_{1-\alpha/2}^2\,. \tag{9.59}$$

Підставляючи в (9.59) значення $m = 474$, $n = 3492$, $\alpha = 0{,}05$, одержуємо такий надійний інтервал параметра p:

$$0{,}1248 < p < 0{,}1475, \tag{9.60}$$

що підтверджує припущення про зростання частки підлітків зі сколіозом не менше ніж на 10% у порівнянні з попереднім обстеженням.

Відзначимо, що розв'язок цієї задачі, наведений Афіфі (1982), є неправильний, тому що оцінка типу (9.51) була розглянута для відомого значення параметра p, що не відповідає дійсності. Умови застосування нерівності (9.51) обговорюються на початку даного прикладу. Ідея наведеного тут розв'язку запозичена з відомого довідника Семендяєва (1982). ●

9.10. Необхідний обсяг вибірки

Після обрання для оцінки параметра генеральної сукупності способу утворення вибірки (див. розд. 4.6) розраховуємо необхідний обсяг вибірки за умови заданої ширини надійного інтервалу й надійної ймовірності $\beta=1-\alpha$. Розглянемо задачу визначення обсягу вибірки для оцінки математичного сподівання за вибірковим середнім, припускаючи, що неперервна випадкова величина X розподілена за нормальним законом.

Нехай середньоквадратичне відхилення σ відомо. Тоді з виразу (9.22) для інтервальної оцінки математичного сподівання й заданої ширини інтервалу $2\Delta=2\,z_{\alpha/2}\sigma/\sqrt{n}$ випливає, що

$$n>\left(\frac{z_{1-\alpha/2}\sigma}{\Delta}\right)^2. \tag{9.61}$$

Якщо обчислений за (9.61) обсяг вибірки n занадто великий у порівнянні із практично наявними можливостями, то його можна зменшити, зменшивши при цьому надійну ймовірність або збільшивши надійний інтервал.

Нехай тепер середньоквадратичне відхилення σ не відоме. У цьому випадку попередньо беремо невелику пробну вибірку обсягу n_1 й обчислюємо незсунену оцінку $\hat{s}_1^2$ параметра σ^2. У результаті маємо надійний інтервал (9.23)

$$m-\frac{\hat{s}_1}{\sqrt{n_1}}t_{n_1-1,\,1-\alpha/2}<\mu<m+\frac{\hat{s}_1}{\sqrt{n_1}}t_{n_1-1,\,1-\alpha/2}. \tag{9.62}$$

За заданої ширини надійного інтервалу $2\Delta=2\,t_{n_1-1,\,\alpha/2}\,\hat{s}_1/\sqrt{n_1}$ з виразу (9.62) знаходимо оцінку шуканого обсягу вибірки

$$n>\left(\frac{t_{n_1-1,\,1-\alpha/2}\hat{s}_1}{\Delta}\right)^2. \tag{9.63}$$

Перейдемо тепер до визначення необхідного обсягу вибірки для дискретної випадкової величини. Розглянемо найпростіший дослід, у якому спостерігається дискретна випадкова величина, – підкидання монети. Уведемо подію A={поява

герба}. Число появ події A у серії з n дослідів, тобто частота $n(A) = m$, задовольняє біномний розподіл (див. розд. 5.1) з відомим параметром p=0,5, і дійсна оцінка, що випливає із (9.51), тобто

$$-z_{1-\alpha/2} < \frac{m - \frac{n}{2}}{\sqrt{n\frac{1}{2}\left(1-\frac{1}{2}\right)}} < z_{1-\alpha/2}\,. \tag{9.64}$$

Нехай очікувана частота m у серії з n дослідів дорівнює

$$m = \frac{n}{2} \pm \Delta\,, \tag{9.65}$$

де Δ – величина відхилення m від середнього значення числа дослідів $n/2$. Тоді, після підстановки (9.65) в (9.64), одержимо

$$-z_{1-\alpha/2} < \frac{2\Delta}{\sqrt{n}} < z_{1-\alpha/2}\,, \tag{9.66}$$

звідки для величини відхилення Δ маємо такий надійний інтервал:

$$-\frac{\sqrt{n}}{2} z_{1-\alpha/2} < \Delta < \frac{\sqrt{n}}{2} z_{1-\alpha/2}\,. \tag{9.67}$$

Подвійну нерівність (9.67) після піднесення до квадрата

$$\Delta^2 < \frac{n}{4} z_{1-\alpha/2}^2 \tag{9.68}$$

можна розв’зати відносно числа дослідів n, у яких спостерігається задане відхилення Δ:

$$n > \frac{4\Delta^2}{z_{1-\alpha/2}^2}\,. \tag{9.69}$$

Із (9.67) випливає, що у разі збільшення числа дослідів абсолютна величина відхилення $|\Delta|$ зростає не швидше ніж $\sqrt{n}\frac{z_{1-\alpha/2}}{2}$. Графічно це означає, що $|\Delta|$ із заданою ймовірністю $\beta = 1 - \alpha$ знаходиться в області, обмеженій знизу й зверху відповідними кривими (на рис. 9.8 рівень значущості α=0,5, $z_{1-\alpha/2} \approx 1{,}96$),

$$\Delta_{-}(n;\alpha) = -\frac{z_{1-\alpha/2}}{2}\sqrt{n}\,, \tag{9.70}$$

$$\Delta_{+}(n;\alpha) = +\frac{z_{1-\alpha/2}}{2}\sqrt{n}\,. \tag{9.71}$$

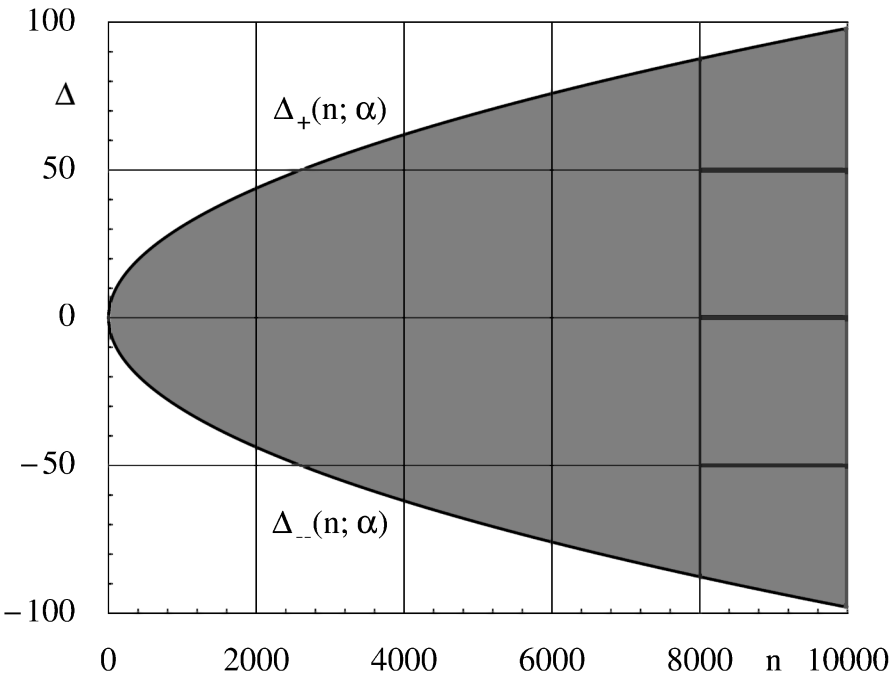


Рис. 9.8. Область, що обмежує відхилення очікуваної частоти m від середнього значення числа дослідів $n/2$, залежно від кількості дослідів n (9.65)

Із виразу (9.65) легко помітити, що величина $\frac{\Delta}{n}$ визначає абсолютну точність наближення статистичною ймовірністю m/n події A, класичної ймовірності, що дорівнює $1/2$:

$$\frac{m}{n} = \frac{1}{2} \pm \frac{\Delta}{n}\,. \tag{9.72}$$

Із (9.67) тепер випливає, що для величини $\frac{\Delta}{n}$ справедлива така подвійна нерівність:

$$-\frac{z_{1-\alpha/2}}{2\sqrt{n}} < \frac{\Delta}{n} < \frac{z_{1-\alpha/2}}{2\sqrt{n}}\,. \tag{9.73}$$

Іншими словами, із заданою ймовірністю $\beta = 1 - \alpha$ значення $\frac{\Delta}{n}$ знаходяться в області, обмеженій знизу й зверху кривими (див. рис. 9.9):

$$\delta_{-}(n;\alpha) = -\frac{z_{1-\alpha/2}}{2\sqrt{n}}, \tag{9.74}$$

$$\delta_{+}(n;\alpha) = +\frac{z_{1-\alpha/2}}{2\sqrt{n}}. \tag{9.75}$$

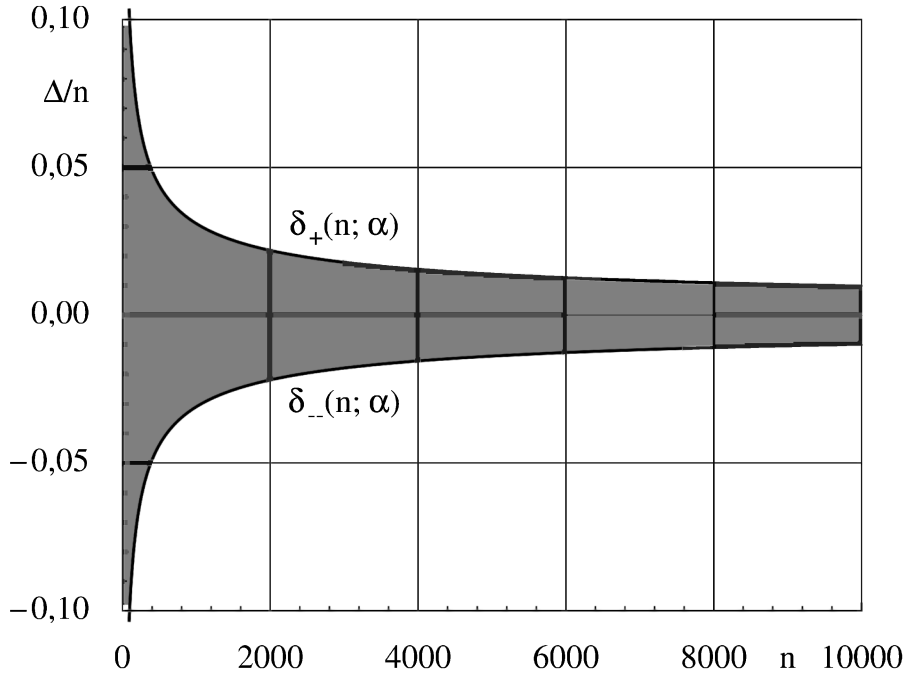


Рис. 9.9. Область, що обмежує відхилення статистичної імовірності m/n від математичного сподівання 0,5, залежно від кількості дослідів n (9.72)

Якщо задана точність $\left|\frac{\Delta}{n}\right| = \delta$ для виразу (9.72), тобто

$$\left|\frac{\Delta}{n}\right| < \frac{z_{1-\alpha/2}}{2\sqrt{n}} < \delta, \tag{9.76}$$

то легко визначити умову досягнення цієї точності:

$$n > \frac{z_{1-\alpha/2}^{2}}{4\delta^{2}}. \tag{9.77}$$

10. Перевірка статистичних гіпотез

10.1. Класифікація гіпотез

Складовими частинами будь-якого науково-дослідного процесу є:

1) формулювання проблеми;
2) постановка досліду (експерименту, спостережень);
3) аналіз результатів досліду;
4) тлумачення результатів.

Після виконання цих етапів, як правило, формулюється висновок – спочатку у вигляді гіпотези[1], що потім перевіряється, тобто зазначений ланцюжок знову повторюється. Наявність або відсутність даних, що свідчать на користь сформульованого висновку, тобто виконання або невиконання заздалегідь обмовлених умов, змушує дослідника прийняти або відхилити гіпотезу.

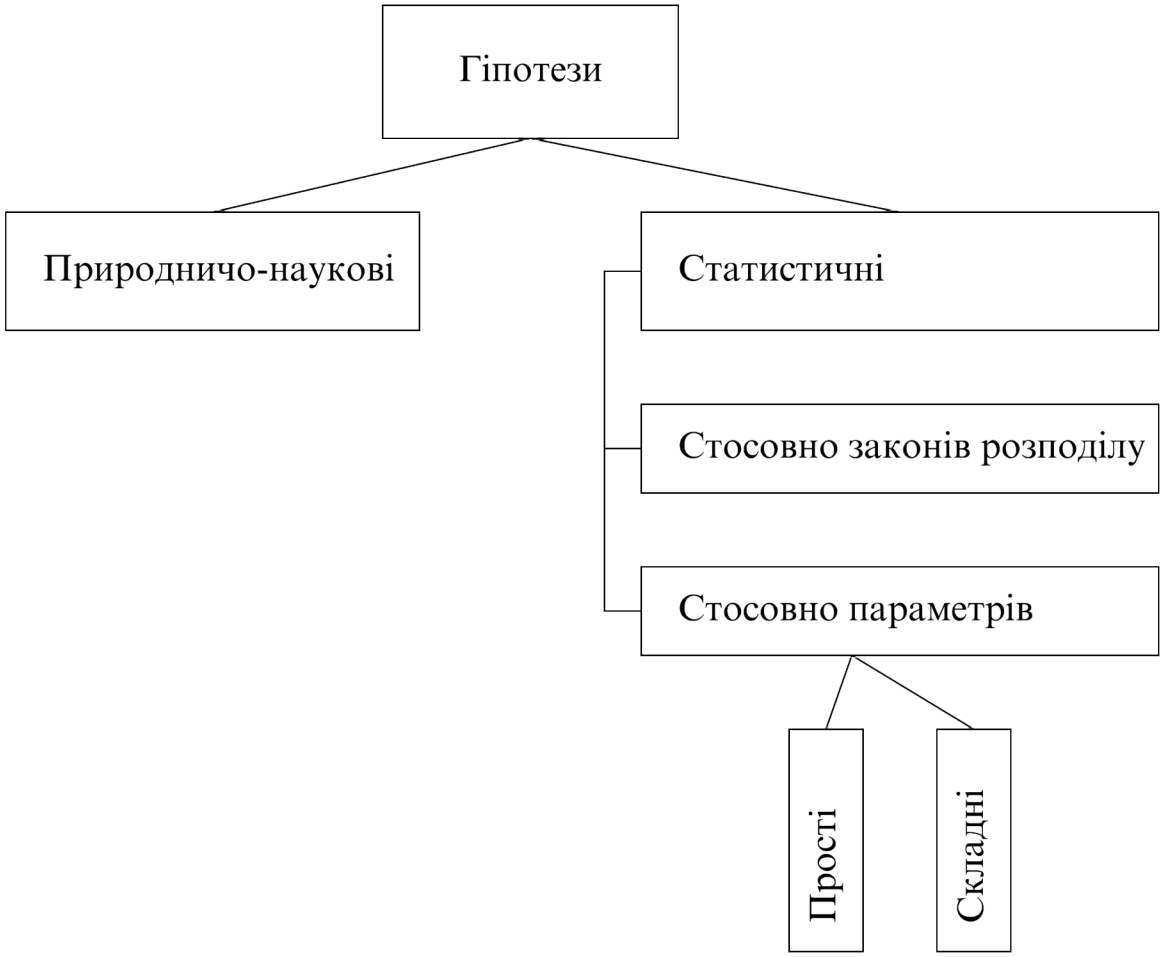


Рис. 10.1. Класифікація гіпотез

Існує два основні види гіпотез, які варто розрізняти (рис. 10.1): природничо-наукові і статистичні.

[1] Грец. hypothesis – основа, допущення.

Природничо-наукова гіпотеза – це припущення про закономірний (причинно-наслідковий) зв'язок явищ. Природничо-наукова гіпотеза перевіряється (обґрунтовується) за результатами досліду.

Статистична гіпотеза – припущення про якісні або кількісні властивості генеральної сукупності, що перевіряється за вибіркою. Уже сама методологія перевірки статистичної гіпотези, добре нам відома за вибірковим методом, − вивчення властивостей цілого за спостереженням його частини, повинна наводити на думку про те, що „абсолютної“ вірогідності підтвердження або відхилення статистичної гіпотези за результатами досліду очікувати не слід.

Приклад 10.1. Протягом XIX в. у Франції двічі велися дебати про самозародження життя. З 1802 р. з цією ж ідеєю протягом 30 років боровся Ж. Кюв'є[2] , що вийшов переможцем у суперечці з Жоффрі.

Потім, уже в 60-х рр., Л. Пастер брав участь у дебатах з Ф. Пуше, що відстоював концепцію самозародження життя. Теза Л. Пастера полягала у тому, що життя не самозароджується ніколи. Це важлива ознака природничо-наукової гіпотези, що або правильна, або не правильна. Про часткову вірогідність природничо-наукової гіпотези мова в принципі не може йти. ●

Як правило, говорячи про властивості генеральної сукупності, маємо на увазі одну або кілька ознак (випадкових величин), що характеризують дану генеральну сукупність. Оскільки випадкова величина визначається своїм описом (див. розд 4.19), повним або неповним, то й статистичну гіпотезу можна формулювати щодо опису випадкової величини.

Гіпотези щодо повного опису випадкових величин у генеральній сукупності називаються **гіпотезами про закони розподілу**.

Гіпотези щодо неповного опису випадкових величин у генеральній сукупності називаються **гіпотезами про параметри розподілу**.

Гіпотеза про параметр розподілу випадкової величини, що полягає в однозначному припущенні щодо значення параметра, називається **простою гіпотезою**.

Гіпотеза про параметр розподілу випадкової величини, що полягає в неоднозначному припущенні щодо значення параметра, називається **складною**. Іншими словами, це таке припущення, якому, у разі його підтвердження, буде задовольняти не одне, а множина значень параметра.

Приклад 10.2. Статистичними гіпотезами будуть такі припущення:

1) досліджувана ознака в деякій генеральній сукупності розподілена за нормальним законом (гіпотеза про закон розподілу);

2) в генеральній сукупності $\sigma^2=5$ (проста гіпотеза про параметр розподілу);

3) в генеральній сукупності $\mu>10$ (складна гіпотеза про параметр розподілу, оскільки параметр μ може мати будь-яке значення, більше 10, тобто ця гіпотеза

---

[2] Ж. Кюв’є (1769–1832), французький зоолог, один із реформаторів порівняльної анатомії, палеонтології, систематики тварин та ін. Почесний член Петербурзької АН (1802). Увів поняття типу в зоології. Установив принцип „кореляції органів“, на основі якого реконструював будову багатьох тварин, які вимерли. Не визнавав мінливості видів, пояснюючи зміну викопних фаун теорією катастроф.

складається з нескінченного числа простих гіпотез). ●

Прийнято називати гіпотезу, що перевіряється, **основною** або **нульовою**[3] (**нуль-гіпотезою)** і позначати H_0. Припущення, протилежне гіпотезі H_0, називається **альтернативною гіпотезою**. Замість однієї альтернативної гіпотези можуть розглядатися декілька альтернативних. Вони позначаються так: $H_1, H_2, \ldots$.

Приклад 10.2 (продовження).

1) H_0 : $\{X \sim N(\mu, \sigma)\}$;

2) H_0 : $\{\sigma = 5\}$, H_1 : $\{\sigma < 5\}$, H_2 : $\{\sigma > 5\}$;

3) H_0 : $\{\mu > 10\}$, H_1 : $\{\mu \le 10\}$. ●

Зауваження 10.1. Складні гіпотези формулюють у випадку, коли важливо зафіксувати перевищення параметром граничного рівня. Наприклад, статистична гіпотеза H_0 у прикл. 10.2, 3 може формулюватися в такій ситуації. У водному середовищі починають відбуватися деякі небажані процеси, якщо середній рівень його забруднення промисловими стоками перевищує значення, що дорівнює 10 од. Чи буде середній рівень забруднення водного середовища дорівнювати 12 або 15 од., не є важливе, оскільки пусковий механізм небажаних процесів уже діє. Факт досягнення й перевищення граничного значення, що дорівнює 10 од., складає суть статистичної гіпотези H_0. ▲

Приклад 10.3 (Гласс, 1976). Історія перевірки статистичних гіпотез бере початок з XVIII ст. Очевидно, найперший приклад статистичної гіпотези з'явився в написаній в 1710 р. Дж. Арбутнотом (1667 – 1735) роботі „Доводы в пользу божественных пророчеств. Выведенные на основе постоянных и систематических наблюдений над рождением обоих полов“. Відзначивши, що записи протягом 82 років свідчать про перевагу народжених хлопчиків, Арбутнот довів, що ці дані спростовують гіпотезу про те, що народження чоловіків і жінок однакове за ймовірністю. На думку Арбутнота, більша частота народження чоловіків – результат втручання Провидіння, оскільки більш імовірно, що чоловіка можуть убити на війні або він загине від непосильної роботи, не досягши зрілості. Сучасний дослідник у подібній ситуації міг би спробувати перевірити просту гіпотезу $H_0 : \{p = 0{,}5\}$ для параметра біномного розподілу (див. прикл. 9.5 і продовження його розгляду в розд. 9.12). ●

10.2. Схема перевірки природничо-наукової гіпотези

Підставою для перевірки слугує наше переконання в пізнаваності світу дослідним шляхом, тобто в тім, що природа дає правильні відповіді на коректно поставлені питання. Фольклор природознавства говорить, що природа у відповідь на питання дослідника відповідає “ні” голосно, а “так” – тихо. Отже, **гіпотеза має бути відхилена, якщо є суперечний приклад, але не обов'язково має бути прийнята, якщо такого прикладу знайти не можна**.

---

[3] Назва пояснюється грою слів. Перевіркою гіпотез спочатку займалась філософія – наука, завданням якої був збір доказів для анулювання гіпотез.

На цьому й заснована **логічна схема перевірки природничо-наукової гіпотези** (Тутубалін, 1982).

1. Формулюється гіпотеза H_0, що перевіряється.

2. Підбирається деяка подія S, що у випадку істинності гіпотези H_0 відбутися в досліді не може, тобто S – неможлива подія, називана **критичною**. Критична подія буде індикатором („лакмусовим папірцем“) під час перевірки наукової гіпотези. Відношення критичної події до гіпотези можна записати ще так:

$$P\left(\left\{S|H_0\right\}\right)=0. \qquad (10.1)$$

3. Проводиться дослід, результатом якого можуть бути події S і $\overline{S}$.

3.1. Якщо в досліді **настала критична подія** S (почуте голосне „ні“), тоді **гіпотеза** H_0 **відхиляється** (**відкидається**).

3.2. Якщо в досліді **не настала критична подія** S, то можна не достаньо обґрунтовано сказати, що **гіпотеза** H_0 **приймається** (**підтверджується**), але, оскільки „й незаряджена рушниця раз на рік стріляє“, то варто обережно сказати, що гіпотеза H_0 (поки!) не відкидається (тихе „так“).

Зауваження 10.2. Нульова ймовірність того, що в результаті досліду відбулася подія S, що відповідає гіпотезі H_0, записується у формі (10.1), що збігається з формою запису умовної ймовірності (див. розд. 4.3), але слід мати на увазі, що це не зовсім вдалий, хоча й дуже зручний, запис, оскільки гіпотезі H_0 ймовірність приписати не можна. Вона або правильна, або ні (див. прикл. 10.1). ▲

Зауваження 10.3. Опис логічної схеми перевірки природничо-наукової гіпотези уможливив питання про те, як довго потрібно проводити дослід для надійного її підтвердження. Адже якщо, скажімо, в 10 послідовних дослідах критична подія не наставала, то чи може дослідник бути впевнений у тому, що і в наступному 11-му досліді критична подія не настане. А втім навіть удосконалювання техніки проведення досліду й уточнення умов, за яких дослід проводиться, не може виключити можливість перегляду вимог до самої критичної події. Коли ж цьому може бути покладений край? Існує декілька поглядів. Перший (песимістичний, висловлений у дещо грубій, але досить точній формі) полягає в тому, що поки живі прихильники старої теорії, нову „впровадити“ не вдасться, навіть якщо результати досліду її начебто підтверджують. За словами М. Планка (1933), „велика наукова ідея рідко впроваджується шляхом поступового переконання й обернення своїх супротивників. У дійсності справа відбувається так, що опоненти поступово ідуть з життя, а наступне покоління з самого початку освоюється з новою ідеєю.“ Запас міцності позицій прихильників старої теорії може виявитися досить різноманітним – можна сперечатися з приводу чистоти досліду, точності формулювання висновків, посилатися на авторитети й т.д. Інший погляд (більш оптимістичний) формулюється у вигляді ланцюжка змін щодо нової теорії: „це нісенітниця“ → „у цьому щось є“ → „дивно, як я сам до цього не дійшов?“ (інший варіант: „я про це знав/казав і раніше“)[4] .

---

[4] Вчений XVIII ст. ірландець Кірван (саме він запропонував взяти теплоємність води за одиницю, від цієї пропозиції бере початок теплова одиниця калорія, яка і на сьогодні не втратила сво-

На наш погляд, можна дати таку інтерпретацію прийняття природничо-наукової гіпотези. Якщо в деякій дуже довгій серії n дослідів (вони можуть проводитися протягом достатньо тривалого, за мірками людського життя, проміжку часу різними дослідниками) спостерігається тільки подія $\overline{S}$ (отже, критична подія S не спостерігалася жодного разу), то частота цієї події дорівнює довжині серії, тобто $n(\overline{S}) = n$. Відтак можна говорити, що прийняття природничо-наукової гіпотези відбувається тоді, коли статистична ймовірність події $\overline{S}$ дорівнює одиниці (умова стабілізації відносної частоти тут виконується в „чистому“ вигляді) або коли це є вірогідна подія:

$$P^*(\overline{S}) = \frac{n(\overline{S})}{n} = \frac{n}{n} = 1. \ \blacktriangle \qquad (10.2)$$

10.3. Схема перевірки статистичної гіпотези

Логічна схема перевірки статистичної гіпотези виходить зі схеми перевірки наукової гіпотези з урахуванням основної ідеї вибіркового методу. Оскільки достовірності статистичної гіпотези очікувати не доводиться, то „сферу“ її дії дещо звузимо, тобто будемо вважати, що властивість, що виражається нею, виконується не для всіх, а майже для всіх об'єктів генеральної сукупності. Це означає, що для гіпотези H_0, що перевіряється, критична подія S повинна бути малоймовірною (див. розд. 3.4), тобто

$$P\left(\left\{S|H_0\right\}\right) = \alpha, \qquad (10.3)$$

де α – **рівень значущості**. З погляду класичної ймовірності (див. розд. 3.6), рівень значущості – це частка об'єктів генеральної сукупності, для яких H_0 не виконується.

Тепер настав час для застосування принципу практичної достовірності (див. розд. 3.2). Він – як чехівська рушниця, що з'явилася в першому акті й має вистрілити в другому. Нагадаємо, що в цьому принципі мова йшла про те, як варто поводитися з малоймовірною випадковою подією в одному або декількох дослідах. Саме така ситуація (другий акт п'єси) має місце при перевірці статистичної гіпотези. По-перше, є малоймовірна подія S (критична подія). По-друге, є серія дослідів, що буде розглядатися як один укрупнений дослід (пояснення того, чому потрібно використовувати саме такий укрупнений дослід, дамо дещо пізніше). З „укрупненням“ досліду ми вже стикалися (див. розд. 8.8).

Отже, **логічна схема перевірки статистичної гіпотези** така (Тутубалін, 1982).

1. Формулюється гіпотеза H_0.

го значення) був прихильником флогістонної теорії, на захист якої у 1778 р. написав трактат. Щоб переконати Кірвана, відомі французькі вчені Бертоллє, Морво, Лаплас, Фуркруа та Лавуазьє склали на неї пункт за пунктом заперечення і надіслали автору. Ознайомившись з їх доводами, Кірван направив французьким вченим листа, в якому повідомив, що він повністю „складає зброю“ і сам напише спростування на свою книгу.

2. Підбирається критична подія S, що у випадку істинності гіпотези H_0 є малоймовірна. Для цієї події попередньо (!) задається ймовірність настання – рівень значущості α (ступінь малоймовірності, але так, звичайно, говорити не можна).

3. Проводиться дослід, результатом якого можуть бути події S і $\overline{S}$.

3.1. Якщо в досліді **критична подія** S **настала,** то **гіпотеза** H_0 **відхиляється** (**не підтверджується**) **на рівні значущості** α, (ще говорять, що **спостерігається значуще на рівні** α **відхилення від гіпотези** H_0).

3.2. Якщо в досліді **критична подія** S **не настала**, то **гіпотеза** H_0 **приймається** (**підтверджується**) **на рівні значущості** α (ще говорять, що **відхилення від гіпотези** H_0, **значущого на рівні** α**, не спостерігається**).

Зауваження 10.4. Звернемо особливу увагу на формулювання статистичної гіпотези H_0, що перевіряється. Для різних модельних ситуацій буде пропонуватися відповідний метод перевірки (фактично розбіжність в методах зводяться до способів вибору критичної події). Однак, незалежно від передбачуваного результату перевірки, **загальне правило постановки гіпотези** H_0 таке: варто формулювати її у вигляді твердження про те, що „деякий невідомий розподіл відповідає певному стандартному“, „невідоме значення деякого параметра дорівнює певному значенню“, „невідомі значення параметрів двох розподілів є однакові“ і т.ін. У подальшому зміст цього правила з'ясується, поки ж обмежимося лише тим, що самі методи вибору критичної події враховують саме таке формулювання статистичної гіпотези. ▲

Звичайно **критична подія** S **полягає в тому**, **що в результаті досліду** деяка спеціальним способом підібрана випадкова величина $\widetilde{g}$ (статистика) **набере значення з** деякої **критичної множини** (критичної **області**), яку також будемо позначати літерою S. Раніше ми називали статистиками тільки вибіркові числові характеристики (див. розд 4.14). Але це не єдині статистики. Взагалі кажучи, статистикою називається будь-яка величина, що обчислюється за вибіркою. Зрозуміло, що статистика, як і вибіркова числова характеристика, буде визначеною за вибіркою випадковою величиною. Іншими словами, статистика для кожної статистичної гіпотези визначається як деяка функція

$$\widetilde{g} = \widetilde{g}(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n) \tag{10.4}$$

вибіркових значень (вибірки) і є випадкова величина, розподіл якої визначається розподілом випадкової величини X і обсягом вибірки, тобто в умовах даного досліду розподіл $\widetilde{g}$ може вважатися відомим (див. розд. 8.6). Нагадаємо, що маленькими літерами латинського алфавіту ми позначаємо значення випадкової величини, а значком тильда (~) зверху – випадкові величини (див. розд. 5.2). Над вибірковими значеннями $x_1, x_2, \ldots, x_n$ поставлений значок тильда, оскільки за організації декількох вибірок ці значення будуть змінюватися, і, отже, якщо їх розглядати не як результат одиничної вибірки, то самі вони є випадкові величини. Одержання одного значення статистики в серії дослідів і є той укрупнений дослід, про який мова йшла на початку розділу. Статистика $\widetilde{g}$ може мати один

із стандартних розподілів: $N(0,1)$, $\tilde{\chi}^2_\nu$, $\tilde{t}_\nu$ або F_{ν_1,ν_2} (див. розд. 8.1 – 8.5). Критична область S виражається через статистику $\tilde{g}$ й, залежно від гіпотези H_0, набуває одну із форм (рис. 10.2):

а) $S = \{\tilde{g} \le g_\alpha\}$; (10.5)

б) $S = \{\tilde{g} \ge g_{1-\alpha}\}$; (10.6)

в) $S = \{\tilde{g} \le g_{\alpha/2}\} + \{\tilde{g} \ge g_{1-\alpha/2}\}$, (10.7)

де g_α, $g_{1-\alpha}$, $g_{\alpha/2}$, $g_{1-\alpha/2}$ – квантилі статистики $\tilde{g}$.

Випадки а) і б) являють собою **однобічні критичні області**, а випадок в) – **двобічну критичну область**.

Основний етап перевірки статистичної гіпотези (етап 3 логічної схеми перевірки статистичної гіпотези) полягає в обчисленні значення статистики $g^* = g(x_1, x_2, \ldots, x_n)$ для даної вибірки та перевірці того, чи потрапляє обчислене значення g^* до відповідної критичної області S для $\tilde{g}$. **Якщо значення g^* потрапляє до критичної області, то гіпотеза H_0 відхиляється, якщо значення g^* не потрапляє до критичної області S, то гіпотеза H_0 приймається**.

Як бачимо, перевірка настання критичної події зводиться до перевірки виконання однієї з нерівностей (10.5), (10.6), (10.7).

За допомогою поняття квантиля можна дати інше трактування цього етапу. Оскільки будь-яке значення випадкової величини, у тому числі й вибіркове значення g^* статистики $\tilde{g}$, є квантилем з відповідним індексом (див. розд. 4.17), то будемо вважати, що $g^* = g_q$, де індекс q визначається за таблицями квантилів для $\tilde{g}$. Далі скористаємося властивістю 3 квантилів для того, щоб виразити настання подій S і $\bar{S}$ через відношення не між значеннями квантилів, а між індексами квантилів:

а) $$\begin{pmatrix} S = \{g_q < g_\alpha\} \\ \bar{S} = \{g_q \ge g_\alpha\} \end{pmatrix} \Leftrightarrow \begin{pmatrix} S = \{q < \alpha\} \\ \bar{S} = \{q \ge \alpha\} \end{pmatrix}; \quad (10.8)$$

б) $$\begin{pmatrix} S = \{g_q > g_{1-\alpha}\} \\ \bar{S} = \{g_q \le g_{1-\alpha}\} \end{pmatrix} \Leftrightarrow \begin{pmatrix} S = \{q > 1-\alpha\} \\ \bar{S} = \{q \le 1-\alpha\} \end{pmatrix}; \quad (10.9)$$

в) $$\begin{pmatrix} S = \{g_q < g_{\alpha/2}\} + \{g_q > g_{1-\alpha/2}\} \\ \bar{S} = \{g_{\alpha/2} \le g_q \le g_{1-\alpha/2}\} \end{pmatrix} \Leftrightarrow \Leftrightarrow \begin{pmatrix} S = \{q < \alpha/2\} + \{q > 1-\alpha/2\} \\ \bar{S} = \{\alpha/2 \le q \le 1-\alpha/2\} \end{pmatrix}. \quad (10.10)$$

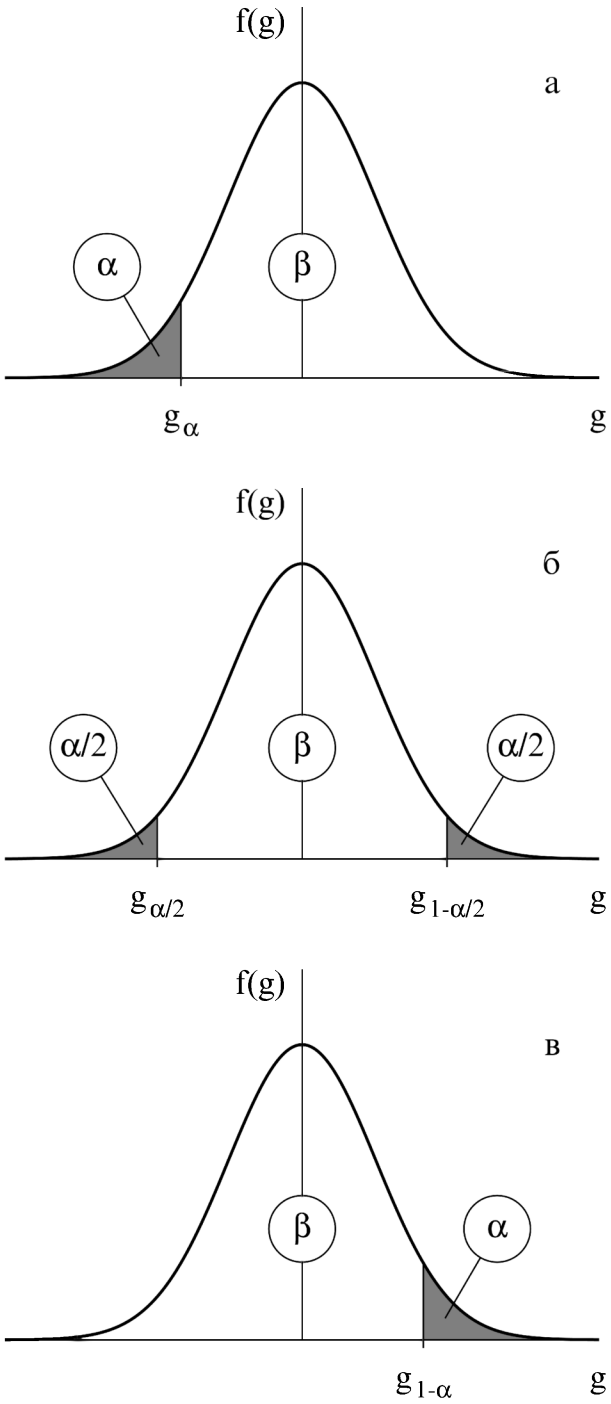


Рис. 10.2. Критичні області S, обмежені квантилями відповідної статистики $\widetilde{g}$ для заданого рівня значущості α: а – (10.5), (10.8), (10.11); б – (10.6), (10.9), (10.12); в – (10.7), (10.10), (10.13)

За цими відношеннями на основі індексу квантиля вводиться величина, називана ***P-значенням*** і визначена для розглянутих вище випадків у такий спосіб (Афіфі, 1982):

а) $P = q$;

б) $P = 1 - q$;

в) $P = 2\min(q, 1 - q)$.

Тобто настання подій S і $\overline{S}$ тепер виражається через відношення між рівнем значущості α і P-значенням:

$$\text{а)} \begin{pmatrix} S = \{g_q < g_\alpha\} \\ \overline{S} = \{g_q \geq g_\alpha\} \end{pmatrix} \Leftrightarrow \begin{pmatrix} S = \{P < \alpha\} \\ \overline{S} = \{P \geq \alpha\} \end{pmatrix}; \tag{10.11}$$

$$\text{б)} \begin{pmatrix} S = \{g_q > g_{1-\alpha}\} \\ \overline{S} = \{g_q \leq g_{1-\alpha}\} \end{pmatrix} \Leftrightarrow \begin{pmatrix} S = \{P < \alpha\} \\ \overline{S} = \{P \geq \alpha\} \end{pmatrix}; \tag{10.12}$$

$$\text{в)} \begin{pmatrix} S = \{g_q < g_{\alpha/2}\} + \{g_q > g_{1-\alpha/2}\} \\ \overline{S} = \{g_{\alpha/2} \leq g_q \leq g_{1-\alpha/2}\} \end{pmatrix} \Leftrightarrow \begin{pmatrix} S = \{P < \alpha\} \\ \overline{S} = \{P \geq \alpha\} \end{pmatrix}. \tag{10.13}$$

Якщо *P*-значення менше ніж α, то гіпотеза H_0 відхиляється з рівнем значущості α, у протилежному випадку гіпотеза H_0 приймається на рівні значущості α.

Прийняття або відхилення гіпотези не є логічним доведенням її істинності або хибності. Розглянемо це питання докладніше. Нехай сформульовані гіпотези H_0 та H_1 і визначена критична область S. Під час перевірки статистичної гіпотези H_0 може виникнути одна із чотирьох ситуацій, перелічених у табл. 10.1.

Таблиця 10.1

Рішення відносно гіпотези H_0	Істинність гіпотези H_0	
	Правильна	Неправильна
Відкидається	Помилка I роду, імовірність α	Правильне рішення
Приймається	Правильне рішення, імовірність 1-α	Помилка II роду

Приклад 10.4 (Біометрія, 1982). За результатами досліджень на тваринах токсичності деякого препарату може бути прийнято одне із двох рішень: A – випустити препарат у продаж, B – знищити як токсичний. Помилки, пов'язані з діями A і B, зовсім різні, різна й важливість їх уникнення. Нехай прийняте рішення A, у той час як правильним є рішення B, тобто препарат, визнаний безпечним, є небезпечним для пацієнта. Така помилка може призвести до смерті пацієнта, що вживає препарат, тобто є небезпечна. Нехай прийняте рішення В, у той час як правильним є рішення А, тобто в результаті неточностей експерименту нетоксичний препарат визнано небезпечним. Наслідки такої помилки виражаються у фі-

нансовому збитку, що, звичайно, неприємно, проте випадкове відхилення зовсім нешкідливого препарату менш небезпечне, ніж можлива загибель пацієнтів. ●

Помилки I і II роду зовсім різні за своїм значенням. **Розв'язання питання про те, яка з них небезпечніша (вибір значення α), перебуває поза статистикою й належить вже до методології тої експериментальної науки, результати якої статистично обговорюються**.

Величина рівня значущості визначається важливістю висновків, які повинен зробити експериментатор. Ухвалення рішення (вибір α) у звичайних ситуаціях ґрунтується на всьому досвіді розвитку природознавства. Спочатку нормою вважалося вибирати α=0,0027, ця величина фігурує в правилі 3σ (див. розд. 7.4): можна знехтувати частиною площі під нормальною кривою, вилученої з обох сторін від математичного сподівання μ більш ніж на три середньоквадратичних відхилення σ. Накопичення матеріалу в різних прикладних областях показало, однак, що при цьому занадто часто приймається неправильна нульова гіпотеза. Якщо вибрати $\alpha = 0{,}05$, то, як виявилося, здебільшого неправильно відкидається нульова гіпотеза. Можна рекомендувати таке правило:

а) якщо $P > 0{,}05$, то приймається нульова гіпотеза;

б) якщо $P < 0{,}01$, то нульова гіпотеза відхиляється й приймається альтернативна;

в) якщо $0{,}01 < P < 0{,}05$ 0,01, то результат уважається невизначеним.

В останньому випадку, однак, потрібне проведення додаткових дослідів. А втім дослідник має право, з огляду на всю сукупність знань з досліджуваного питання, прийняти те або інше рішення. Він може це зробити, взявши на себе відповідальність за цей крок, але при цьому повинен пам'ятати, що рівень значущості критерію є ймовірність помилкового відхилення нульової гіпотези.

Варто зазначити, що велике P-значення (наприклад, близьке до одиниці) аж ніяк не свідчить про велику правдоподібність гіпотези H_0. Це може бути, наприклад, наслідком:

1) недостатності дослідних даних;

2) зсуненості вибіркових значень.

Чим серйозніші наслідки помилки I роду, тим меншим повинен бути рівень значущості. Проте обирати його дуже близьким до нуля немає сенсу, оскільки різко зросте ймовірність помилки II роду. Наприклад, у разі перевірки токсичності препарату (прикл. 10.4) цей спосіб буде відхиляти будь-який новий лікарський препарат. Отже, необхідно вибирати оптимальний рівень значущості, ретельно зважуючи важливість наслідків помилок I і II роду. **Під час перевірки будь-якої статистичної гіпотези рішення ніколи не приймається з абсолютною впевненістю**, завжди існує ризик ухвалення неправильного рішення. Саме в контролі й оцінці цього ризику полягає суть перевірки статистичної гіпотези.

10.4. Гіпотези про закон розподілу випадкових величин

Припустимо, що ми бажаємо встановити, чи суперечать дослідні дані гіпотезі про те, що випадкова величина X розподілена за деяким законом. Наскільки важ-

ливе є розв'язання питання про закон розподілу, стане ясно з подальшого. Зараз обмежимося тим, що без знання закону розподілу, як правило, неможливе проведення статистичної обробки даних. Наприклад, всі інтервальні оцінки параметрів, розглянуті в розд. 9, ґрунтуються на тому, що досліджувана ознака розподілена за нормальним законом. Попередні висновки про закон розподілу засновані на результатах первинної обробки вибіркових даних, тобто на рядах розподілу частот або відносних частот для дискретної випадкової величини або гістограмах частот, відносних частот або вибіркової функції густини для неперервної випадкової величини (див. розд. 4.8, 4.10), після чого формулюється гіпотеза:

$$H_0\text{: \{ випадкова величина } X \text{ має функцію розподілу } F(x)\}.$$

Для розв'язання питання про прийняття гіпотези H_0 застосовують різні критерії, з яких найбільш простим є **критерій Пірсона**.

Застосування критерію χ^2 для дискретних випадкових величин

Припустимо, що проведені n незалежних дослідів, у яких дискретна випадкова величина X набула певні значення, тобто одержана вибірка обсягу n. Після первинної обробки вибіркових даних ми можемо побудувати ряди розподілу частот або відносних частот. Графічне зображення рядів може навести на думку про те, що досліджувана випадкова величина розподілена за деяким стандартним законом, наприклад біномним або Пуассона (див. розд. 5.1, 5.2). Тому необхідно з'ясувати, наскільки істотно (статистично значимо) відрізняються теоретичні ймовірності від вибіркових відносних частот. Для того щоб ідея статистичного висновку стала ясною, об'єднаємо дві таблиці – ряду розподілу (див. розд. 4.7, табл. 4.6) і відносних частот (див. розділ 4.7, табл. 4.7). В результаті одержимо табл. 10.2:

Таблиця 10.2

x_1	x_2	...	x_k	...	$x_{k=K}$
$p_1^* = \frac{n_1}{n}$	$p_2^* = \frac{n_2}{n}$	...	$p_k^* = \frac{n_k}{n}$	...	$p_K^* = \frac{n_{k=K}}{n}$
p_1	p_2	...	p_k	...	$p_{k=K}$

Зауважимо, що число вибіркових значень менше (точніше кажучи, не більше) числа можливих значень дискретної випадкової величини X (не змішуйте поняття обсягу вибірки n і числа значень k, див. розд 4.4), оскільки в досліді деякі значення дискретної випадкової величини X можуть і не спостерігатися. Тому в другому рядку таблиці, де наведені відносні частоти значень, можливо, потрібно вписати нульові значення. Пропускати ці нульові значення, перевіряючи статистичну гіпотезу про закон розподілу дискретної випадкової величини, не можна, тобто нульові вибіркові відносні частоти необхідно враховувати в критерії.

Далі висувається гіпотеза:

H_0: {дискретна випадкова величина X має ряд розподілу (x_k, p_k), $k = \overline{1, K}$},

тобто ряд розподілу виходить із першого й третього рядків табл. 10.2, а відхилення відносних частот від теоретичних ймовірностей p_k пояснюються випадковими причинами й обмеженістю вибірки.

У **критерії χ^2 Пірсона** статистикою служить величина

$$n\sum_{k=1}^{K}\frac{(p_k - p_k^*)^2}{p_k} = \sum_{k=1}^{K}\frac{(np_k - n_k)^2}{np_k} \sim \tilde{\chi}_\nu^2, \tag{10.14}$$

розподілена за законом Пірсона із числом степенів вільності $\nu = K - q$, де q – число незалежних умов, що накладаються на параметри теоретичного розподілу. Якщо вибіркове значення $(\chi_\nu^2)^*$ статистики (10.14) потрапляє в **однобічну критичну область** $S = \{\tilde{\chi}_\nu^2 \geq \chi_{\nu,1-\alpha}^2\}$ (відповідне P-значення менше наперед заданого рівня значущості α), то **гіпотеза H_0 відхиляється**, у протилежному випадку **гіпотеза H_0** приймається. Оскільки занадто малі значення $(\chi_\nu^2)^*$ можуть бути не обов'язково наслідком збігу з теоретичним розподілом (див. розд. 8.6), то **критична область** може вибиратися **двобічною** $S = \{\tilde{\chi}_\nu^2 \leq \chi_{\nu,\alpha/2}^2\} + \{\tilde{\chi}_\nu^2 \geq \chi_{\nu,1-\alpha/2}^2\}$.

Зупинимося дещо докладніше на обчисленні числа степенів вільності. Оскільки частоти завжди задовольняють умову

$$\sum_{k=1}^{K} n_k = n \tag{10.15}$$

або, що те ж саме, для відносних частот виконується умова

$$\sum_{k=1}^{K}\frac{n_k}{n} = 1, \tag{10.16}$$

то число степенів вільності уже стає менше на одиницю, тобто $\nu = m - 1$. Серед інших умов має місце збіг вибіркового середнього m з математичним сподіванням μ

$$\sum_{k=1}^{K} \frac{n_k}{n} x_k = \sum_{k=1}^{K} p_k^* x_k = m = \mu, \qquad (10.17)$$

або (і) збіг вибіркової незсуненої дисперсії $\hat{s}^2$ з дисперсією σ^2

$$\frac{n}{n-1} \sum_{k=1}^{K} \frac{n_k}{n} (x_k - m)^2 = \frac{n}{n-1} \sum_{k=1}^{K} p_k^* (x_k - m)^2 = \hat{s}^2 = \sigma^2, \qquad (10.18)$$

Виконуючи кожну з цих умов, число степенів свободи необхідно зменшувати на одиницю.

Приклад 10.3 (продовження). Є всі підстави припустити, що розподіл дітей за статтю є біномним, тому формулюється гіпотеза:

H_0 : {розподіл дітей за статтю є біномним}.

Для перевірки гіпотези складемо ряди із відносних частот і теоретичних ймовірностей $C_8^m \cdot p^m \cdot q^{8-m} = C_8^m \cdot (0{,}5)^8, m = 0, 1, \ldots, 8$ для біномного розподілу (табл. 10.3).

Таблиця 10.3

m	0	1	2	3	4	5	6	7	8
$\frac{n_m}{n}$	0,004051	0,02798	0,1004	0,2006	0,2818	0,2247	0,1145	0,03939	0,006443
p_m	0,003906	0,03125	0,1094	0,2188	0,2734	0,2188	0,1094	0,03125	0,003906

Вибіркове значення статистики (10.14) критерію Пірсона $(\chi_\nu^2)^* \approx 0{,}007025$. Під час обчислення параметрів теоретичного розподілу не використовуються вибіркові значення, тому число степенів вільності $\nu = K = 9$; $(\chi_\nu^2)^* < \chi_{\nu,1-\alpha}^2 \approx 16{,}9$ для рівня значущості α=0,05, отже, гіпотеза H_0 приймається. ●

Застосування критерію χ^2 для неперервних випадкових величин

Після первинної обробки вибіркових даних ми маємо интервальні ряди частот, відносних частот і вибіркової функції щільності (див. розд. 4.4, табл. 4.3). Якщо є підстави для порівняння одержаного вибіркового розподілу з тим або іншим стандартним розподілом, то висуваємо гіпотезу:

H_0: { неперервна випадкова величина X має функцію розподілу $F(x)$}.

Ймовірності попадання значення випадкової величини в інтервали $[a_k, a_{k+1})$ можна обчислити точно (див. рис. 10.3, а):

$$p_k = S_k = \int_{a_k}^{a_{k+1}} f(x)\,dx = F(a_{k+1}) - F(a_k) \tag{10.19}$$

або приблизно (див. рис. 10.3, б):

$$p_k = S_k = f(\xi)(a_{k+1} - a_k) \approx f\left(\frac{a_k + a_{k+1}}{2}\right)\cdot(a_{k+1} - a_k), \tag{10.20}$$

$$p_k = S_k = f(\xi)(a_{k+1} - a_k) \approx \frac{1}{2}\big(f(a_k) + f(a_{k+1})\big)\cdot(a_{k+1} - a_k). \tag{10.21}$$

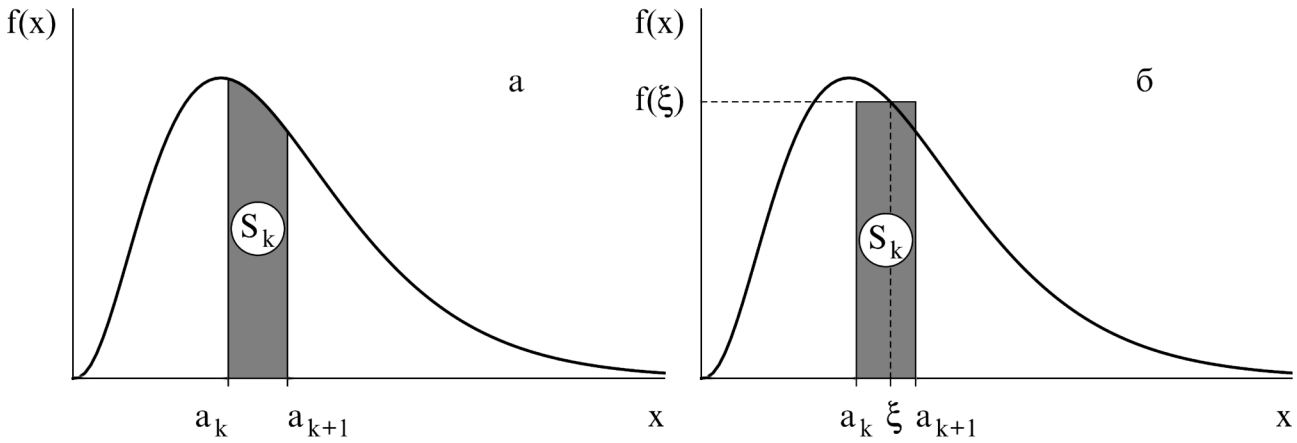


Рис. 10.3. Обчислення ймовірності p_k попадання значення випадкової величини в інтервал $[a_k, a_{k+1})$: а – точне (10.19); б – наближене (10.20)

Нижче наводимо відносні частоти p_k^* і теоретичні ймовірності p_k (табл. 10.4).

Таблиця 10.4

$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_k, a_{k+1})$	...	$[a_K, a_{K+1})$
$p_1^* = \frac{n_1}{n}$	$p_2^* = \frac{n_2}{n}$	...	$p_k^* = \frac{n_k}{n}$	...	$p_K^* = \frac{n_{k=K}}{n}$
p_1	p_2	...	p_k	...	p_K

Якщо функція густини $f(x)$ і функція розподілу $F(x)$ визначені на всій числовій осі (наприклад, у випадку гіпотези: $\{X$ розподілена за нормальним законом$\}$), необхідно в табл. 10.4 ввести додаткові інтервали: $(-\infty, a_1)$ і $[a_{K+1}, +\infty)$, для яких імовірності p_0 та p_{K+1} обчислюються в такий спосіб:

$$p_0 = \int_{-\infty}^{a_1} f(x)\,dx = F(a_1) - F(-\infty) = F(a_1), \tag{10.22}$$

$$p_{K+1} = \int_{a_K}^{+\infty} f(x)\,dx = F(+\infty) - F(a_{K+1}) = 1 - F(a_{K+1}). \tag{10.23}$$

В результаті маємо табл. 10.5.

Таблиця 10.5

$(-\infty, a_1)$	$[a_1, a_2)$	$[a_2, a_3)$	...	$[a_k, a_{k+1})$	...	$[a_K, a_{K+1})$	$[a_{K+1}, +\infty)$
0	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$	...	$\frac{n_K}{n}$	0
p_0	p_1	p_2	...	p_k	...	p_K	p_{K+1}

У **критерії** χ^2 **Пірсона** для неперервної випадкової величини статистикою служить величина

$$n\sum_{k=0}^{K+1}\frac{(p_k - p_k^*)^2}{p_k} = \sum_{k=0}^{K+1}\frac{(np_k - n_k)^2}{np_k} \sim \tilde{\chi}_\nu^2, \tag{10.24}$$

розподілена за законом $\tilde{\chi}_\nu^2$ із числом степенів вільності $\nu = K + 2 - q$, де q – число незалежних умов, що накладаються на параметри теоретичного розподілу. Увипадку нормального закону математичне сподівання μ і дисперсія σ^2 невідомі, тому, обчислюючи величини p_k слід застосовувати їхні вибіркові оцінки, а число степенів вільності $\nu = K + 2 - 1$ (одна одиниця „пішла“на умову для частот) варто зменшити ще на дві одиниці, тобто $\nu = K - 1$.

Якщо вибіркове значення статистики $(\chi_\nu^2)^*$ попадає в **однобічну критичну область** $S = \{\chi_{\nu,*}^2 \geq \chi_{\nu,1-\alpha}^2\}$ (відповідне P-значення менше наперед заданого рівня значущості α), то **гіпотеза** H_0 **відхиляється**, у протилежному випадку **гіпоте-**

за H_0 **приймається**. Занадто малі значення $(\chi^2_\nu)^*$ не обов'язково можуть бути результатом прийнятного збігу з теоретичним розподілом (див. розд. 10.1), тому **критична область** обирається **двобічною** $S = \{\chi^2_{\nu,*} \le \chi^2_{\nu,\alpha/2}\} + \{\chi^2_{\nu,*} \ge \chi^2_{\nu,1-\alpha/2}\}$.

Приклад 10.5 (Афіфі, 1982). У n=113 пацієнтів, що знаходилися в критичному стані, вимірявся систолічний тиск X=SP (**S**ystolic **P**ressure), мм рт. ст.: ΔSP=145, SP_{min}=26, SP_{max}=171, m_{SP}=106,25, $\hat{s}^2_{SP}$=941,46, $\hat{s}^2_{SP}$=30,68.

У процесі групування вибіркових даних число інтервалів K=11 визначалося за формулою (4.3). На гістограмах (рис. 10.4) зображені вибіркові функції густини f^*_k й розподілу F^*_k. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{SP}, \hat{s}_{SP})$.

На основі порівняння дослідних даних і теоретичного розподілу формулюємо **гіпотезу** про нормальний закон розподілу $H_0 : \{SP \sim N(m_{SP}, \hat{s}_{SP})\}$. Перевірка гіпотези здійснюється з рівнем значущості α=0,05. Обчислене за (10.24) значення статистики $(\chi^2_\nu)^* \approx 16{,}77 < \chi^2_{\nu,1-\alpha} \approx 18{,}31$, $\nu = K - 1$, отже, гіпотеза H_0 приймається. Це дає підставу для виконання інтервальної оцінки математичного сподівання й дисперсії за формулами (9.62), (9.80). Після обчислення значення квантилів $t_{n-1,1-\alpha/2} \approx 1{,}98$, $\chi^2_{n-1,\alpha/2} \approx 84{,}60$, $\chi^2_{n-1,1-\alpha/2} \approx 143{,}18$ одержуємо шукані інтервальні оцінки у вигляді подвійних нерівностей

$$106{,}25 - \frac{30{,}68}{\sqrt{113}} \cdot 1{,}98 < \mu_{SP} < 106{,}25 + \frac{30{,}68}{\sqrt{113}} \cdot 1{,}98,$$

$$100{,}54 < \mu_{SP} < 111{,}97;$$

$$\frac{941{,}46 \cdot (113-1)}{143{,}18} < \sigma^2_{SP} < \frac{941{,}46 \cdot (113-1)}{84{,}60},$$

$$736{,}29 < \sigma^2_{SP} < 1246{,}12,. \bullet$$

Приклад 10.6 (Афіфі, 1982). У n=113 пацієнтів, що знаходилися в критичному стані, вимірявся діастолічний тиск X=DP (**D**yastolic **P**ressure), мм рт. ст:. ΔDP=98, DP_{min}=10, DP_{max}=108, m_{DP}=58,49, $\hat{s}^2_{DP}$=341,34, $\hat{s}_{DP}$=18,48.

У процесі групування вибіркових даних число інтервалів K=12 визначалося за формулою (4.3). На гістограмах (рис. 10.5) зображені вибіркові функції густини f^*_k й розподілу F^*_k. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{DP}, \hat{s}_{DP})$.

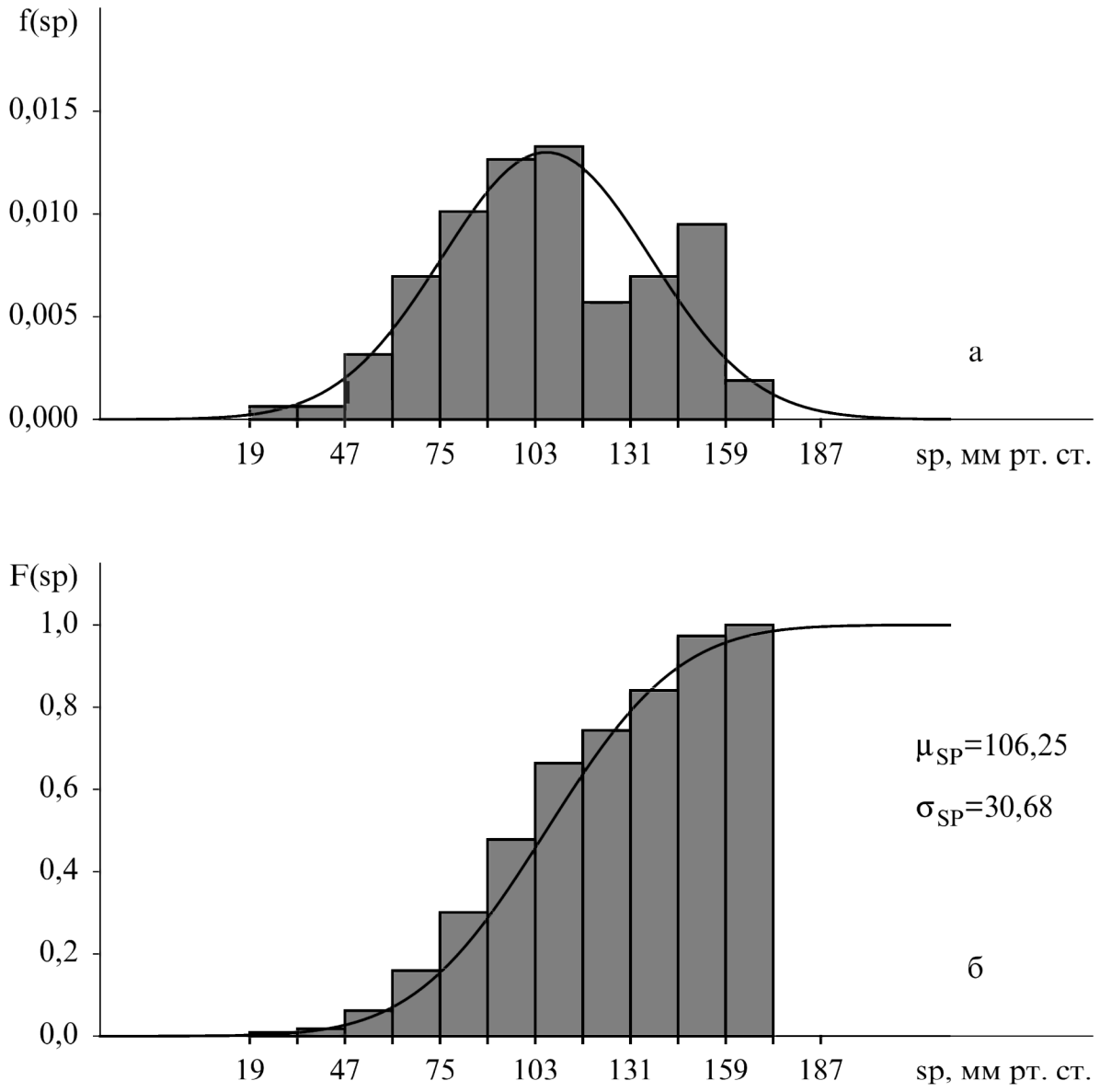


Рис. 10.4. Результати обробки вибіркових даних для прикл. 10.5:
а – функція густини; б – функція розподілу

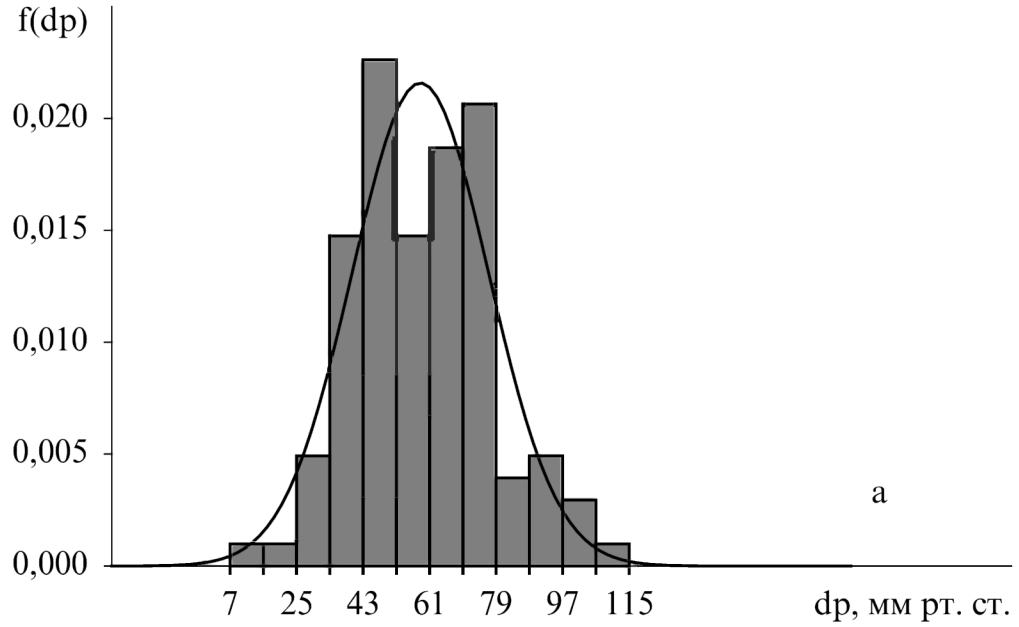
На основі порівняння дослідних даних і теоретичного розподілу формулюється **гіпотеза** про нормальний закон розподілу $H_0 : \{DP \sim N(m_{DP}, \hat{s}_{DP})\}$. Перевірка гіпотези H_0 здійснюється з рівнем значущості α=0,05. Обчислене за (10.25) значення статистики $(\chi^2_\nu)^* \approx 13{,}14 < \chi^2_{\nu,1-\alpha} \approx 19{,}68$, $\nu = K - 1$, отже, гіпотеза приймається. Це дає підставу для виконання інтервальної оцінки математичного сподівання й дисперсії за формулами (9.62), (9.80). Після обчислення значення квантилів $t_{n-1,1-\alpha/2} \approx 1{,}98$, $\chi^2_{n-1,\alpha/2} \approx 84{,}60$, $\chi^2_{n-1,1-\alpha/2} \approx 143{,}18$, одержуємо шукані інтервальні оцінки у вигляді подвійних нерівностей

$$58{,}49 - \frac{18{,}48}{\sqrt{113}} \cdot 1{,}98 < \mu_{DP} < 58{,}49 + \frac{18{,}48}{\sqrt{113}} \cdot 1{,}98,$$

$$55{,}05 < \mu_{DP} < 61{,}93;$$

$$\frac{341{,}34 \cdot (113-1)}{143{,}18} < \sigma^2_{DP} < \frac{341{,}34 \cdot (113-1)}{84{,}60},$$

$$267{,}14 < \sigma^2_{DP} < 452{,}12. \quad \bullet$$



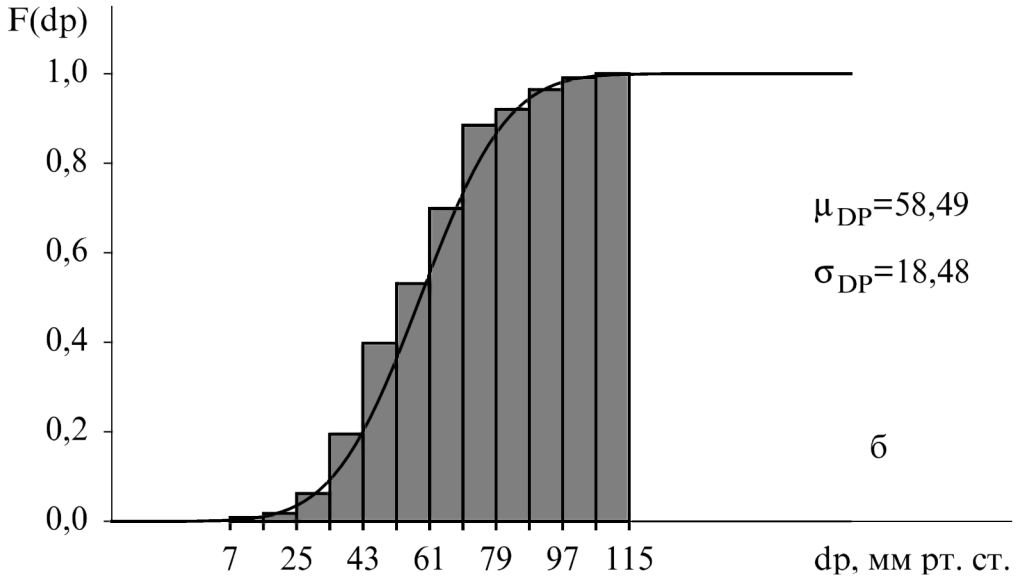
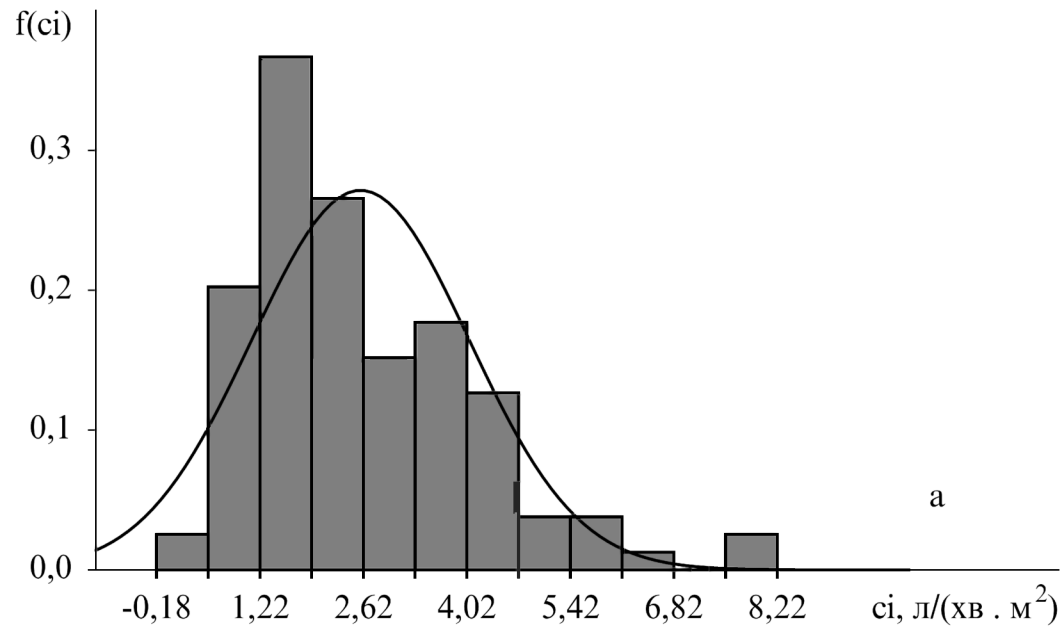


Рис. 10.5. Результати обробки вибіркових даних для прикл. 10.6: а – функція густини; б – функція розподілу

Приклад 10.7 (Афіфі, 1982). У n=113, що знаходилися в критичному стані, вимірявся серцевий індекс $X=CI$ (**C**ardial **I**ndex), л/(хв·м^2), що визначається як продуктивність серця, л/хв, поділена на площу тіла, м^2: ΔCI=7,46, CI_{min}=0,17, CI_{max}=7,63, m_{CI}=2,58, $\hat{s}^2_{CI}$=2,15, $\hat{s}_{CI}$=1,47.

У процесі групування вибіркових даних число інтервалів K=12 визначалося за формулою (4.3). На гістограмах (рис. 10.6) зображені вибіркові функції густини f^*_k й розподілу F^*_k. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{CI}, \hat{s}_{CI})$.



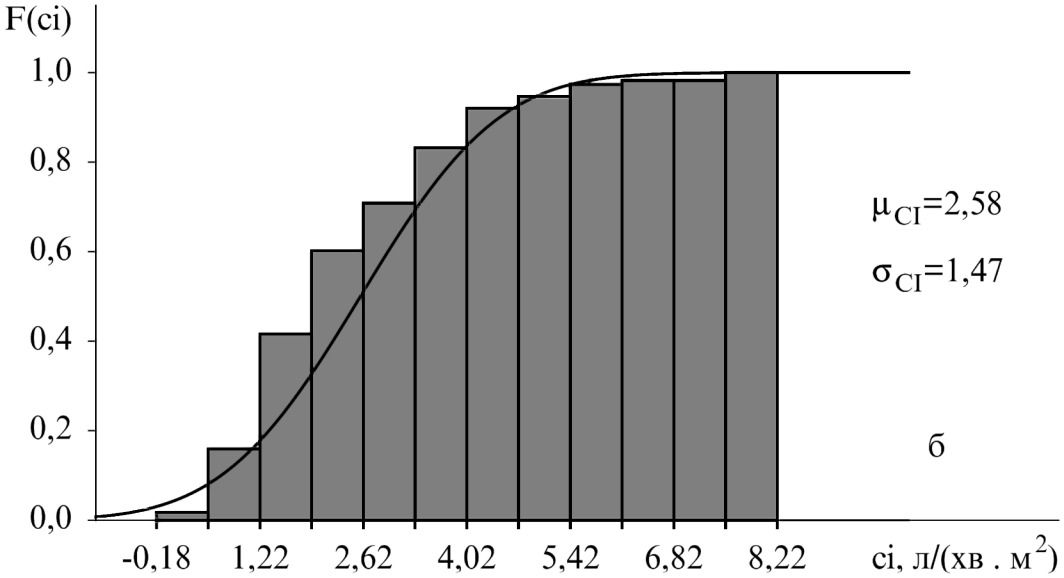


Рис. 10.6. Результати обробки вибіркових даних для прикл. 10.7: а – функція густини; б – функція розподілу для первинної змінної

Порівняння дослідних даних і теоретичного розподілу, взагалі кажучи, не дає підстав для **гіпотези** про нормальний закон розподілу $H_0:\{CI \sim N(m_{CI}, \hat{s}_{CI})\}$, але гіпотеза все ж формулюється. Перевірка гіпотези H_0 здійснюється з рівнем значущості α=0,05. Обчислене за (10.24) значення статистики $(\chi^2_\nu)^* \approx 13{,}14 <$ $< (\chi^2_\nu)^* \approx 13{,}14 < \chi^2_{\nu,1-\alpha} \approx 19{,}68$, $\nu = K-1$, отже, гіпотеза відхиляється.

Перетворення $X' = CI' = lg\,CI$ дозволяє одержати симетричні гістограми вибіркових функцій f_k^* і F_k^* (рис. 10.7). Під час групування даних число інтервалів K=12. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{CI'}, \hat{s}_{CI'})$. Для перетвореної змінної: CI'_{min}= –0,77, CI'_{max}=0,88, $\Delta CI'$=1,65, $m_{CI'}$=0,34, $\hat{s}^2_{CI'}$=0,072, $\hat{s}_{CI'}$=0,27.

На основі порівняння дослідних даних і теоретичного розподілу формулюємо **гіпотезу** про нормальний закон розподілу $H_0:\{CI' \sim N(m_{CI'}, \hat{s}_{CI'})\}$. Перевірка гіпотези H_0 здійснюється з рівнем значущості α=0,05. Обчислене за (10.24) значення статистики $(\chi^2_\nu)^* \approx 13{,}14 < \chi^2_{\nu,1-\alpha} \approx 19{,}68$, $\nu = K-1$, отже, гіпотеза приймається. Це дає підставу для виконання інтервальної оцінки математичного сподівання й дисперсії за формулами (9.62), (9.80). Після обчислення значення квантилів $t_{n-1,1-\alpha/2}$≈1,98, $\chi^2_{n-1,\alpha/2}$≈84,60, $\chi^2_{n-1,1-\alpha/2}$ ≈143,18, одержуємо шукані інтервальні оцінки у вигляді подвійних нерівностей

$$0{,}33 - \frac{0{,}27}{\sqrt{113}} \cdot 1{,}98 < \mu_{CI'} < 0{,}33 + \frac{0{,}27}{\sqrt{113}} \cdot 1{,}98,$$

$$0{,}28 < \mu_{CI'} < 0{,}38;$$

$$\frac{0{,}072 \cdot (113-1)}{143{,}18} < \sigma^2_{CI'} < \frac{0{,}072 \cdot (113-1)}{84{,}60},$$

$$0{,}056 < \sigma^2_{CI'} < 0{,}096.$$

Оскільки параметри первинної $X = CI$ і перетвореної $X' = CI' = \lg X$ змінних зв'язані співвідношенням

$$\lg \mu_{CI} = \mu_{CI'} + 1{,}15 \cdot \sigma^2_{CI'} \qquad \mu_{CI} = 10^{\mu_{CI'} + 1{,}15 \cdot \sigma^2_{CI'}},$$

яке випливає безпосередньо з визначення математичного сподівання й дисперсії, то інтервальна оцінка може бути проведена й для первинної змінної

$$10^{0{,}29 + 1{,}15 \cdot 0{,}056} < \mu_{CI} < 10^{0{,}39 + 1{,}15 \cdot 0{,}096}, \qquad 2{,}24 < \mu_{CI} < 3{,}13.$$

Безпосереднє обчислення інтервальної оцінки для первинної змінної за (9.89) приводить до подвійної нерівності

$$2{,}56 - \frac{1{,}46}{\sqrt{113}} \cdot 1{,}98 < \mu_{CI} < 2{,}56 + \frac{1{,}46}{\sqrt{113}} \cdot 1{,}98, \qquad 2{,}28 < \mu_{CI} < 2{,}83,$$

що узгоджується з оцінкою, одержаною за допомогою перетворення змінної (див. попередню подвійну нерівність), і підтверджує стійкість інтервального оцінювання математичного сподівання для симетричного вибіркового розподілу. ●

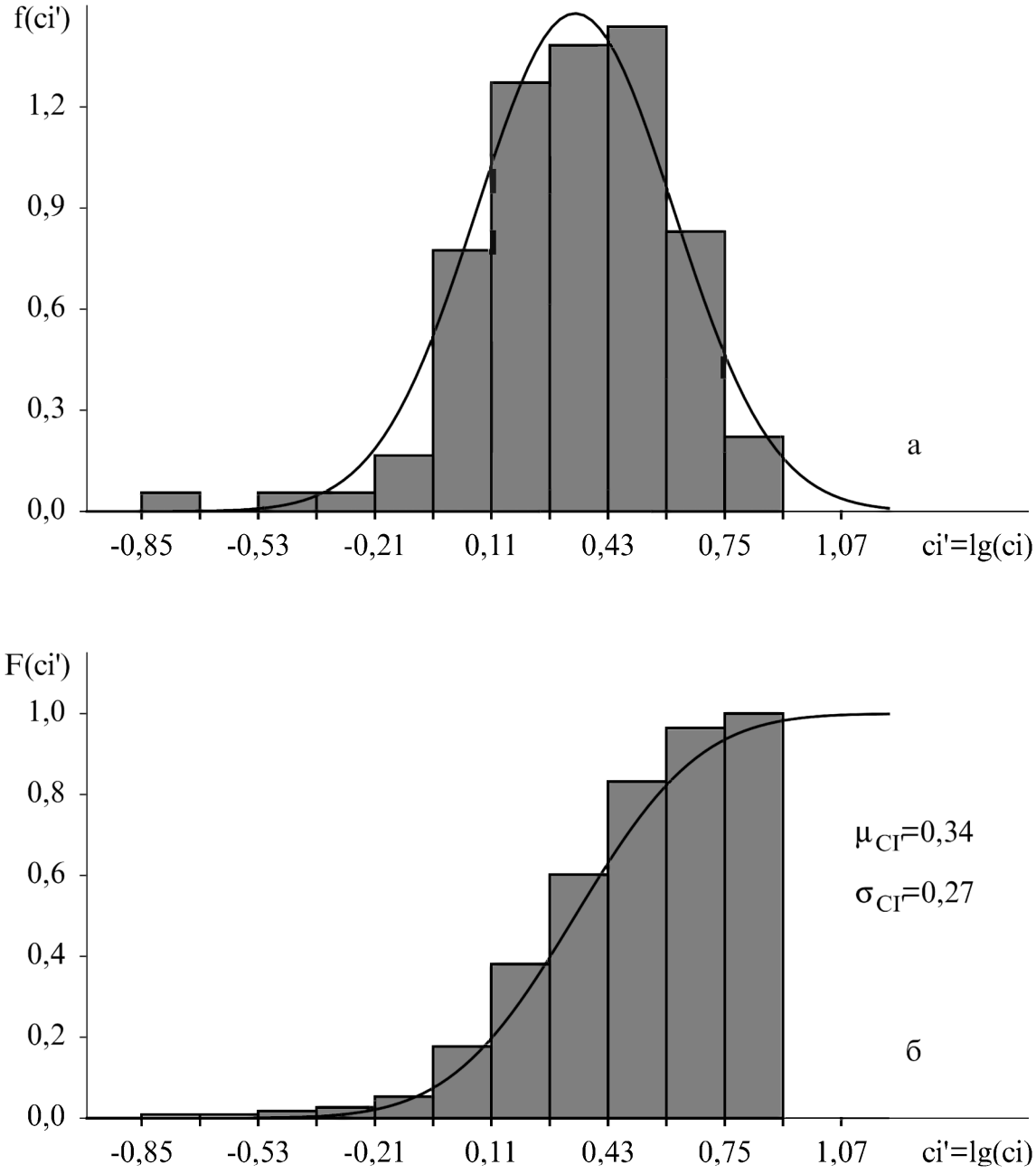


Рис. 10.7. Результати обробки вибіркових даних для прикл. 10.7: а – функція густини; б – функція розподілу для перетвореної змінної

Зауваження 10.5. Перехід від початкових величин до перетворених, як у прикл. 10.6, є досить поширений. Відомі добре розроблені рекомендації для по-

дібних перетворень, що грунтуються на формі вибіркових розподілів f_k^* (Біометрія, 1982). ▲

Приклад 10.8 (Афіфі, 1982). У n=113, що знаходилися в критичному стані, вимірювався середній час циркуляції $X=MCT$ (**M**ean **C**irculation **T**ime), с: ΔMCT =7,46, $MCT_{\min}$ =8,1, $MCT_{\max}$ =50,9, m_{MCT} =22,78, $\hat{s}^2_{MCT}$ =109,53, $\hat{s}_{MCT}$ =10,47.

У процесі групування вибіркових даних число інтервалів K=11 визначалося за формулою (4.3). На гістограмах (рис. 10.8) зображені вибіркові функції густини f_k^* й розподілу F_k^*. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{MCT}, \hat{s}_{MCT})$.

Порівняння дослідних даних і теоретичного розподілу, взагалі кажучи, не дає підстав для прийняття **гіпотези** про нормальний закон розподілу $H_0 : \{MCT \sim N(m_{MCT}, \hat{s}_{MCT})\}$, але гіпотеза все ж формулюється. Перевірка гіпотези H_0 здійснюється з рівнем значущості α=0,05. Обчислене за (10.25) значення статистики $(\chi^2_\nu)^* \approx 41{,}67 < \chi^2_{\nu,1-\alpha} \approx 19{,}68$, $\nu = K-1$, отже, гіпотеза відхиляється.

Перетворення $X' = MCTI = \lg MCT$ дозволяє одержати симетричні гістограми вибіркових функцій f_k^* і F_k^* (рис. 10.9). Під час групування вибіркових даних число інтервалів K=12. Суцільними лініями показані функція густини й функція нормального розподілу $N(m_{MCT'}, \hat{s}_{MCT'})$. Для перетвореної змінної: $MCT_{\min}$ =0,91, $MCT_{\max}$ =1,78, $\Delta MCT'$=0,86, $m_{CI'}$=1,32, $\hat{s}^2_{CI'}$=0,037, $\hat{s}_{CI'}$=0,19.

На основі порівняння дослідних даних і теоретичного розподілу формулюється **гіпотеза** про нормальний закон розподілу $H_0 : \{MCT' \sim N(m_{MCT'}, \hat{s}_{MCT'})\}$. Перевірка гіпотези H_0 здійснюється з рівнем значущості α=0,05. Обчислене за (10.24) значення статистики $(\chi^2_\nu)^* \approx 8{,}3 < \chi^2_{\nu,1-\alpha} \approx 19{,}68$, $\nu = K-1$, отже, гіпотеза приймається. Це дає підставу для виконання інтервальної оцінки математичного сподівання й дисперсії за формулами (9.62), (9.80). Після обчислення значення квантилів $t_{n-1,1-\alpha/2} \approx 1{,}98$, $\chi^2_{n-1,\alpha/2} \approx 84{,}60$, $\chi^2_{n-1,1-\alpha/2} \approx 143{,}18$ одержуємо шукані інтервальні оцінки у вигляді подвійних нерівностей

$$0{,}33 - \frac{0{,}27}{\sqrt{113}} \cdot 1{,}98 < \mu_{CI'} < 0{,}33 + \frac{0{,}27}{\sqrt{113}} \cdot 1{,}98,$$

$$0{,}28 < \mu_{CI'} < 0{,}38;$$

$$\frac{0{,}072 \cdot (113-1)}{143{,}18} < \sigma^2_{CI'} < \frac{0{,}072 \cdot (113-1)}{84{,}60},$$

$$0{,}056 < \sigma^2_{CI'} < 0{,}096.$$

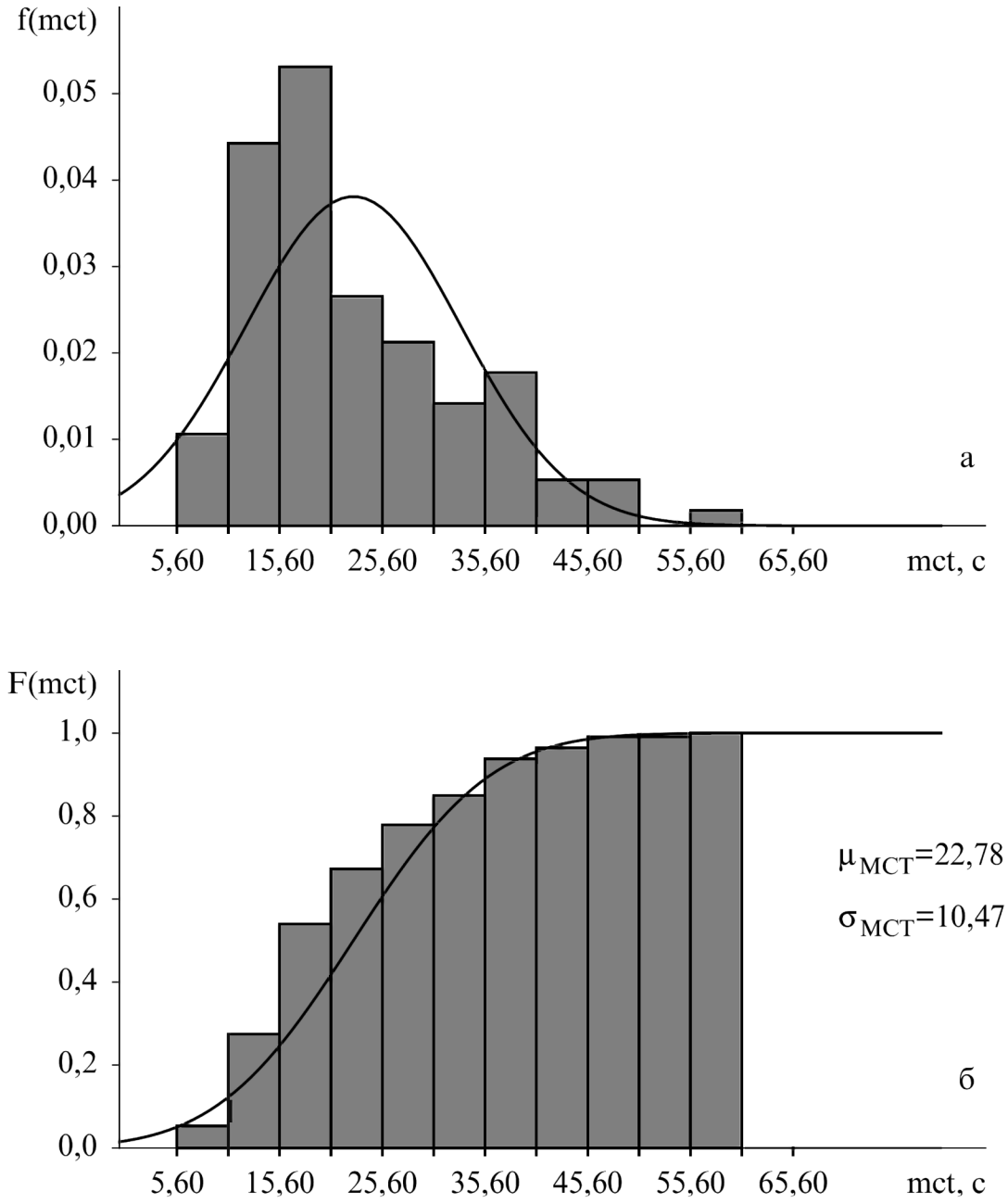


Рис. 10.8. Результати обробки вибіркових даних для прикл. 10.8: а – функція густини; б – функція розподілу для первинної змінної

Оскільки параметри первинної $X = MCT$ і перетвореної $X' = MCT' = \lg X$ змінних зв'язані співвідношенням

$$\lg \mu_{CI} = \mu_{CI'} + 1{,}15 \cdot \sigma^2_{CI'}, \quad \mu_{CI} = 10^{\mu_{CI'} + 1{,}15 \cdot \sigma^2_{CI'}},$$

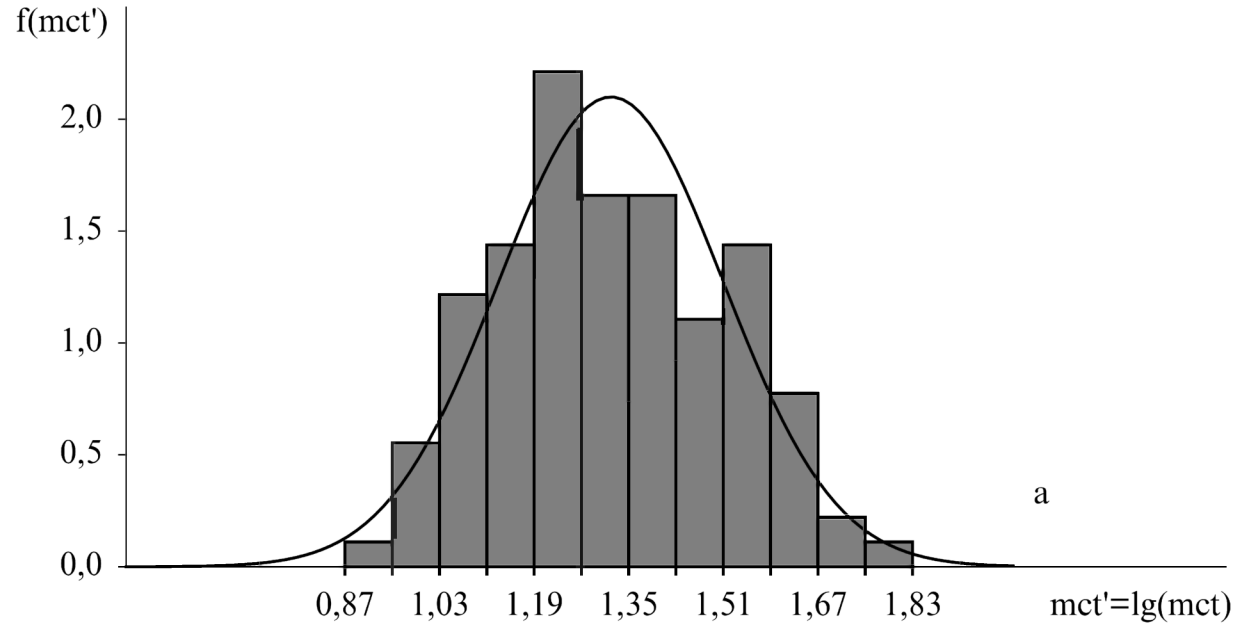
яке випливає безпосередньо з визначення математичного сподівання й дисперсії, то інтервальна оцінка може бути проведена й для первинної змінної

$$10^{0{,}29 + 1{,}15 \cdot 0{,}056} < \mu_{CI} < 10^{0{,}39 + 1{,}15 \cdot 0{,}096}, \qquad 2{,}24 < \mu_{CI} < 3{,}13.$$

Безпосереднє обчислення інтервальної оцінки для первинної змінної за (9.89) приводить до подвійної нерівності

$$2{,}56 - \frac{1{,}46}{\sqrt{113}} \cdot 1{,}98 < \mu_{CI} < 2{,}56 + \frac{1{,}46}{\sqrt{113}} \cdot 1{,}98, \qquad 2{,}28 < \mu_{CI} < 2{,}83,$$

що узгоджується з оцінкою, одержаною за допомогою перетворення змінної (див. попередню подвійну нерівність), і підтверджує стійкість інтервального оцінювання математичного сподівання для симетричного вибіркового розподілу. ●



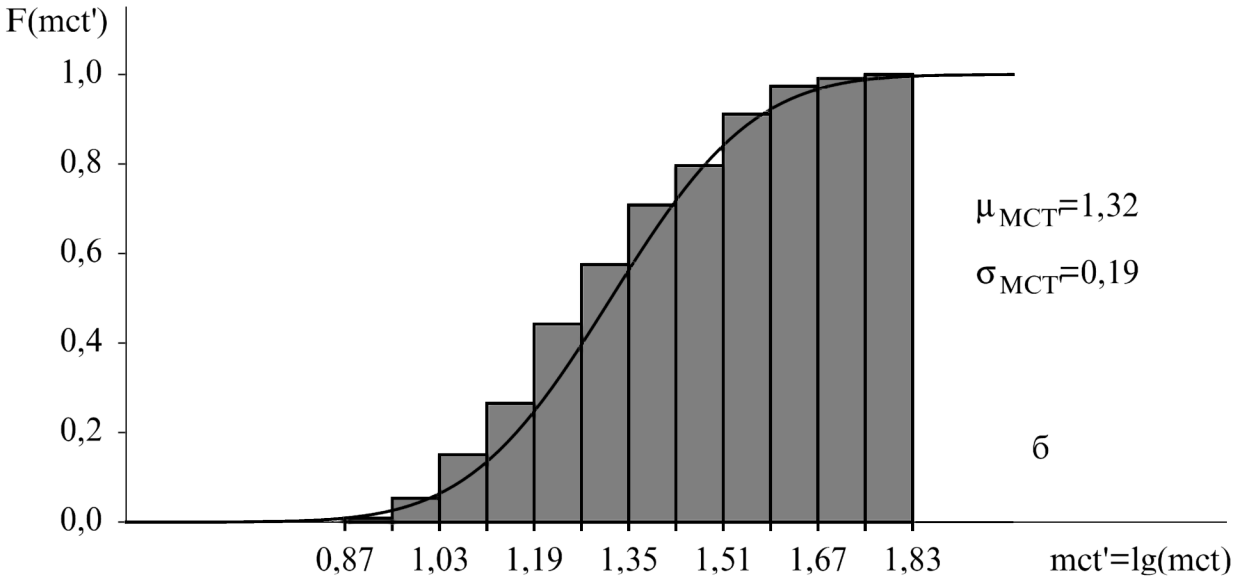


Рис. 10.9. Результати обробки вибіркових даних для прикл. 10.8: а – функція густини; б – функція розподілу для перетвореної змінної

10.5. Перевірка гіпотез про математичне сподівання й дисперсію неперервної випадкової величини

Якщо гіпотеза про нормальний розподіл неперервної випадкової величини X уже прийнята, то можна перевірити гіпотезу $H_0 : \{\mu_X = \mu_0\}$ про рівність математичного сподівання деякому значенню μ_0, відомому з яких-небудь міркувань, наприклад з дослідів, проведених іншим дослідником. Для цього досить перевірити, чи попадає значення μ_0 в надійний інтервал для математичного сподівання (9.25), (9.26), тобто перевірити справедливість подвійної нерівності

$$m - \frac{\hat{s}}{\sqrt{n}} t_{n-1,1-\alpha/2} < \mu_0 < m + \frac{\hat{s}}{\sqrt{n}} t_{n-1,1-\alpha/2}. \qquad (10.25)$$

Перевірку гіпотези можна виконати й у вигляді перевірки справедливості однобічної нерівності

$$|m - \mu_0| < \frac{\hat{s}}{\sqrt{n}} t_{n-1,1-\alpha/2}. \qquad (10.26)$$

Аналогічно перевіряється гіпотеза $H_0 : \{\sigma_X^2 = \sigma_0^2\}$ про рівність дисперсії деякому значенню σ_0^2. Для цього досить перевірити, чи попадає значення σ_0^2 в надійний інтервал для дисперсії (9.29), (9.30), тобто перевірити справедливість подвійної нерівності

$$\frac{\hat{s}^2(n-1)}{\chi^2_{n-1,\,1-\alpha/2}} < \sigma_0^2 < \frac{\hat{s}^2(n-1)}{\chi^2_{n-1,\,\alpha/2}}. \qquad (11.27)$$

Зауваження 10.6. Для даних вибіркових значень статистичні гіпотези $H_0 : \{\mu_X = \mu_0\}$ й $H_0 : \{\sigma_X^2 = \sigma_0^2\}$ будуть прийняті за неєдиного вибору значень μ_0 і σ_0^2. Спробуйте дати пояснення цього факту. ▲

10.6. Порівняння параметрів неперервних випадкових величин у дослідах „контроль група – дослідна група“

Розглянемо таку модельну ситуацію. Нехай є дві однотипні генеральні сукупності, одна з яких піддається деякому впливу, а інша – ні. При цьому в обох генеральних сукупностях можуть відбуватися подібні природні процеси, здатні змінювати характеристики об’єктів.

Припустимо, що для спостереження за поведінкою об'єктів обох генеральних сукупностей обрана деяка кількісна ознака. Позначимо її для генеральної сукуп-

ності, що не піддається впливу („контрольної“), через X, а для генеральної сукупності, що піддається впливу – через Y. Після завершення досліду з обох генеральних сукупностей витягаємо вибірки, загалом кажучи, різних обсягів n_X і n_Y. Вибіркові дані для двох генеральних сукупностей можна зобразити у вигляді записів в „лабораторному журналі“, як у табл. 10.6 і 10.7.

Таблиця 10.6

i	X
1	x_1
2	x_2
…	…
i	x_i
…	…
n_X	x_{n_X}

Таблиця 10.7

i	Y
1	y_1
2	y_2
…	…
i	y_i
…	…
n_Y	y_{n_Y}

У принципі можлива й інша організація досліду. З однієї генеральної сукупності витягаємо дві вибірки, причому не обов'язково одного обсягу. Першу вибірку спостерігаємо в природному стані („контрольна“), а на другу чинимо деякий вплив („дослідна“). Така організація може виявитися зручнішою, ніж у першому випадку. Однак для зручності можна вважати, що вони одержані з різних генеральних сукупностей, що не розрізняються за своїми властивостями.

Потрібно за двома наборами вибіркових даних (табл. 10.6 і 10.7) визначити, чи позначається даний вплив на ознаці, що досліджується. Основна гіпотеза в описаній постановці досліду формулюється мовою дослідника в такий спосіб:

$$H_0\text{: \{за досліджуваною ознакою виявити вплив не можна\}}. \quad (10.28)$$

У найбільш простій постановці гіпотези вважається, що розподіл ознаки в обох генеральних сукупностях є нормальний (див. розд. 6.3), тобто потрібно перевірити додаткові гіпотези про закони розподілу (див. розд. 10.4):

$$H_0 : \{X \sim N(\mu_X, \sigma_X), \quad \mu_X = m_X, \quad \sigma_X = \hat{s}_X\}, \quad (10.29)$$

$$H_0 : \{Y \sim N(\mu_Y, \sigma_Y), \quad \mu_Y = m_Y, \quad \sigma_Y = \hat{s}_Y\}, \quad (10.30)$$

Якщо гіпотези (10.29), (10.30) відхиляються, то перевірити гіпотезу (10.28) не можна. Якщо гіпотези приймаються, то хід подальших міркувань такий. Вплив не змінив у якісному відношенні закон розподілу (обидва розподіли підпорядковуються нормальному закону), але можливі кількісні зміни за рахунок зміни дисперсії (змінюється розсіювання ознаки, тобто ширина функції густини) або математичного сподівання (функція густини зсувається уздовж осі вправо або вліво). Тому потрібно перевірити додаткові гіпотези:

$$H_0 : \{\sigma_X^2 = \sigma_0^2\}, \tag{10.31}$$

$$H_0 : \{\mu_X = \mu_0\} . \tag{10.32}$$

Відзначимо, що попередні міркування про формулювання гіпотез (10.31), і (10.32) ґрунтуються на порівнянні незсунених вибіркових дисперсій $\hat{s}_X^2$ і $\hat{s}_Y^2$ вибіркових середніх значень m_X і m_Y. Ці значення були одержані ще до перевірки гіпотез (10.29), (10.30).

Перевірка гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$

Побудова критерію перевірки ґрунтується на тім, що:

1) для нормально розподілених випадкових величин X і Y з дисперсією σ^2 й незсуненими вибірковими дисперсіями $\hat{s}_X^2$ й $\hat{s}_Y^2$, обчисленими за вибірками обсягів n_X і n_Y, величини $(n_X - 1)\hat{s}_X^2 / \sigma^2$ й $(n_Y - 1)\hat{s}_Y^2 / \sigma^2$ мають розподіл Пірсона $\tilde{\chi}_\nu^2$ з числом степенів вільності $\nu_X = n_X - 1$ й $\nu_Y = n_Y - 1$ відповідно, тобто

$$\frac{(n-1)\tilde{\hat{s}}_X^2}{\sigma^2} \sim \tilde{\chi}_{n_X - 1}^2; \tag{10.33}$$

$$\frac{(n-1)\tilde{\hat{s}}_Y^2}{\sigma^2} \sim \tilde{\chi}_{n_Y - 1}^2; \tag{10.34}$$

2) у припущенні справедливості гіпотези H_0 відношення вибіркових дисперсій $\hat{s}_X^2$, $\hat{s}_Y^2$ для випадкових величин X і Y має розподіл Фішера F із числом степенів вільності ν_X чисельника й числом степенів вільності ν_Y знаменника, тобто

$$\frac{\tilde{\hat{s}}_X^2}{\tilde{\hat{s}}_Y^2} = \frac{\dfrac{\tilde{\chi}_{n_X - 1}^2 \sigma^2}{n_X - 1}}{\dfrac{\tilde{\chi}_{n_Y - 1}^2 \sigma^2}{n_Y - 1}} = \frac{\dfrac{\tilde{\chi}_{n_X - 1}^2}{n_X - 1}}{\dfrac{\tilde{\chi}_{n_Y - 1}^2}{n_Y - 1}} \sim F_{\nu_X, \nu_Y} . \tag{10.35}$$

Для заданого рівня значущості α можна побудувати надійний інтервал значень величини F_{ν_X, ν_Y}, яких вона набуває із ймовірністю $\beta = 1 - \alpha$:

$$F_{\nu_X, \nu_Y, \alpha/2} < F_{\nu_X, \nu_Y} < F_{\nu_X, \nu_Y, 1-\alpha/2} . \tag{10.36}$$

Якщо вибіркове значення $(F_{\nu_X,\nu_Y})^*$ попадає в надійний інтервал (10.36), то **гіпотеза** H_0 : $\{\sigma_x^2 = \sigma_y^2\}$ **приймається**, у протилежному випадку **гіпотеза відхиляється**.

Для скорочення обчислень звичайно користуються досить очевидною властивістю F-розподілу, що випливає безпосередньо з виразу для величини F. Саме якщо за обчислення вибіркового значення $(F_{\nu_X,\nu_Y})^*$ в чисельник (10.35) зліва від знака рівності поміщати більше значення, а в знаменник менше, то для прийняття гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ досить виконання однобічної нерівності

$$(F_{\nu_X,\nu_Y})^* < F_{\nu_X,\nu_Y,1-\alpha/2} \ . \tag{10.37}$$

Критерій F є нестійкий до відхилень розподілів досліджуваних величин від нормального закону. Тому його не можна застосувавати, якщо є сумніви щодо нормальності розподілів.

Перевірка гіпотези $H_0 : \{\mu_X = \mu_Y\}$

Побудова критерію перевірки гіпотези ґрунтується на тім, що:

1) неперервні випадкові величини X і Y розподілені нормально;

2) відомий результат перевірки гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$.

Якщо гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ прийнята, тобто $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, то незсунені вибіркові дисперсії $\hat{s}_X^2$ і $\hat{s}_Y^2$ є точковими оцінками одного і того ж параметра σ^2. Середньозваженою (узагальненою) точковою оцінкою параметра σ^2 за двома вибірками буде величина

$$\hat{s}^2 = \frac{\nu_X \hat{s}_X^2 + \nu_Y \hat{s}_Y^2}{\nu_X + \nu_Y}. \tag{10.38}$$

За її допомогою будується критерій перевірки гіпотези $H_0 : \{\mu_X = \mu_Y\}$ (двовибірковий t-критерій), заснований на статистиці

$$\frac{\tilde{m}_X - \tilde{m}_Y}{\tilde{\hat{s}}\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim \tilde{t}_\nu, \tag{10.39}$$

що має розподіл Стьюдента із числом степенів вільності $\nu = \nu_X + \nu_Y$.

Якщо гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ відхилена, то замість двовибіркового t-критерію потрібно застосовувати критерій Уелча, у якому статистикою є величина

$$\frac{\widetilde{m}_X - \widetilde{m}_Y}{\sqrt{\frac{\widetilde{\hat{s}}_X^2}{n_X} + \frac{\widetilde{\hat{s}}_Y^2}{n_Y}}} \sim \widetilde{t}_\nu, \tag{10.40}$$

що має t-розподіл Стьюдента із числом степенів вільності

$$\nu = \frac{\left(\frac{\hat{s}_X^2}{n_X} + \frac{\hat{s}_Y^2}{n_Y}\right)^2}{\left(\frac{\hat{s}_X^4}{n_X^2(n_X - 1)} + \frac{\hat{s}_Y^4}{n_Y^2(n_Y - 1)}\right)}. \tag{10.41}$$

Для заданого рівня значущості α можна побудувати надійний інтервал значень (10.39), (10.40), прийнятих із ймовірністю $\beta = 1-\alpha$:

$$t_{\nu,\alpha/2} < \widetilde{t}_\nu < t_{\nu,1-\alpha/2}. \tag{10.42}$$

Оскільки число степенів вільності ν (10.42) не обов'язково ціле, то для обчислення квантилів $t_{\nu,\alpha/2}$ і $t_{\nu,1-\alpha/2}$ варто використовувати інтерполяцію за значеннями квантилів для найближчих цілих значень ν.

Гіпотеза $H_0 : \{\mu_X = \mu_Y\}$ приймається або відхиляється, залежно від того, попадає чи не попадає у відповідний надійний інтервал (10.42) вибіркове значення t_ν^* (10.39) або (10.40).

Розглянуті t-критерії **є стійкі** до помірних відхилень досліджуваних розподілів від нормального закону.

Зауваження 10.7. У дослідах „контроль група – дослідна група“ прийняття гіпотези про рівність дисперсій $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ може бути звичайним явищем, а прийняття гіпотези про рівність математичних сподівань $H_0 : \{\mu_X = \mu_Y\}$ залежить від того, наскільки вдалим є вибір ознаки для вивчення даного впливу. Якщо ознака реагує на даний вплив, то гіпотеза $H_0 : \{\mu_X = \mu_Y\}$ має відхилятися. Що робити в цьому випадку? Відповідь проста – оцінити, наскільки в середньому відрізняється значення досліджуваної ознаки в генеральних сукупностях, що відповідають контрольній і дослідній групам. Для цього потрібно перевірити таку гіпотезу $H_0 : \{\mu_X - \mu_Y = \delta\}$. ▲

Зауваження 10.8. Відхилення додаткових гіпотез (10.29), (10.30) про закон розподілу ознаки до та після впливу не означає принципової неможливості перевірити гіпотезу(10.28). Розібраний нами метод перевірки гіпотези (10.28) належить до **параметричних,** тому що для нього потрібне виконання точкової оцінки

параметрів X і Y, тобто математичних сподівань μ_X і μ_Y та дисперсій σ_X^2 і σ_Y^2, що добре видно з формулювання гіпотез (10.30) і (10.31). У статистиці розроблені й інші методи перевірки гіпотез вигляду (10.29) та інших, називаних **непараметричними**. Ці методи потребують перевірки додаткових гіпотез про закон розподілу. Ми не будемо розглядати подібні методи перевірки статистичних гіпотез, обмежимося тільки параметричними. ▲

Перевірка гіпотези $H_0 : \{\mu_X - \mu_Y = \delta\}$

Для допоміжної випадкової величини $Z=X-Y$ можна використовувати статистику

$$\frac{(\mu_X - \mu_Y) - (\tilde{m}_X - \tilde{m}_Y)}{\tilde{\hat{s}}\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim \tilde{t}_\nu , \tag{10.43}$$

якщо гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ прийнята, і статистику

$$\frac{(\mu_X - \mu_Y) - (\tilde{m}_X - \tilde{m}_Y)}{\sqrt{\dfrac{\tilde{\hat{s}}_X^2}{n_X} + \dfrac{\tilde{\hat{s}}_Y^2}{n_Y}}} \sim \tilde{t}_\nu , \tag{10.44}$$

якщо гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ відхилена.

Для заданого рівня значущості α можна побудувати надійні інтервали значень величин (10.43) і (10.44):

$$\begin{gathered}(m_X - m_Y) - t_{\nu,1-\alpha/2}\,\hat{s}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} < \\ < \mu_X - \mu_Y < \\ < (m_X - m_Y) + t_{\nu,1-\alpha/2}\,\hat{s}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\end{gathered} , \tag{10.45}$$

$$\begin{gathered}(m_X - m_Y) - t_{\nu,1-\alpha/2}\sqrt{\frac{\hat{s}_X^2}{n_X} + \frac{\hat{s}_Y^2}{n_Y}} < \\ < \mu_X - \mu_Y < \\ < (m_X - m_Y) + t_{\nu,1-\alpha/2}\sqrt{\frac{\hat{s}_X^2}{n_X} + \frac{\hat{s}_Y^2}{n_Y}}\end{gathered} . \tag{10.46}$$

Для перевірки гіпотези $H_0 : \{ \mu_X - \mu_Y = \delta \}$ досить перевірити виконання нерівностей (10.45) або (10.46) для значення δ.

10.8. Порівняння параметрів безперервних випадкових величин у дослідах „до впливу – після впливу“

Розглянемо таку модельну ситуацію. Нехай деяка генеральна сукупність піддається певному впливу. Наприклад, для рослин таким впливом можуть бути добрива, гербіциди, пестициди, для тварин – зміни в харчуванні й т.д. Вплив на генеральну сукупність вважається відомим і сумнівів не викликає. Для дослідника питання полягає в тому, яка кількісна ознака „відгукується“ на даний вплив і наскільки вона змінюється. Якщо така задача буде розв’язана, то можна спробувати застосувати одержаний результат в оберненій постановці. Тобто за зміни значення даної ознаки робити висновок про певний тип і інтенсивність впливу на об'єкти генеральної сукупності.

Перейдемо до формального опису досліду. Нехай відгук об’єктів даної генеральної сукупності на певний вплив вивчається за деякою кількісною ознакою неперервної випадкової величини. Витягнемо з генеральної сукупності вибірку обсягу n. Позначимо ознаку до початку впливу (наприклад, у момент часу $t = t_1$) через X. Виміряємо значення ознаки X у вибіркових об'єктів. Потім об’єкти вибірки піддамо впливу. Позначимо ознаку після закінчення впливу (наприклад, у момент часу $t = t_2$) через Y. Виміряємо значення ознаки Y у вибіркових об'єктів.

За двома наборами вимірів значень ознак X і Y, зображеними у вигляді записів в лабораторному журналі (табл. 10.8) необхідно визначити, чи відгукується досліджуваний показник на вплив. Звернемо увагу на те, що кожний запис (рядок) в лабораторному журналі належить до одного об'єкта вибірки.

Таблиця 10.8

Дослід	X	Y
1	x_1	y_1
2	x_2	y_2
…	…	…
i	x_i	y_i
…	…	…
n	x_n	y_n

Зауваження 10.9. В принципі, можна уявити й іншу організацію досліду. Саме до початку впливу з генеральної сукупності витягається вибірка, вимірами досліджуваної ознаки у об’єктів вибірки заповнюється стовпчик X в лабораторному журналі, потім об’єкти повертаються в генеральну сукупність. Після закінчення впливу на генеральну сукупність ті ж самі об’єкти витягують (така можливість має бути забезпечена) і в них вимірюються значення досліджуваної ознаки, якими заповнюється стовпчик Y в лабораторному журналі. Така ор-

ганізація досліду може виявитися зручнішою, якщо, наприклад, вплив може спостерігатися тільки в природних умовах. ▲

Основна гіпотеза мовою дослідника формулюється в такий спосіб:

$$H_0\text{: \{за досліджуваною ознакою виявити вплив не можна\}}. \qquad (10.47)$$

У найбільш простому варіанті гіпотези вважається, що розподіл ознаки в обох генеральних сукупностях підпорядковується нормальному закону (див. розд. 6.3), тобто потрібно перевірити додаткові гіпотези про закон розподілу (див. розд. 6.4):

$$H_0 : \{X \sim N(\mu_X, \sigma_X), \quad \mu_X = m_X, \quad \sigma_X = \hat{s}_X\}, \qquad (10.48)$$

$$H_0 : \{Y \sim N(\mu_Y, \sigma_Y), \quad \mu_Y = m_Y, \quad \sigma_Y = \hat{s}_Y\}. \qquad (10.49)$$

Якщо гіпотези (10.48), (10.49) відхиляються, то перевірити гіпотезу (10.48) не можна (див. заув. 10.7 до розд. 10.7). Якщо гіпотези приймаються, то хід подальших міркувань такий. Вплив не змінив у якісному відношенні закон розподілу (обидва розподіли підпорядковуються нормальному закону), але можливі кількісні зміни за рахунок зміни дисперсії (змінюється розсіювання ознаки, тобто ширина функції густини) або (і) математичного сподівання (функція густини зрушується уздовж осі вправо або вліво). Тому потрібно перевірити додаткові гіпотези:

$$H_0 : \{\sigma_X^2 = \sigma_0^2\}, \qquad (10.50)$$

$$H_0 : \{\mu_X = \mu_0\}. \qquad (10.51)$$

Відзначимо, що попередні міркування про формулювання гіпотез (10.50), (10.51) ґрунтуються на порівнянні незсунених вибіркових дисперсій $\hat{s}_X^2$ і $\hat{s}_Y^2$ вибіркових середніх значень m_X і m_Y. Ці значення були одержані ще до перевірки гіпотез (10.48), (10.49).

Зауваження 10.10. Прийняття гіпотез (10.50), (10.51) зовсім не означає, що всі об'єкти генеральної сукупності не змінили значення досліджуваної ознаки під досліджуваним впливом, тобто „не відгукнулися“ на вплив. У одних об'єктів значення ознаки може понизитися, а в інших, навпаки, − підвищитися. Прийняття гіпотези говорить лише про те, що в середньому генеральна сукупність „не відгукнулася“ за досліджуваною ознакою на даний вплив. ▲

Перевірка гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$

Побудова критерію перевірки ґрунтується на тому, що:

1) випадкові величини X і Y розподілені нормально, причому математичні сподівання μ_X та μ_Y можуть бути невідомі;

2) статистика

$$\frac{\tilde{\hat{s}}_X^2 - \tilde{\hat{s}}_Y^2}{\sqrt{\frac{4\tilde{\hat{s}}_X^2 \tilde{\hat{s}}_Y^2}{n-2}\left(1-\tilde{r}_{X,Y}^2\right)}} \sim \tilde{t}_\nu \qquad (10.52)$$

розподілена за законом Стьюдента із числом степенів вільності $\nu = n-2$, $r_{X,Y}$ – вибірковий коефіцієнт кореляції (точкова оцінка коефіцієнта кореляції $\rho_{X,Y}$).

Для заданого рівня значущості α можна побудувати надійний інтервал значень величини $\tilde{t}_{n-2}$, яких вона набуває з імовірністю $\beta = 1-\alpha$:

$$t_{n-2,\alpha/2} < \tilde{t}_{n-2} < t_{n-2,1-\alpha/2}, \qquad (10.53)$$

Якщо вибіркове значення $(t_{n-2})^*$ (10.53) потрапляє в надійний інтервал (10.53), то **гіпотеза** $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ **приймається**, у протилежному випадку **гіпотеза відхиляється**.

Для скорочення обчислень звичайно послуговуються симетричністю розподілу Стьюдента. Саме для прийняття гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ досить виконання однобічної нерівності

$$\left| t_{n-2}^* \right| < t_{n-2,1-\alpha/2}. \qquad (10.54)$$

Зауваження 10.9. Той факт, що вибіркові дисперсії $\hat{s}_X^2$ і $\hat{s}_Y^2$ можуть відрізнятися, ще не доводить, що гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ правильна чи ні. Гіпотеза, що перевіряється, може бути сформульована в іншій формі, що проясняє суть питання. Це формулювання таке. Якщо неперервні випадкові величини X і Y розподілені нормально, причому $\sigma_X^2 = \sigma_Y^2$, то чи може бути так, що в деякій частці $\beta=1-\alpha$ великої серії вибірок значення незсунених вибіркових дисперсій $\hat{s}_X^2$ і $\hat{s}_Y^2$ відрізняються так само, як і в даній вибірці. ▲

Перевірка гіпотези $H_0 : \{\mu_X = \mu_Y\}$

Побудова критерію перевірки гіпотези ґрунтується на тому, що:

1) випадкові величини X і Y розподілені нормально;

2) відомий результат перевірки гіпотези $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$.

Якщо гіпотеза $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ прийнята, тобто $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, то вихідна гіпотеза формулюється у вигляді $H_0 : \{\mu_Z = \mu_Y - \mu_X = 0\}$ для допоміжної непе-

рервної випадкової величини $Z = X - Y$. Статистикою критерію перевірки гіпотези H_0 служить величина

$$\frac{\widetilde{m}_Z}{\widetilde{\widehat{s}}_Z / \sqrt{n}} \sim \widetilde{t}_\nu, \tag{10.55}$$

розподілена за законом Стьюдента із числом степенів вільності $\nu = n - 1$, де m_Z і $\widehat{s}_Z^2$ – відповідно вибіркове середнє та вибіркова дисперсія величини Z:

$$m_Z = \frac{1}{n}\sum_{i=1}^{n}((y_i - x_i)), \tag{10.56}$$

$$\widehat{s}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(y_i - x_i) - m_Z]^2. \tag{10.57}$$

Для заданого рівня значущості α можна побудувати надійний інтервал значень величини (10.55), яких вона набуває з імовірністю $\beta = 1-\alpha$:

$$t_{n-1,\,\alpha/2} < \widetilde{t}_{n-2} < t_{n-1,\,1-\alpha/2}. \tag{10.58}$$

Якщо вибіркове значення t_{n-2}^* (10.55) попадає в надійний інтервал (10.58), то **гіпотеза** $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ **приймається**, у протилежному випадку **гіпотеза відхиляється**.

Розглянутий t-критерій перевірки гіпотези $H_0 : \{\mu_X = \mu_Y\}$ **є стійкий** до помірних відхилень досліджуваних розподілів від нормального закону.

Зауваження 10.12. У дослідах „до впливу – після впливу“ прийняття гіпотези про рівність дисперсій $H_0 : \{\sigma_X^2 = \sigma_Y^2\}$ може бути звичайним явищем, а прийняття гіпотези про рівність математичних сподівань $H_0 : \{\mu_X = \mu_Y\}$ залежить від того, наскільки вдалим є вибір ознаки для вивчення даного впливу. Якщо ознака відгукується на даний вплив, то гіпотеза $H_0 : \{\mu_X = \mu_Y\}$ відхиляється. Що робити в цьому випадку? Відповідь проста – оцінити, наскільки в середньому відрізняється значення досліджуваної ознаки в генеральних сукупностях, що відповідають контрольній і дослідній групам. Для цього потрібно перевірити таку гіпотезу $H_0 : \{\mu_X - \mu_Y = \delta\}$. ▲

Перевірка гіпотези $H_0 : \{\mu_X - \mu_Y = \delta\}$

Для допоміжної випадкової величини $Z = X - Y$ можна використовувати статистику (10.53), що дозволяє побудувати надійний інтервал для не рівного нулю математичного сподівання $\mu_Z = \mu_X - \mu_Y$:

$$m_Z - t_{n-1,\,1-\alpha/2}\frac{\hat{s}_Z}{\sqrt{n}} < \mu_X - \mu_Y < m_Z + t_{n-1,\,1-\alpha/2}\frac{\hat{s}_Z}{\sqrt{n}}, \qquad (10.59)$$

Потім потрібно перевірити, чи належить надійному інтервалу (10.59) задане значення δ.

10.8. Гіпотези про коефіцієнт кореляції

Гіпотеза про коефіцієнт кореляції $H_0 : \{\rho_{X,Y} = 0\}$

Вибірковий коефіцієнт кореляції $r_{X,Y}$ є дуже нестійка точкова оцінка коефіцієнта кореляції $\rho_{X,Y}$. Це означає, що за невеликого обсягу вибірки n навіть помітна відмінність від нуля значення вибіркового коефіцієнта кореляції $r_{X,Y}$ може виявитися статистично незначущою (тобто за значенням $r_{X,Y}$ не можна зробити висновок про наявність кореляції). Для надійних оцінок коефіцієнта кореляції $\rho_{X,Y}$, як правило, необхідно кілька десятків або навіть сотень дослідів. Тому потрібно з'ясувати, чи достатня величина $r_{X,Y}$ для обґрунтованого висновку про наявність кореляційного зв'язку. Для цього перевіряється **гіпотеза про рівність коефіцієнта кореляції нулю** $H_0 : \{\rho_{X,Y} = 0\}$. **Статистикою критерію перевірки** в припущенні **про нормальність розподілів неперервних випадкових величин** X і Y є така величина (див. вибірковий розподіл 12 у розд. 8.6 і властивості коефіцієнта кореляції в розд. 7.6.)

$$\sqrt{n-2}\,\frac{\tilde{r}_{X,Y}}{\sqrt{1-\tilde{r}^2_{X,Y}}} \sim \tilde{t}_{n-2}, \qquad (10.60)$$

що має розподіл Стьюдента із числом степенів вільності $\nu = n-2$. Отже, для заданого рівня значущості α можна побудувати надійний інтервал значень величини (10.60), яких вона набуває з імовірністю $\beta = 1-\alpha$:

$$t_{n-2,\,\alpha/2} < \tilde{t}_{n-2} < t_{n-2,\,1-\alpha/2}. \qquad (10.61)$$

Якщо вибіркове значення t^*_{n-2} статистики (10.61) попадає в надійний інтервал (10.61), то **гіпотеза** $H_0 : \{\rho_{X,Y} = 0\}$ **приймається**, у протилежному випадку **гіпотеза** H_0 **відхиляється**, і **коефіцієнт кореляції статистично значимо відрізняється від нуля**.

Зважаючи на симетрію розподілу Стьюдента, достатньо перевірити виконання однобічної нерівності

$$\left| t^*_{n-2} \right| < t_{n-2,\,1-\alpha/2}. \qquad (10.62)$$

Розглянутий t-критерій перевірки гіпотези $H_0 : \{\rho_{X,Y} = 0\}$ **є стійкий** до помірних відхилень досліджуваних розподілів від нормального закону.

Зауваження 10.13. Перевірка гіпотези $H_0 : \{\rho_{X,Y} = 0\}$, як це випливає з (10.60), може бути здійснена й безпосередньо за вибірковим значенням $r_{X,Y}$. Тобто **гіпотеза** $H_0 : \{\rho_{X,Y} = 0\}$ **приймається**, якщо

$$\left| r_{X,Y} \right| < \sqrt{\frac{t_{n-2,1-\alpha/2}}{n-2+t_{n-2,1-\alpha/2}}}, \tag{10.63}$$

у протилежному випадку **гіпотеза** H_0 **відхиляється**. ▲

Приклад 10.7 (продовж. прикл. 10.5 і 10.6). Кореляційне поле для вибіркових значень $X = SP$ і $Y = DP$ (рис.10.10) указує на лінійний зв'язок між ними. Вибіркове значення коефіцієнта кореляції $r_{XY} \approx 0{,}82$. Перевіряючи гіпотезу $H_0 : \{\rho_{X,Y} = 0\}$, враховуємо, що величини X і Y розподілені за нормальним законом. Вибіркове значення статистики (10.60) $t^*_{n-2} \approx 14{,}86$ у порівнянні зі значенням квантиля $t_{n-2,1-\alpha/2} \approx 0{,}975$ дозволяє відхилити гіпотезу про рівність нулю коефіцієнта кореляції. ●

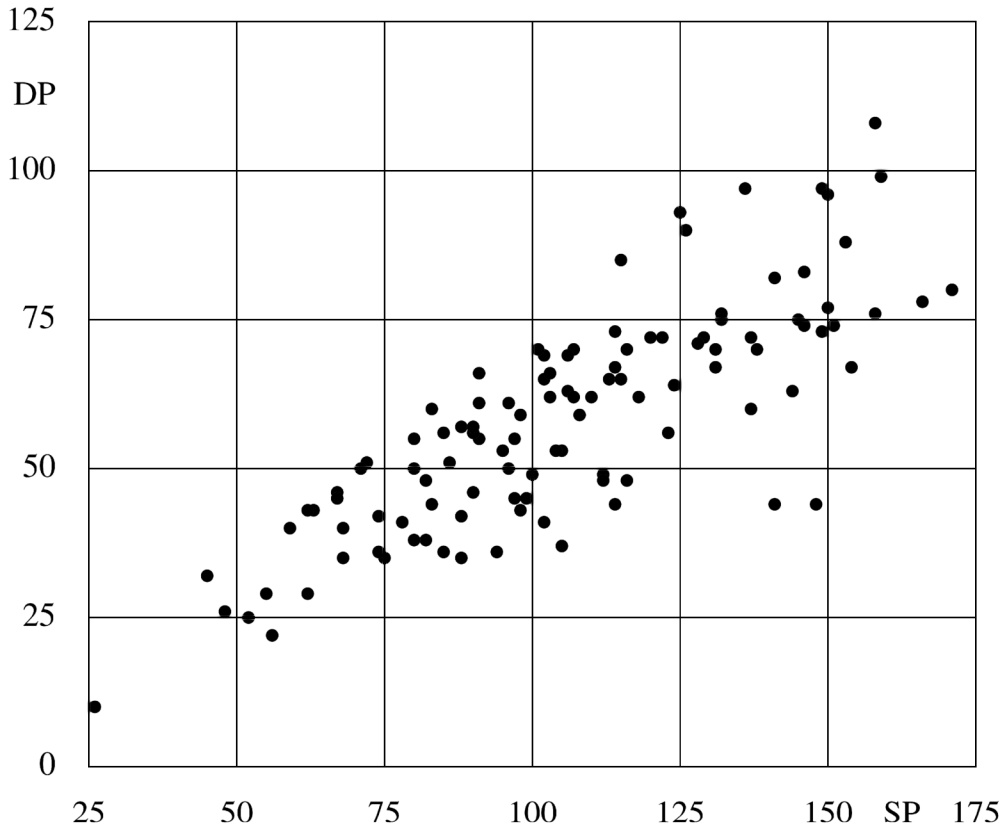


Рис. 10.10. Кореляційне поле між величинами SP і DP (прикл. 10.7)

Приклад 10.10 (продовж. прикл. 10.7 і 10.8). Кореляційне поле для величин $X = MCT$ і $Y = CI$ вказує на нелінійний зв'язок (рис. 10.11), отже, коефіцієнт кореляції не може бути застосований для вимірювання зв'язку між X і Y. До того ж величини X і Y не розподілені за нормальним законом і навіть за наявністю лінійного зв'язку інтервальна оцінка коефіцієнта кореляції була б неможливою.

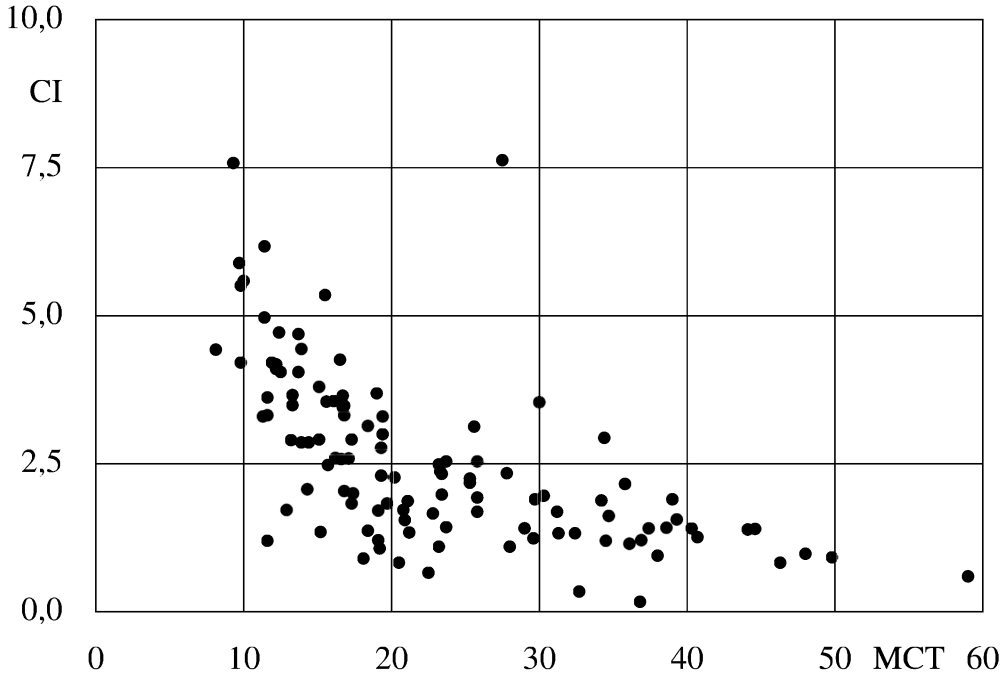


Рис. 10.11. Кореляційне поле між первинними величинами MCT і CI (прикл. 10.8)

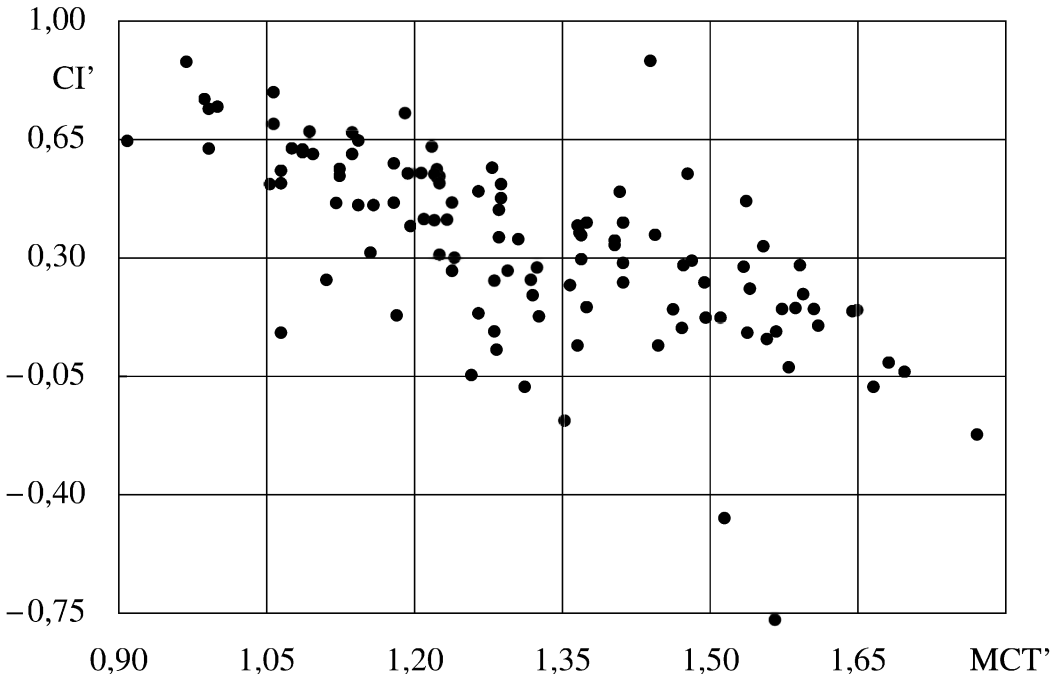


Рис. 10.12. Кореляційне поле між перетвореними величинами MCT' і CI' (прикл. 10.8)

За допомогою перетворень $X' = MCT' = \lg X$ і $Y' = CI' = \lg YI$ досліджувані величини нормалізуться, а зв'язок між ними, як вказує кореляційне поле (рис. 10.12), можна вважати лінійним. Проте значення вибіркового коефіцієнта кореляції $r_{\lg X, \lg Y} = r_{X',Y'} \approx 0{,}66$ незначно відрізняється від значення вибіркового коефіцієнта кореляції $r_{X,Y} \approx 0{,}6$.

Вибіркове значення статистики $t^*_{n-2} \approx 9{,}256$ у порівнянні зі значенням квантиля $t_{n-2,\,1-\alpha/2} \approx 0{,}975$ дозволяє відхилити гіпотезу про рівність нулю коефіцієнта кореляції. ●

Приклад 7.4 (продовження). Значення вибіркового коефіцієнта кореляції $r_{X,Y} \approx 0{,}82$, $\frac{1}{2}\ln[(1 + r_{X,Y})/(1 - r_{X,Y})] \approx 1{,}1568$; для рівня значущості α=0,05 (β=0,95) значення квантиля стандартизованого нормального розподілу $z_{1-\alpha/2} = z_{0,975} \approx 1{,}96$, тоді

$$y_{\alpha/2} = y_{0,025} = \frac{1}{2}\ln\frac{1 + r_{X,Y}}{1 - r_{X,Y}} - \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \approx 0{,}9708,$$

$$y_{1-\alpha/2} = y_{0,975} = \frac{1}{2}\ln\frac{1 + r_{X,Y}}{1 - r_{X,Y}} + \frac{z_{1-\alpha/2}}{\sqrt{n-3}} \approx 1{,}3428,$$

$$\operatorname{th} y_{0,025} \approx 0{,}7491, \quad \operatorname{th} y_{0,975} \approx 0{,}8723.$$

Отже, $0{,}7491 < \rho_{X,Y} < 0{,}8723$. ●

Приклад 7.5 (продовження). $r_{\lg X, \lg Y} \approx 0{,}66$; $y_{0,025} \approx 0{,}9728$, $y_{0,975} \approx 0{,}6068$; $\operatorname{th} y_{0,025} \approx 0{,}7526$, $\operatorname{th} y_{0,975} \approx 0{,}5419$; $0{,}5419 < \rho_{\lg X, \lg Y} < 0{,}7526$. ●

Гіпотеза про коефіцієнт кореляції $H_0 : \{\rho_{X,Y} = \rho_0\}$

За відхилення гіпотези $H_0 : \{\rho_{X,Y} = 0\}$ може бути перевірена гіпотеза $H_0 : \{\rho_{X,Y} = \rho_0 \neq 0\}$ – досить переконатися в тому, що значення ρ_0 належить надійному інтервалу (8.56).

11.10. Гіпотези про коефіцієнти лінійної регресії

Вибірковий коефіцієнт регресії b_1 є точкова оцінка **коефіцієнта регресії** β_1, тому необхідно з'ясувати, чи достатня точкова оцінка для статистично обґрунто-

ваного висновку про те, що коефіцієнт регресії β_1 відмінний від нуля. Для цього перевіряється гіпотеза $H_0 : \{\beta_1 = 0\}$. Якщо в результаті перевірки гіпотеза підтверджується, то величини X і Y залежні.

Підставою для побудови **критерію перевірки гіпотези** служить **припущення про нормальність розподілу величини** Y, звідки випливає, що величина $n\tilde{s}_d^2 / \sigma^2$ має **розподіл Пірсона** $\tilde{\chi}_\nu^2$ із числом степенів вільності $\nu = n - 2$, де s_d^2 – **дисперсія оцінки**:

$$s_d^2 = \frac{1}{n-2}\sum_{i=1}^{n} d_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 , \qquad (10.64)$$

а $s_d = \sqrt{s_d^2}$ – стандартна помилка оцінки.

Статистикою критерію служить величина

$$\frac{\sum_{i=1}^{n}(\tilde{\hat{y}}_i - \tilde{m}_Y)^2}{\tilde{s}_d^2} \sim \tilde{F}_{\nu_1,\nu_2} , \qquad (10.65)$$

розподілена за законом Фішера $\tilde{F}_{\nu_1,\nu_2}$ із числом степенів вільності чисельника $\nu_1 = 1$ і знаменника $\nu_2 = n - 2$. Отже, для заданого рівня значущості α можна побудувати надійний інтервал значень величини (10.64), яких вона набуває з ймовірністю $\beta = 1-\alpha$:

$$F_{\nu_1,\nu_2,\alpha/2} < F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2,1-\alpha/2} , \qquad (10.66)$$

Якщо вибіркове значення $(F_{\nu_1,\nu_2})^*$, обчислене за (10.65), попадає в надійний інтервал (10.66), то **гіпотеза** $H_0 : \{\beta_1 = 0\}$ **приймається**, у протилежному випадку **гіпотеза відхиляється**. У випадку відхилення гіпотези $H_0 : \{\beta_1 = 0\}$ говорять, що **коефіцієнт регресії статистично значимо відрізняється від нуля**.

F-критерій **є нестійкий** до відхилень досліджуваних розподілів від нормального.

Гіпотези про значення параметрів лінійної регресії

Якщо є підстави припускати, що значення параметрів лінійної регресії дорівнюють деяким значенням, відповідно $\beta_0^{(0)}$ й $\beta_1^{(0)}$, то варто перевірити гіпотези

$H_0 : \{\beta_0 = \beta_0^{(0)}\}$ й $H_0 : \{\beta_1 = \beta_1^{(0)}\}$. Для заданого рівня значущості **гіпотези приймаються**, якщо значення $\beta_0^{(0)}$ й $\beta_1^{(0)}$ попадають у надійні інтервали (8.67), (8.68), у протилежному випадку **гіпотези відхиляються**.

Приклад 10.7 (продовження). Обчислені за (7.68) коефіцієнти лінійної регресії такі: $b_0 = 6{,}3$, $b_1 = 0{,}49$. Відповідна лінія регресії зображена на рис. 10.13. Інтервальні оцінки коефіцієнтів визначені за (8.67), (8.68) для усталеного рівня значущості $\alpha = 0{,}05$: $-2{,}01 \le b_0 \le 14{,}6$, $0{,}42 \le b_1 \le 0{,}57$. Отже, якщо за результатами інших дослідів обчислені коефіцієнти b_0 і b_1 лінійної регресії, які відрізняються від даних, але містяться в наведених вище надійних інтервалах для даного досліду, можна вважати, що ці досліди статистично узгоджуються.

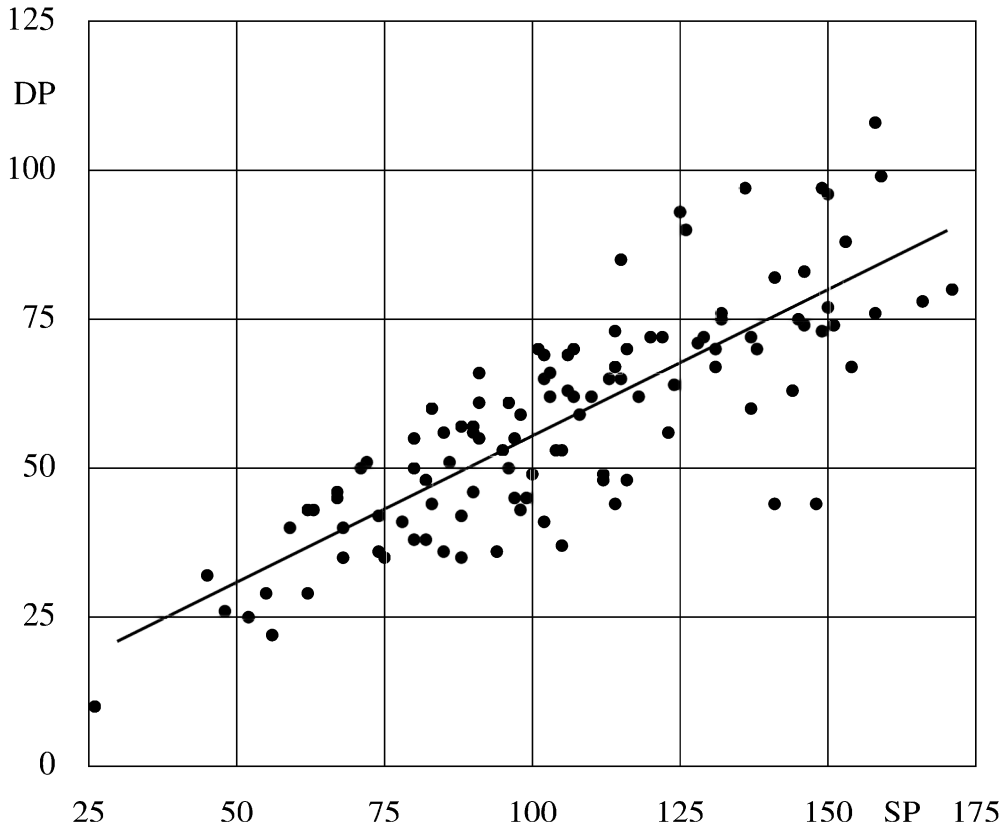


Рис. 10.13. Лінійна регресія між величинами SP і DP (прикл. 10.7)

Значення квантилів $\chi^2_{\nu,1-\alpha}$ розподілу Пірсона Таблиця Д. 1

Число степенів вільності ν	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
35	57,3	53,2	49,8	22,5	20,6	18,5
40	63,7	59,3	55,8	26,5	24,4	22,2
45	70,0	65,4	61,7	30,6	28,4	25,9
50	76,2	71,4	67,5	34,8	32,4	29,7
75	106,4	100,8	96,2	56,1	52,9	49,5
100	135,6	129,6	124,3	77,9	74,2	70,1

Значення квантилів $F_{\nu_1,\nu_2,1-\alpha/2}$ розподілу Фішера для рівня значущості α =0,05 Таблиця Д. 2

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	993,1	1001	1018
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,08	13,90
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,46	8,26
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,23	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,07	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,36	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,89	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,56	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,31	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,12	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	2,96	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,84	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,73	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,64	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,57	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,50	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,44	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,39	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,35	2,09

Закінчення табл. Д. 2

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,31	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,27	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,24	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,21	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,18	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,16	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,13	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,11	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,09	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,07	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	1,94	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,82	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,69	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,57	1,00

Значення квантилів $t_{\nu,1-\alpha/2}$ розподілу Стьюдента Таблиця Д. 3

Число степенів вільності	Рівень значущості α					
ν	0,10	0,05	0,02	0,01	0,002	0,001
1	6,314	12,706	31,821	63,657	318,309	636,619
2	2,920	4,303	6,965	9,925	22,327	31,598
3	2,353	3,182	4,541	5,841	10,215	12,941
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,869
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,408
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,610	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,552	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,767
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
∞	1,645	1,960	2,326	2,576	3,090	3,291

Значення перетворення Фішера *y(r)* Таблиця Д. 4

r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1104	0,1206	0,1307	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Значення функції th *y* Таблиця Д. 5

y	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0,1391	0,1489	0,1586	0,1684	0,1781	0,1877
0,2	0,1974	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821
0,3	0,2913	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714
0,4	0,3799	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4699	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0,5511	0,5581	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0,6351	0,6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969
1,1	0,8005	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8275	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8692	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033
1,5	0,9051	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9201
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9341
1,7	0,9354	0,9366	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,8	0,9468	0,9478	0,9488	0,9498	0,9508	0,9517	0,9527	0,9536	0,9545	0,9554
1,9	0,9562	0,9571	0,9579	0,9587	0,9595	0,9603	0,9611	0,9618	0,9626	0,9633
2,0	0,9640	0,9647	0,9654	0,9661	0,9667	0,9674	0,9680	0,9687	0,9693	0,9699
2,1	0,9705	0,9710	0,9716	0,9721	0,9727	0,9732	0,9737	0,9743	0,9748	0,9753
2,2	0,9757	0,9762	0,9767	0,9771	0,9776	0,9780	0,9785	0,9789	0,9793	0,9797
2,3	0,9801	0,9805	0,9809	0,9812	0,9816	0,9820	0,9823	0,9827	0,9830	0,9833
2,4	0,9837	0,9840	0,9843	0,9846	0,9849	0,9852	0,9855	0,9858	0,9861	0,9863
2,5	0,9866	0,9869	0,9871	0,9874	0,9876	0,9879	0,9881	0,9884	0,9886	0,9888
2,6	0,9890	0,9892	0,9895	0,9897	0,9899	0,9901	0,9903	0,9905	0,9906	0,9908
2,7	0,9910	0,9912	0,9914	0,9915	0,9917	0,9919	0,9920	0,9922	0,9923	0,9925
2,8	0,9926	0,9928	0,9929	0,9931	0,9932	0,9933	0,9935	0,9936	0,9937	0,9938
2,9	0,9940	0,9941	0,9942	0,9943	0,9944	0,9945	0,9946	0,9947	0,9949	0,9950

Список рекомендованої літератури

Климов, Ю.И. Является ли вероятность “нормальной” физической величиной? [Текст] / Ю.И. Климов // Успехи физических наук. – 1992. – Т. 162, № 7. – С. 149. – 182; http://www.ufn.ru.

Аптон, Г. Анализ таблиц сопряженности [Текст]: Пер. с англ. / Г. Аптон. – М.: Финансы и статистика, 1982. – 143 с.

Артемьева, Е.Ю. Сборник задач по теории вероятностей и математической статистике для психологов [Текст] / Е.Ю. Артемьева. – М.: Изд-во МГУ, 1969. – 92 с.

Архипов, С. Еще раз о законе Бенфорда [Текст] / С. Архипов // Техника – молодежи. – 1980. – № 8. – С. 61.

Афифи, А. Статистический анализ: Подход с использованием ЭВМ [Текст]: Пер. с англ. / А. Афифи, С. Эйзен. – М.: Мир, 1982. – 488 с.

Биометрия [Текст] / Н.В. Глотов, Л.А. Животовский, Н.В. Хованов, Н.Н. Хромов-Борисов. – Л.: Изд-во ЛГУ, 1982. – 263 с.

Борщ, В.Л. Збірник задач з математичної статистики (для студентів біолого-медичного інституту) [Текст] / В.Л. Борщ, Ю. П.Совіт. – Д.: ДДУ, 1997. – 80 с.

Борщ, В.Л. Математичні методи в біології в прикладах і задачах [Текст] / В.Л. Борщ, Ю.П. Совіт. – Д.: ДДУ, 1998. – 80 с.

Бронштейн, И.Н.. Справочник по математике для инженеров и учащихся втузов [Текст] / И.Н. Бронштейн, К.А. Семендяев. – М.: Наука, 1986. – 544 с.

Владимирский, Б.М. Математические методы в биологии [Текст] / Б.М. Владимирский. – Ростов: Изд-во Рост. ун-та, 1983. – 304 с.

Герасимович, А.И. Математическая статистика [Текст] / А.И. Герасимович. – Минск: Высшэйш. шк., 1983. – 279 с.

Глас, Дж. Статистические методы в педагогике и психологии [Текст] : Пер. с англ. / Дж. Гласс, Дж. Стэнли. – М.: Прогресс, 1976. – 495 с.

Глотов, Н.В. Сборник задач по биометрии [Текст] / Н.В. Глотов, А.А. Филатов , Н.Н. Хромов–Борисов. – Л.: Изд-во ЛГУ, 1985. – 81 с.

Джонсон, Н. Статистика и планирование эксперимента в технике и науке. Методы обработки данных [Текст]: Пер. с англ. / Н. Джонсон, Ф. Лион. – М.: Мир, 1980.

Дубровский, В.Н. Как возникает распределение Пуассона [Текст] / В.Н. Дубровский // Квант. – 1988. №8. – С. 23 – 25.

Жалдак, М.І. Теорія ймовірностей і математична статистика з елементами інформаційної технології [Текст] / М.І. Жалдак, Н. М.Кузьміна, С.Ю. Берлінська. – К.: Вища шк., 1995. – 351 с.

Каминский, Л.С. Медицинская и демографическая статистика [Текст] / Л.С. Каминский. – М.: Статистика, 1973. – 351 с.

Лакин, Г.Ф. Биометрия [Текст] / Г.Ф. Лакин. – М.: Высш. шк., 1980. – 293 с.

Математическая статистика [Текст] / В.Н. Иванова, В. Н.Калинина, Л.А. Нешумова, И.С. Решетникова. – М.: Высш. шк., 1981. – 327 с.

Мизес, Р. Вероятность и статистика [Текст] / Р. Мизес. – М.; Л.: Госиздат, 1930. – 250 с.

Плохинский, Н.А. Биометрия [Текст] / Н.А. Плохинский. – М.: Изд-во МГУ, 1970. – 367 с.

Рокицкий, П.Ф. Биологическая статистика [Текст] / П.Ф. Рокицкий. – Мн.: Вышэйш. шк., 1973. – 320 с.

Снедекор, Дж.У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии [Текст] / Дж.У. Снедекор. – М.: Изд-во с.-х. лит., 1961. – 503 с.

Суходольский, Т.Р. Основы математической статистики для психологов [Текст] / Т.Р. Суходольский. – Л.: Изд-во ЛГУ, 1972. – 429 с.

Терентьев, П.В. Практикум по биометрии [Текст] / П.В. Терентьев, Н.С. Ростова. – Л.: Изд-во ЛГУ, 1977. – 152 с.

Тернер, Д. Вероятность, статистика и исследование операций [Текст] / Д. Тернер. – М.: Статистика, 1976. – 431 с.

Турчин, В.М. Математическая статистика в примерах и задачах [Текст]: Учеб. Пособие / В. М.Турчин. – Д.: Изд-во ДГУ, 1992. – 164 с.

Турчин, В. М. Математическая статистика. Лабораторный практикум [Текст]: Учеб. Пособие / В.М. Турчин. – Д.: Изд-во ДГУ, 1992. – 184 с.

Турчин, В.М. Математична статистика в прикладах та задачах [Текст] / В.М. Турчин. – Д.: Вид-во ДДУ, 1992. – 164 с.

Турчин, В.М. Теория вероятностей и математическая статистика [Текст]: Учеб. пособие: В 2 т. / В.М. Турчин. – Д.: Изд-во ДГУ, 1996. – Т. 2: Математическая статистика. – 304 с.

Тутубалин, В.Н. Вероятность, компьютеры и обработка результатов эксперимента [Текст] / В.Н. Тутубалин // Успехи физ. наук. – 1993. – Т. 163, № 7. – С. 93 – 109; http://www.ufn.ru.

Тьюки, Дж. Анализ результатов наблюдений: Разведочный анализ [Текст]: Пер. с англ. / Дж. Тьюки. – М.: Мир, 1981. – 693 с.

Урбах, В.Ю. Статистический анализ в биологических и медицинских исследованиях [Текст] / В.Ю. Урбах. – М.: Медицина, 1975. – 294 с.

Феллер, В. Введение в теорию вероятностей и ее приложения [Текст]: В 2 т; Пер. с англ. / В. Феллер. – М.: Мир. – Т.1. – 1984. – 528 с.

Фишер, Р.А. Статистические методы для исследователей [Текст] / Р.А. Фишер. – М.: Госстатиздат. –1958. – 268 с.

Хинчин, А.Я. Учение Мизеса о вероятностях и принципы физической статистики [Текст] / А.Я. Хинчин // Успехи физ. наук. – 1929. – Т. 9, вып. 2. – С. 141 – 166; http://www.ufn.ru.

Чукова, Ю.П. Распределение Пуассона [Текст] / Ю.П. Чукова // Квант. – 1988. – № 8. – С. 15 – 18.

Зміст

Передмова 3

7. Системи двох випадкових величин 5
7.1. Поняття системи випадкових величин 5
7.2. Спільний розподіл двох дискретних випадкових величин 5
7.3. Незалежні дискретні випадкові величини 10
7.4. Дії над дискретними випадковими величинами 16
7.5. Спільний розподіл двох інтервальних і неперервних випадкових величин 25
7.6. Дії над інтервальними та неперервними випадковими величинами 30
7.7. Регресія 36
7.8. Кореляція 44

8. Неперервні випадкові величини, похідні від нормально розподіленої випадкової величини 53
8.1. Ще одне визначення нормального розподілу 53
8.2. Розподіл Пірсона $\tilde{\chi}^2$ 59
8.3. Розподіл Фішера F 61
8.4. Розподіл Стьюдента t 63
8.5. Вибіркові розподіли 63
8.6. Взаємозв'язки між розподілами дискретних і неперервних випадкових величин 70

9. Оцінювання параметрів випадкових величин 74
9.1. Точкове оцінювання параметрів неперервної випадкової величини 74
9.2. Властивості вибіркового середнього значення як точкової оцінки математичного сподівання 75
9.3. Властивості вибіркової дисперсії як точкової оцінки дисперсії 75
9.4. Інтервальне оцінювання параметрів неперервної випадкової величини 78
9.5. Інтервальне оцінювання математичного сподівання нормально розподіленої випадкової величини 79
9.6. Інтервальне оцінювання дисперсії нормально розподіленої випадкової величини 82
9.7. Інтервальна оцінка коефіцієнта кореляції нормально розподілених випадкових величин 83
9.8. Інтервальна оцінка параметрів лінійної регресії 87
9.9. Інтервальне оцінювання параметрів дискретної випадкової . . . . . величини 88
9.10. Необхідний обсяг вибірки 90

10. Перевірка статистичних гіпотез 94
10.1. Класифікація гіпотез 94
10.2. Схема перевірки наукової гіпотези 96
10.3. Схема перевірки статистичної гіпотези 98
10.4. Гіпотези про закон розподілу випадкових величин 103
10.5. Перевірка гіпотез про математичне сподівання й дисперсію неперервної випадкової величини 118
10.6. Порівняння параметрів неперервних випадкових величин у дослідах “контроль – вплив” 118
10.7. Порівняння параметрів безперервних випадкових величин у дослідах “до впливу – після впливу” 124
10.8. Гіпотези про коефіцієнт кореляції 128
10.9. Гіпотези про коефіцієнти лінійної регресії 131

Додаток 134

Список рекомендованої літератури 140

Навчальне видання

Володимир Леонідович Борщ
Юрій Петрович Совіт

Математичні методи в біології.
Основи статистики

Конспект лекцій

Редактор В. О. Насєкан
Техредактор Л. П. Замятіна
Коректор А.А. Грищенко

Підписано до друку 15.10.08. Формат 60×84/16. Папір друкарський.
Друк плоский. Ум. друк. арк. 7,8. Ум. фарбовідб. 7,8. Обл.-вид. арк. 8,5.
Тираж 200 прим. Зам. № ____.

РВВ ДДУ, пр. Гагаріна, 72, м. Дніпропетровськ, 49010.
Друкарня ДДУ, вул. Наукова, 5, м. Дніпропетровськ, 49050.