МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

О. С. МАЗМАНІШВІЛІ

МАТЕМАТИЧНА СТАТИСТИКА

НАВЧАЛЬНИЙ ПОСІБНИК ДО ПРАКТИЧНИХ ЗАНЯТЬ

Рекомендовано Міністерством освіти і науки України
як навчальний посібник для студентів спеціальностей
7.080201 "Інформатика" та 7.080202 "Прикладна математика"

Затверджено Редакційно-видавничою
Радою НТУ "ХПІ"

Харків НТУ "ХПІ" 2010

ББК 22.171
М12
УДК 519.2

Рецензенти:
Є.В. Бодянський, д-р техн. наук, проф., Харківський національний університет радіоелектроніки
Г.І. Загарій, д-р техн. наук, проф., Харківська державна академія залізничного транспорту

Гриф привласнено Міністерством освіти і науки України,
лист № @@@@ від @@@@

М12 Математична статистика: Навчальний посібник до практичних занять /
Мазманішвілі О.С. — Харків: НТУ "ХПІ", 2010.
— 232 с. — Укр. мова.

ISBN 966–593–271–3

Підготовлений для виконання практикуму з курсу "Математична статистика". У посібнику систематизовано матеріали з основних тем дисципліни: теоретичні відомості, необхідні для розв'язування задач, приклади таких розв'язувань.

Призначений для студентів, що навчаються за спеціальністю "Прикладна математика". Буде корисний студентам фізико-математичних, інженерно-технічних і економічних спеціальностей університетів, а також фахівцям.

Подготовлено для выполнения практикума по курсу "Математическая статистика". В пособии систематизированы материалы по основным темам дисциплины: теоретические сведения, необходимые для решения задач, примеры таких решений.

Предназначено для студентов, обучающихся по специальности "Прикладная математика". Будет полезно студентам физико-математических, инженерно-технических и экономических специальностей университетов, а также специалистам.

This textbook contains tasks on Mathematical Statistics sorted according to the topics of the course: theoretical information that is necessary for solving the tasks, examples of the solutions.

This textbook is created for students on Applied Mathematics. It can also be useful for students on Mathematics, Physics, Engineering and Economy.

Іл. 55. Табл. 21. Бібліогр. 42 найм.

ББК 22.171

ISBN 966–593–271–3

Зміст

Вступ

Математична і прикладна статистика широко використовується в науці, економіці, техніці, медицині та інших галузях діяльності завдяки її постійному розвитку, в тому числі програмному.

Відомості на перспективу вивчає теорія ймовірностей, а ретроспективні відомості – статистика.

Теорія ймовірностей – це один з розділів чистої математики. Будується ця теорія дедуктивно, виходячи з деяких аксіом і припущень. Найбільш суворий підхід пов'язаний з використанням теорії множин, теорії міри й інтеграла Лебега. Звичайно починають з побудови "елементарної теорії ймовірностей", в якій розглядаються випадкові події з кінцевим числом виходів. Потім теорія розповсюджується на випадок, коли кількість можливих виходів нескінченна. Застосування теорем до розв'язання різних задач теорії ймовірностей пов'язано з використанням сполучень, перестановок, операцій підсумовування й інтегрування. Методи, що застосовуються в теорії ймовірностей, такі, наприклад, як перетворення Лапласа, використовуються і в інших розділах математики.

У протилежність теорії ймовірностей математична статистика – це розділ прикладної математики. Для неї характерна, головним чином, індуктивна побудова, оскільки в цьому випадку ми йдемо в зворотному напрямку – від спостереження події до гіпотези. При цьому наша аргументація засновується на виведенні ймовірностей, всебічне знання якої, таким чином, виявляється абсолютно необхідним.

Приклад 1. *Типова задача теорії ймовірностей.* Коли підкидається монета, то є відомою ймовірність p, що випаде "герб", і ймовірність $1-p$, що випаде "цифра". Яка ймовірність того, що в результаті N кидань "герб" випаде n разів?

Використовуючи біноміальний розподіл, ми отримаємо наступний результат:

$$\Pr(n) \;=\; C_N^n p^n (1-p)^{N-n}.$$

Приклад 2. *Типова задача математичної статистики.* Монета підкидається N разів, при цьому "герб" випадає n разів. Що можна сказати про невідомий параметр p?

Очевидно, неможна сподіватися отримати на це питання також певну відповідь, як і в попередньому випадку. З самого початку ми знаємо, що $0 \le p \le 1$. Крім того, $p \neq 0$, якщо $n \neq 0$, і $p \neq 1$, якщо $n \neq N$.

Розглянемо поняття *найбільш правдоподібного значення* параметра. У цьому випадку ми могли б сказати, що *найбільш правдоподібне значення p дорівнює n/N*. Потім слід би розглянути також й інші правдоподібні значення p. У результаті отримаємо малий інтервал

$$p_1 < \frac{n}{N} < p_2,$$

який, як ми сподіваємося, буде містити істинне значення p.

Нехай $\delta p = p_2 - p_1$. Чим більше δp, тим з більшою достовірністю p попаде до вказаного інтервалу. З іншого боку, більш широкий інтервал дає нам меншу інформацію відносно самої величини p. Таким чином, у статистичному аналізі завжди присутня принципова невизначеність. У всякому випадку ми можемо розраховувати, що оцінимо цю невизначеність.

До статистичних методів звертаються в тих випадках, коли доводиться розглядати не одиничні, а масові явища. Первинною обробкою відомостей займається загальна статистика, а опрацьовуванням відомостей на основі застосування математичних методів – математична статистика. Остання є наукою про методи кількісного аналізу масових явищ.

Перші кроки в математичній статистиці були зроблені в XVIII ст., вони були пов'язані зі статистикою народонаселення і з питаннями страхування. У кінці XVIII ст. почалася серйозна робота щодо теорії помилок вимірювань, яка спричинила на початку XIX ст. створення далеко просунених її основ. Біологічні дослідження стали в XIX ст. поштовхом для постановки численних питань, що призвели на початку XX ст. до відмежування математичної статистики в окрему науку. Зараз математична статистика застосовується буквально у всіх сферах людської діяльності. Видано безліч літератури, яка висвітлює методи математичної статистики як загального, так і спеціалізованого напрямку, частина наведена в кінці цього посібника.

Видання за своїм задумом має на меті: дати студентам зручний для роботи і практично випробуваний матеріал, який навчає практичним методам і техніці розв'язування різних задач математичної статистики.

Посібник містить матеріал традиційних розділів, що складають у сукупності зміст дисципліни "Математична статистика" як другою частини дисципліны "Теорія ймовірностей і математична статистика" . Матеріал згрупований за наступними основними темами:

1. Основні поняття і задачі математичної статистики.
2. Спеціальні закони розподілу математичної статистики.
3. Статистична теорія оцінювання параметрів розподілу.
4. Статистична перевірка параметричних гіпотез.
5. Статистична перевірка непараметричних гіпотез.
6. Елементи лінійного і регресійного аналізу.

Кожний з перерахованих розділів складається з теоретичної частини, обсяг якої достатній для засвоєння відповідного матеріалу, і частин, що містять як розгорнені розв'язання прикладів, так і задачі для розв'зання та додатки, які містять необхідні статистичні таблиці. Таким чином, цей посібник може служити розв'язником з дисципліни "Математична статистика", при цьому практична частина посібника представлена задачами, які є типовими для техніки, економіки і виробництва, а теоретична частина повністю їх забезпечує. Навчальний матеріал надається в такому обсязі, що оволодіння ним дає доступ до використання сучасних пакетів прикладних програм з статистичного опрацьовування даних.

Таким чином, цей посібник систематизує та організовує навчальний матеріал з практичних аспектів опанування методів та прийомів сучасної математичної статистики. Для зручності роботи, а також можливості самостійного поглибленого вивчен-

ня й контролю засвоєння матеріалу видання скомпоновано з окремих самостійних тем, що адаптовані до відповідних розділів дисципліни "Математична статистика".

В кожному розділі наведені необхідні початкові та довідкові дані, принципи побудування методів розв'язання відповідних задач, типові алгоритми, сформульовані завдання для самостійної роботи. Практикум базується на знаннях, що їх отримують студенти в стандартному обсязі спеціальності "Прикладна математика".

Посібник є розв'язником задач за курсом математичної статистики, він містить як традиційні приклади та задачі за курсом, так і нові, які укладач взяв з наукової практики.

При укладанні матеріалу посібника малась на увазі також можливість його адаптації до навчальних програм різного обсягу та тривалості.

Цей "Посібник до практичних занять" може розглядатись як друга (односеместрова) частина "Посібника" для двосеместрової дисципліни "Теорія ймовірностей та математична статистика", перша (також односеместрова) частина якої – "Теорія ймовірностей". Викладання матеріалу та всі визначення в цих двох частинах узгоджені.

1. Основні поняття і задачі математичної статистики

1.1. Предмет і задачі математичної статистики

Теорія ймовірностей і математична статистика займаються кількісним і якісним аналізом закономірностей випадкових масових явищ. При розгляді задач теорії ймовірностей виходять з припущення, що ймовірності виникнення окремих подій відомі і задані, тобто припускалось, що закони розподілу випадкових величин або їх чисельні характеристики є відомими. Оперуючи цими поняттями, знаходили ймовірності, закони розподілу і числові характеристики інших більш складних подій і випадкових величин (ВВ). На практиці ймовірності виникнення подій, закони розподілу випадкових величин або параметри цих законів розподілу невідомі. Для їх визначення (оцінювання) необхідно проводити експеримент, спеціальні дослідження. Математична статистика розробляє методи математичного опрацювання результатів випробувань з метою отримання відомостей про ймовірності настання окремих подій, про закони розподілу випадкових величин або про параметри цих законів. При обробці результатів експерименту статистичними методами основні поняття теорії ймовірностей : імовірність настання випадкової події, закони розподілу ВВ, параметри законів розподілу ВВ і т. д. виступають як деякі математичні моделі реальних закономірностей. Таким чином, теорія ймовірностей розробляє математичні моделі для опису реальних закономірностей випадкових масових явищ, формує систему поглядів на статистичне опрацьовування результатів експерименту.

Основою статистичних методів є експериментальні дані, які часто називають *статистичними даними*. Статистичними даними називають відомості про кількість об'єктів, що мають ті чи інші ознаки. Наприклад, статистичними даними є дані про відхилення розмірів деталі від номінальних розмірів; дані про число викликів на телефонну станцію між 8 і 9 годинами ранку; дані про продуктивність праці робітників підприємства за звітний період і т. д. Перераховані дані є числовими характеристиками масових випадкових явищ (сортність деталей, навантаження АТС, продуктивність праці), тому предметом математичної статистики є випадкові явища, а її головним завданням – кількісний і якісний аналізи цих явищ.

Основні задачі математичної статистики полягають в розробці методів :

1) організації і планування статистичних спостережень;

2) збору статистичних даних;

3) "згортки інформації", тобто методів угрупування і скорочення статистичних даних з метою зведення великого числа таких даних до невеликого числа параметрів, які в стислому вигляді характеризують усю сукупність, що досліджується;

4) аналізу статистичних даних;

5) прийняття рішень і висновків на основі аналізу статистичних даних;

6) прогнозування випадкових явищ.

1.2. Генеральна і вибіркова сукупності

Одним з основних методів статистичного спостереження є *вибірковий метод.* Розглянемо основні поняття цього методу.

Нехай для дослідження закономірностей випадкового явища зроблено n дослідів, внаслідок яких отримано ряд спостережень $x_1, x_2, \ldots, x_n$. Потрібно опрацювати цей ряд статистично. Для будь-якого статистичного опрацювання необхідно спочатку побудувати *математичну модель* ряду спостережень, тобто вказати, які величини випадкові, які не випадкові, які залежні, які незалежні і т.д.

Для результатів спостережень $x_1, x_2, \ldots, x_n$ можна побудувати різні математичні моделі. Розглянемо модель, в якій ряд спостережень дається формулою

$$x_i \;=\; f(t_i) + \varepsilon_i, \tag{1.1}$$

де t_i – значення деякої детермінованої функції, що характеризує i-й дослід; $f(t_i)$ – деяка функція певного або невідомого вигляду; ε_i – випадкова величина, яку звичайно називають помилкою i-го експерименту.

Відносно помилок ε_i в моделі ряду спостережень, що задається формулою (1.1), можна використовувати також різні припущення. Наприклад, можна вважати, що вимірювання x_i супроводжуються систематичними помилками, тобто $\mathrm{M}[\varepsilon_i] \neq 0$. До того ж можна передбачати, що або ці систематичні помилки не залежать від i (постійні) : $\mathrm{M}[\varepsilon_i] = const$, або змінюються згідно з певним законом: $\mathrm{M}[\varepsilon_i] = \psi(t_i)$.

Припущення про вигляд функції $f(t)$ та про характер помилок ε_i в моделі ряду спостережень, що задається формулою (1.1), визначають методику опрацювання цього ряду. Чим більш складні припущення будуть висунуті відносно $f(t)$ і помилок ε_i в моделі (1.1), тим складніше будуть методи його статистичного опрацювання.

Відносно випадкових помилок ε_i, як правило, передбачають, що вони незалежні, а проведені вимірювання виконані в однакових і стабільних умовах (однорідні вимірювання), тобто $\mathrm{D}[\varepsilon_1] = \mathrm{D}[\varepsilon_2] = \ldots = \mathrm{D}[\varepsilon_n] = \sigma^2$. У цьому випадку кажуть, що вимірювання $x_1, x_2, \ldots, x_n$ *рівноточні.*

Частіше за все вважають, що помилки вимірювання ε_i підкоряються нормальному закону розподілу з параметрами $\mathrm{M}[\varepsilon_i] = 0$ та $\mathrm{M}[\varepsilon_i^2] = \sigma^2$, або в більш скороченому запису : $\varepsilon_i \to N(0;\sigma)$. При цьому припущенні щільність імовірності випадкової величини X, що досліджується, має вигляд

$$f_{\mathrm{x}} \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right). \tag{1.2}$$

Побудувавши нормальну модель ряду спостережень, проводять оцінку параметрів a і σ імовірносної моделі ряду спостережень.

Найбільш точні відомості про ВВ X можна одержати, якщо отримати максимально можливу кількість вимірювань цієї випадкової величини.

Визначення 1. *Генеральною сукупністю* називається сукупність усіх будь-яких спостережень, які могли б бути зроблені при даному реальному комплексі умов вимірювань. Кількість членів, що утворюють генеральну сукупність, називається *обсягом генеральної сукупності.*

Можна виділити три основні види генеральної сукупності :

1) звичайна і реально існуюча, наприклад, кількість бракованих виробів у деякій партії;

2) нескінченна і реально існуюча, наприклад, безліч дійсних чисел, що містяться між нулем і одиницею;

3) уявна (гіпотетична) кінцева або нескінченна. Наприклад, результати $x_1, x_2, \ldots, x_n$ вимірювань деякої постійної фізичної величини є елементами уявної нескінченної сукупності.

Генеральна сукупність є також поняттям модельним. Можна робити різні припущення (будувати моделі) про функцію розподілу F_{x} випадкової величини X або про параметри цієї моделі.

Визначення 2. *Вибірковою сукупністю* або просто *вибіркою обсягом* n називається сукупність n об'єктів, відібраних з генеральної сукупності, що досліджується.

Ряд розподілів $x_1, x_2, \ldots, x_n$ прийнято розглядати як вибірку обсягом n з обмеженої або нескінченної генеральної сукупності.

Визначення 3. Метод, який полягає в тому, що на основі характеристик і властивостей вибірки $x_1, x_2, \ldots, x_n$ робляться висновки про числові характеристики і закон розподілу випадкової величини X, називається *вибірковим методом.*

Для того, щоб висновки про закони розподілу ймовірностей випадкової величини X або про їх числові характеристики були об'єктивні, необхідно, щоб вибірка була представницька (репрезентативна), тобто щоб досить добре представляла випадкову величину, що досліджується. Важливо, щоб при вилученні вибірки кожний елемент генеральної сукупності мав однакову з іншими елементами ймовірність бути включеним у вибірку. Іншими словами, вибір елементів з генеральної сукупності повинен бути випадковим.

За технікою відбору елементів з генеральної сукупності у вибіркову сукупність всі вибірки поділяються на повторні і неповторні. Якщо кожний обстежений елемент повертається обернено в генеральну сукупність і тому може брати участь у подальшому відборі, то вибірка називається *повторною. Безповторна* вибірка полягає в тому, що відібрані елементи обернено в генеральну сукупність не повертаються.

Середнє арифметичне ознаки X у генеральній сукупності називають *генеральним середнім* $\overline{x}$, а його дисперсію – *генеральною дисперсією* σ_0^2.

Середнє арифметичне і дисперсія ознаки X у вибірці називаються *вибірковим середнім* x^* і *вибірковою дисперсією* σ^{*2} (зірочкою часто вказують на те, що розглядаються середні, що віднесені до вибірки).

1.3. Статистичний ряд

Припустимо, що вивчається деяка дискретна або неперервна випадкова величина, закон розподілу якої є невідомим. Для оцінки закону розподілу цієї випадкової величини, або його числових характеристик, проводиться ряд незалежних спостережень x_1, x_2, ..., x_n. Статистичний матеріал, отриманий внаслідок вимірювань, подають у вигляді таблиці, що складається з двох рядків, в першому з яких подані номери вимірювань, а в другому – результати вимірювань.

i – номер вимірювання	1	2	3	$\cdots$	n
x_i – результат вимірювання	x_1	x_2	x_3	$\cdots$	x_n

Таблицю такого виду називають *статистичним рядом*. Вона являє собою первинну форму подання статистичного матеріалу. Якщо інформація у виді простого статистичного ряду при великому числі вимірювань важко доступна для огляду, то за ним важко оцінити закон розподілу випадкової величини, що досліджується.

Для візуальної оцінки закону розподілу випадкової величини X, що досліджується, дані групуються. Якщо вивчається дискретна випадкова величина, то значення спостережень розташовуються в порядку зростання і підраховуються *частоти* m_i або *частості* $p_i = m_i/n$ появи однакових значень випадкової величини X. У результаті отримуємо згруповані статистичні ряди наступного вигляду:

x_i – результат вимірювання	x_1	x_2	x_3	$\cdots$	x_n
m_i – частоти	m_1	m_2	m_3	$\cdots$	m_n

Контроль: $\sum\limits_{i=1}^{n} m_i = n.$

x_i – результат вимірювання	x_1	x_2	x_3	$\cdots$	x_n
$p_i = m_i/n$ – частості	p_1	p_2	p_3	$\cdots$	p_n

Контроль: $\sum\limits_{i=1}^{n} p_i = \sum\limits_{i=1}^{n} m_i/n = 1.$

Якщо вивчається неперервна випадкова величина, то групування полягає в розбитті інтервалу спостережених значень ВВ на k часткових інтервалів, що дорівнюють довжині $[x_0; x_1]$, $[x_1; x_2]$, $[x_2; x_3]$, ..., $[x_{k-1}; x_k]$, і підрахунку частоти m_i або частості $p_i = m_i/n$ влучення спостережених значень у часткові інтервали. Кількість інтервалів вибирається довільно, але звичайно не менше 5 і не більше 15. У результаті складається інтервальний статистичний ряд наступного вигляду:

Інтервали спостережених значень	$[x_0; x_1]$	$[x_1; x_2]$	...	$[x_{k-1}; x_k]$
Частості $p_i = m_i/n$	m_1/n	m_2/n	...	m_k/n

Контроль: $\sum\limits_{i=1}^{k} p_i = \sum\limits_{i=1}^{k} m_i/n = 1.$

Визначення. Перелік спостережених значень випадкової величини X (або інтервалів спостережених значень і відповідних їм частостей $p_i = m_i/n$) називається *статистичним законом розподілу випадкової величини* X.

У теорії ймовірностей під законом розподілу випадкової величини розуміють відповідність між можливими значеннями (або інтервалами можливих значень випадкової величини) і їх імовірностями, а в математичній статистиці статистичний закон розподілу встановлює відповідність між спостереженими значеннями (або інтервалами спостережених значень) випадкової величини і відповідними їм частостями. Статистичні закони розподілу випадкових величин та їхне графічне зображення дозволяють зробити якісну візуальну оцінку закону розподілу випадкової величини, що досліджується.

1.4. Емпірична функція розподілу

Емпіричною функцією розподілу випадкової величини X називають функцію $F_n^*(x)$, що визначає для кожного значення x частість події $\{\mathrm{X} < x\}$:

$$F_n^*(x) \;=\; \frac{n_x}{n}, \tag{1.3}$$

де n_x – кількість x_i, менших за x; n – обсяг вибірки. Значення емпіричної функції розподілу для статистики визначається наступним твердженням.

Теорема Бернуллі

Нехай $F_n^*(x)$ – емпірична функція розподілу, побудована за вибіркою обсягу n з генеральної сукупності з функцією розподілу $F_{\mathrm{X}}(x)$. Тоді для будь-якого $x \in (-\infty, \infty)$ і будь-якого $\varepsilon > 0$ справедливо

$$\lim_{n\to\infty} \Pr\left(\,|F_n^*(x) - F_{\mathrm{X}}(x)| \;<\; \varepsilon\right) \;=\; 1. \tag{1.4}$$

З теореми Бернуллі випливає, що при досить великому обсязі вибірки функції $F_n^*(x)$ і $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$ мало відрізняються одна від одної. Відмінність емпіричної функції розподілу від теоретичної полягає в тому, що теоретична функція розподілу визначає ймовірність події $\{\mathrm{X} < x\}$, а емпірична функція визначає відносну частість цієї події.

Таким чином, при кожному x емпірична функція $F_n^*(x)$ збігається за ймовірністю до $F_{\mathrm{X}}(x)$ і, отже, при великому обсязі вибірки може слугувати наближеним значенням (оцінкою) функції розподілу генеральної сукупності в кожній точці x.

Її коротка назва – *кумулята*. Вона являє собою частку тих результатів експерименту, які не перевершують подане поточне значення.

Емпірична функція розподілу має усі властивості інтегральної функції розподілу.

З визначення емпіричної функції розподілу випливає, що:

1) значення емпіричної функції розподілу належать відрізку [0; 1];

2) $F_n^*(x)$ – функція, що не спадає;

3) якщо $x_{\min}$ – найменше, а $x_{\max}$ – найбільше спостережене значення, то $F_n^*(x) = 0$ при $x \le x_{\min}$ і $F_n^*(x) = 1$ при $x > x_{\max}$.

Основне значення емпіричної функції розподілу полягає в тому, що вона використовується як оцінка функції розподілу $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$. Зі збільшенням обсягу вибірки розбіжності між емпіричною функцією розподілу та функцією розподілу ймовірностей будуть зменшуватися. А при $n \to \infty$ ця функція, згідно з законом великих чисел, перетвориться у функцію розподілу ймовірностей значень випадкової величини.

Отже, емпірична функція розподілу має важливе практичне значення.

Приклад

Побудувати емпіричну функцію розподілу за статистичним розподілом випадкової величини X.

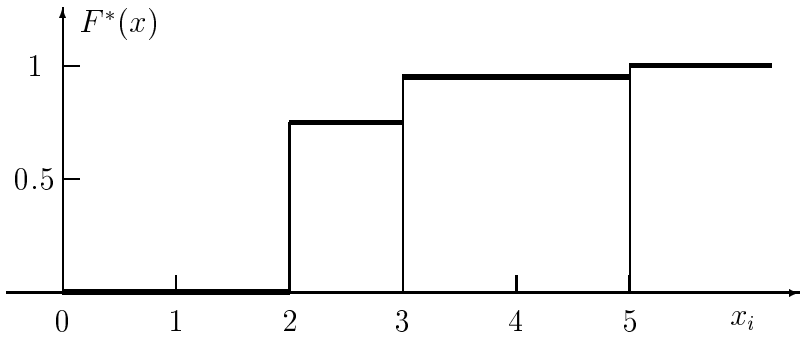


Рисунок 1.1 — Емпірична функція розподілу

Спостережені значення випадкової величини X	2	3	5
Частості $p_i = m_i/n$	0,75	0,20	0,05

Розв'язання

Відносна частота події $\{X < x\}$ дорівнює $F^*(x)$. Отже,

$$F_n^*(x) = \begin{cases} 0 & \text{при} \quad x \le 2; \\ 0,75 & \text{при} \quad 2 < x \le 3; \\ 0,95 & \text{при} \quad 3 < x \le 5; \\ 1 & \text{при} \quad x > 5. \end{cases}$$

Графік емпіричної функції розподілу наведений на рис. 1.1.

1.5. Графічне зображення статистичних рядів

Для наочності згруповані статистичні ряди часто зображують у вигляді графіків і діаграм. Найбільш поширеними є *полігон, гістограма, кумулята* та *огива*. Полігон, кумулята і огива застосовуються для зображення як дискретних, так і інтервальних статистичних рядів, гістограма – для зображення тільки інтервальних рядів.

Внаслідок побудови полігону або гістограми можна отримати перше уявлення про *форму розподілу*, під якою мається на увазі форма його графіка при нескінченній вибірці, тобто форма *кривої розподілу*.

Встановленню і вивченню форм кривих генеральних сукупностей вибіркових даних у математичній статистиці приділяється значна увага. На практиці розрізняють *одновершинні* (унімодальні) та *багатовершинні* (багатомодальні) розподіли.

Характеристика рядів розподілу передбачає з'ясування умов, під впливом яких сформувався розподіл, що вивчається, виразом його основних особливостей, числових характеристик.

Для побудови гістограми відносних частот (частостей) на осі абсцис відкладаємо часткові інтервали спостережених значень випадкової величини X, на кожному з яких будуємо прямокутник, площа якого дорівнює частості даного часткового інтервалу.

Якщо на гістограмі частостей з'єднати середини верхніх сторін елементарних прямокутників, то отримана замкнена ламана утворить *полігон* розподілу частостей.

З принципу побудови гістограми і полігона розподілу частостей випливає, що площа під гістограмою і полігоном частостей дорівнює $S = 1\,(\textit{одиниць}^2)$. У теорії ймовірностей гістограмі та полігону частостей відповідає графік густини розподілу.

Якщо в гістограмі замість частот або частостей записати відповідно накопичені частоти або частості, то вийде *кумулянтний ряд*. Для побудови кумуляти на осі абсцис відкладаємо спостережені значення випадкової величини X, на осі ординат – накопичені частості.

Накопиченою частістю в обраній точці x називається сумарна частість членів статистичного ряду, значення яких менше x, тобто значення накопичених частостей є значеннями емпіричної функції розподілу $F^*(x)$.

У теорії ймовірностей кумулятивній залежності відповідає графік функції розподілу $F_{\mathrm{X}}(x) = \Pr(\mathrm{X} < x)$.

Якщо для побудови кумуляти осі координат поміняти місцями, тобто на горизонтальній осі відкладати значення емпіричної функції розподілу $F^*(x)$, а на вертикальній – спостережені значення випадкової величини X, то отримана ламана лінія називається *огивою*.

Приклад

Результати дослідження міцності 200 зразків бетону на стиснення наведені у вигляді інтервального статистичного ряду.

Інтервали міцності, $\textit{кГ/см}^2$	Частота, m_i	Частості, m_i/n
190–200	10	0,05
200–210	26	0,13
210–220	56	0,28
220–230	64	0,32
230–240	30	0,15
240–250	14	0,07

$$n = \sum_{i=1}^{6} m_i = 200; \qquad \sum_{i=1}^{6} \frac{m_i}{n} = 1.$$

Потрібно побудувати гістограму, полігон розподілу частостей і огиву даного статистичного розподілу.

Розв'язання

При побудовах використовуємо значення емпіричної функції розподілу $F^*(x)$:

$$F^*(x) = \begin{cases} 0 & \text{при} \quad x \le 190; \\ 0,05 & \text{при} \quad 190 < x \le 200; \\ 0,18 & \text{при} \quad 200 < x \le 210; \\ 0,46 & \text{при} \quad 210 < x \le 220; \\ 0,78 & \text{при} \quad 220 < x \le 230; \\ 0,93 & \text{при} \quad 230 < x \le 240; \\ 1 & \text{при} \quad 240 < x \le 250; \\ 1 & \text{при} \quad x > 250. \end{cases}$$

На рис. 1.2 зображена гістограма частостей даного статистичного ряду.

На рис. 1.3 зображені графіки кумуляти й огиви даного інтервального ряду.

Вимірювання випадкових величин, що отримуються при проведенні експерименту і за ходом реєструються, як правило, носить хаотичний характер і є важко

оглядовим. Тому на початку проводиться *первинне опрацювання* вибіркової інформації, приклад якої продемонструємо у наступному параграфі.

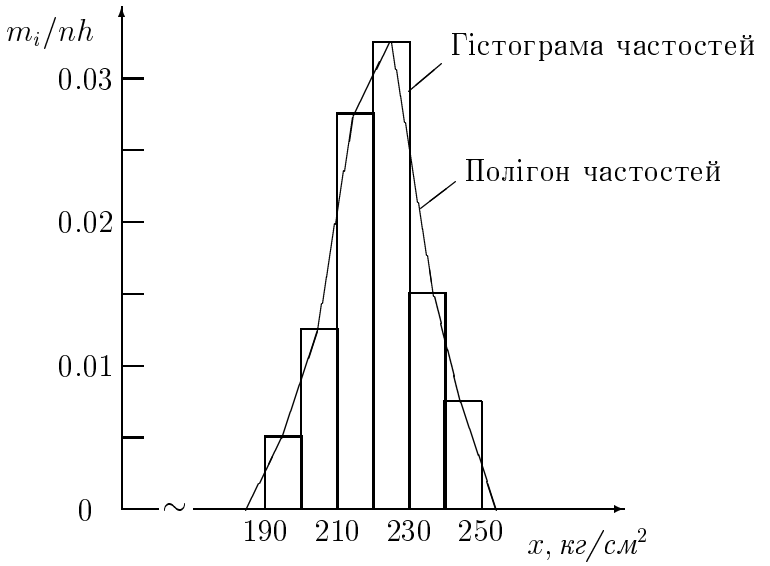


Рисунок 1.2 — Гістограма і полігон розподілу частостей

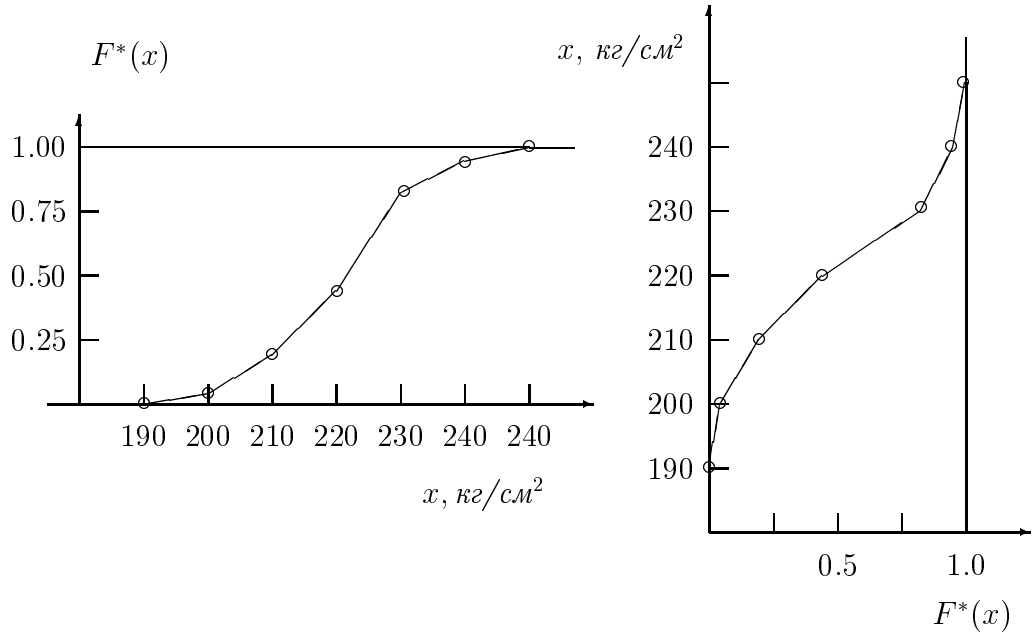


Рисунок 1.3 — Кумулята (ліворуч) і огива (праворуч)

1.6. Приклад графічного опрацювання вибіркової інформації

Припустимо, що районний військовий комісар (військком) отримав дані про контингент чергового призову, зокрема, про зріст молодих людей (таблиця 1.1).

Військком просить свого заступника надати йому інформацію про зріст призовників в такому вигляді, щоб він без розрахунків і великих витрат часу зміг би відповісти на наступні питання:

1) Який має вигляд шеренга призовників, що вишикувані за зростом від меншого до більшого?

2) Який зріст призовників найпоширеніший, скільки відсотків призовників мінімального зросту?

3) Скільки призовників потрібно спрямувати до піхоти, який відсоток це складає від усього контингенту, наскільки щільно заповнюється інтервал зросту, що відводиться для піхотинців?

4) Скільки комплектів обмундирування кожного зросту треба замовити?

Відомості про зріст молодих людей зведені в таблицю 1.1.

Таблиця 1.1 — Зріст призовників в обліковому порядку

Номер за списком	1	2	3	4	5	6	7	8
Зріст, *см*	165	171	182	165	183	180	183	166
Номер за списком	9	10	11	12	13	14	15	
Зріст, *см*	173	184	168	164	170	174	172	

Розв'язання

1. Отримавши це завдання, заступник військкому для відповіді на перше питання упорядкував відомості, побудувавши *ранжированый ряд* розподілу зросту призовників – ряд, в якому всі значення варіант (*варіантой* в статистиці називають виміряне значення) розташовуються за ранжиром (по порядку).

Таблиця 1.2 — Ранжированый ряд розподілу зросту призовників

Номер за ранжиром	1	2	3	4	5	6	7	8
Зріст, *см*	164	165	165	166	168	170	171	172
Номер за ранжиром	9	10	11	12	13	14	15	
Зріст, *см*	173	174	180	182	183	183	184	

Внаслідок переформування початкових даних (таблиця 1.1) виходить таблиця

1.2. Інформація, що міститься в ранжированому ряді розподілу, відповідно ілюструється графіком (рис. 1.4).

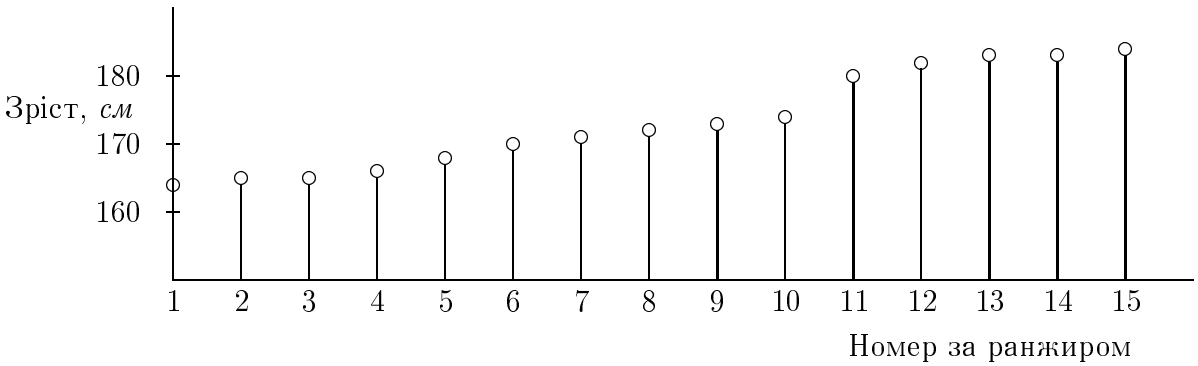


Рисунок 1.4 — Графік ранжированого ряду розподілу зросту призовників

2. Для відповіді на друге питання заступник військкому концентрує інформацію, що міститься в ранжированому ряді розподілу, і будує залежність частоти (кількості об'єктів, які мають однакові значення ознаки, тобто кількості об'єктів з однаковими значеннями варіанти) від зросту.

Крім того, йому знадобиться також значення *частостей* – відносин частот до загальної кількості об'єктів, тобто обсягу вибірки ($n = 15$). Визначимо значення варіанти x_i, частоти – m_i, частості – p_i^*.

Залежність частоти (або частості) від значень ознаки, що змінюється, називається *дискретним варіаційним рядом розподілу*. Тоді

$$p_i^* = m_i/n, \tag{1.5}$$

де n – обсяг вибірки. Всі ці відомості зводяться в таблицю 1.3.

Графік, що ілюструє дискретний варіаційний ряд, називається *полігоном розподілу*. Він представлений на рис. 1.5.

Аналізуючи цей графік, військкому легко виявити, що самим поширеним зростом в даній вибірці є зріст 165 *см* і 183 *см*, при цьому частка призовників (частість) самого малого зросту (164 *см*) становить 0,067, або 6,7 %.

Таблиця 1.3 — Дискретний варіаційний ряд розподілу зросту призовників

x_i	164	165	166	168	170	171	172
m_i	1	2	1	1	1	1	1
p_i^*	1/15	2/15	1/15	1/15	1/15	1/15	1/15
x_i	173	174	180	182	183	184	185
m_i	1	1	1	1	2	1	0
p_i^*	1/15	1/15	1/15	1/15	2/15	1/15	0

3. Для того, щоб полегшити військкому розв'язання його третього питання, заступник військкома групує призовників за їх належністю до нормативів зросту,

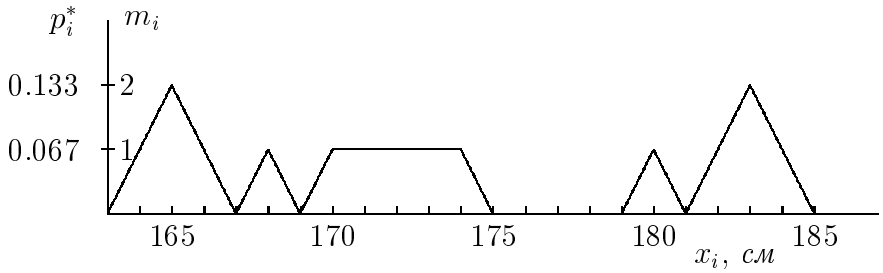


Рисунок 1.5 — Полігон розподілу призовників за ростом

прийнятих для різних родів військ, умовно до групи низького (до 170 *см* включно), середнього (171–180 *см*) і високого (понад 180 *см*) зросту.

Як характеристики розподілу об'єктів за інтервалами ознаки можуть застосовуватися частоти m_i (в одиницях або штуках), частості p_i^* (в кількості об'єктів, що доводяться на одиницю зміни ознаки).

Відповідними аналогами в теорії ймовірностей є кількість подій, імовірність події і густина розподілу ймовірностей.

Залежність перерахованих характеристик від інтервалів ознаки називається *інтервальним рядом розподілу*, а її графічна інтерпретація – *гістограмою розподілу*.

Інтервальний ряд і гістограма розподілу призовників для нашого прикладу подані в таблиці 1.4 і на рисунку 1.6. Масштаб: m_i – 2 *чоловіка* на *см*, p_i^* – 0,133 *од.* на *см*, f_i^* – 0,02 $см^{-1}$ на *см*.

Таблиця 1.4 і рисунок 1.6 наочно показують, що в піхоту слід визначити п'ятьох призовників, що становить 33,3% (частість 0,333), і що на кожний з сантиметрів середнього інтервалу доводиться в середньому 0,0333 частки низького призовника в загальному контингенті, представленому вибіркою.

Таблиця 1.4 — Інтервальний ряд розподілу зросту призовників

Інтервал, *см*	до 171	171–180	вище 180
Частота, m_i	6	5	4
Частість, p_i^*	0,40	0,333	0,267
Густина частостей, f_i^*, $см^{-1}$	0,0571	0,0333	0,0667

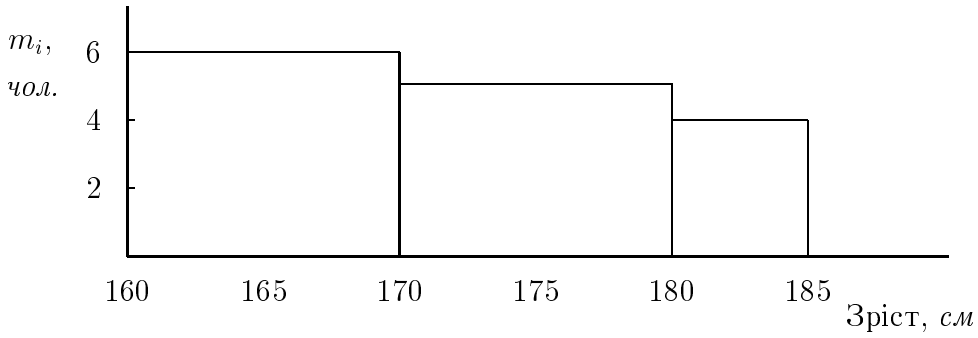


Рисунок 1.6 — Гістограма розподілу призовників за ростом

4. Той самий прийом можна застосовувати і для розв'язання останнього питання, тільки інтервали зміни ознаки (зросту) потрібно вибрати по-іншому. У застосуванні, наприклад, наш інтервальний ряд розподілу і гістограма (умовно) будуть виглядати так, як у таблиці 1.5 і на рисунку 1.7.

За допомогою таблиці 1.5 і рисунку 1.7 останнє завдання заступник військкома розв'язує автоматично – число комплектів обмундирування за зростом становить 3, 3, 4, 1 і 4.

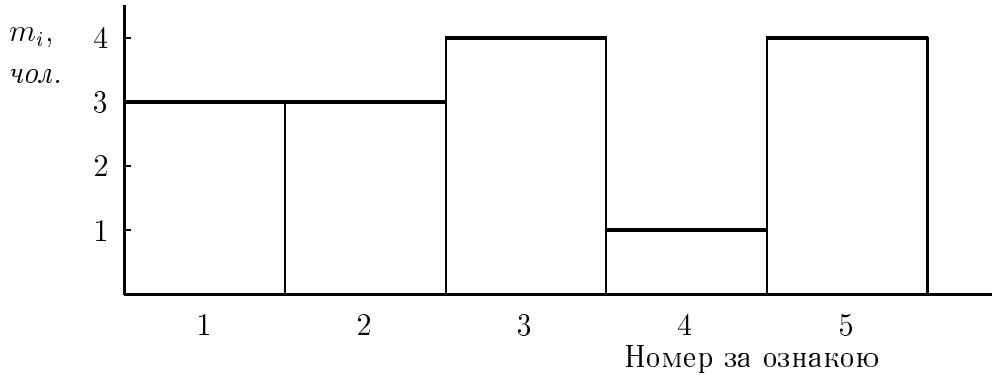


Рисунок 1.7 — Гістограма розподілу призовників за зростом

Таблиця 1.5 — Інтервальний ряд розподілу зросту призовників

Інтервал, *см*	161–165	166–170	171–175	176–180	181–184
Частота, *чол.*	3	3	4	1	4

Таблиця 1.6 — Емпірична (кумулятивна) функція розподілу призовників

Зріст, *см*	163	164	165	166	167	168
Частість	0	1/15	3/15	4/15	4/15	5/15
Зріст, *см*	169	170	171	172	173	174
Частість	5/15	6/15	7/15	8/15	9/15	10/15
Зріст, *см*	175	176	177	178	179	180
Частість	10/15	10/15	10/15	10/15	10/15	11/15
Зріст, *см*	181	182	183	184	185	
Частість	11/15	12/15	14/15	15/15	1	

Як і в теорії ймовірностей, де опис випадкової величини розглядається на декількох смислових рівнях, наприклад, на рівні густини розподілу ймовірностей (або в дискретному варіанті – на рівні дискретного ряду розподілу) і функції розподілу, в математичній статистиці прийняті аналоги цих рівнів.

Поряд з інтервальним рядом і гістограмою розподілу використовується кумулятивна (або емпірична) функція розподілу $F_n^*(x)$ (таблиця 1.6 та рисунок 1.8).

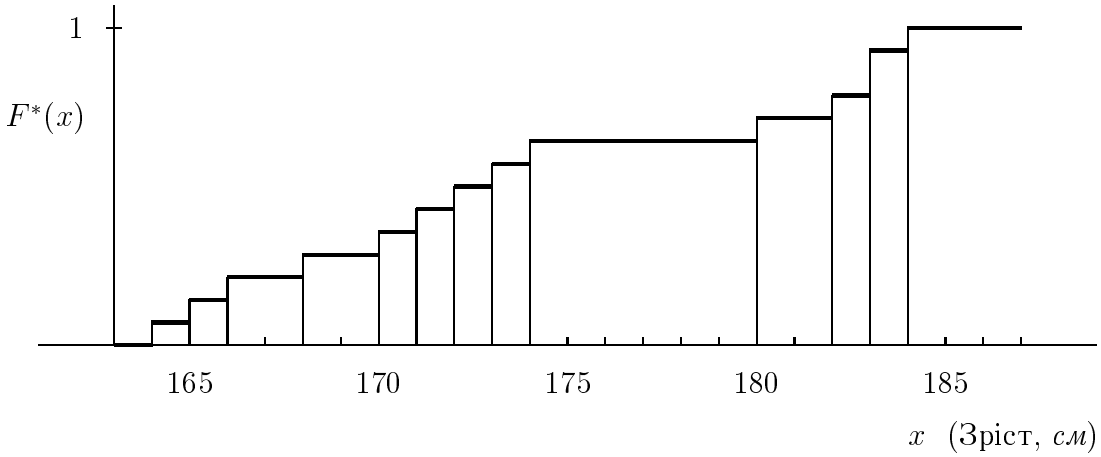


Рисунок 1.8 — Емпірична функція розподілу призовників

Всебічна характеристика рядів розподілу передбачає з'ясування умов, під впливом яких сформувався розподіл, що вивчається, вираження його основних особливостей числовими характеристиками.

1.7. Приклади

Приклад 1.1

Записати у вигляді варіаційного і статистичного рядів вибірку

$$\{5\ \ 3\ \ 7\ \ 10\ \ 5\ \ 5\ \ 2\ \ 10\ \ 7\ \ 2\ \ 7\ \ 7\ \ 4\ \ 2\ \ 4\}$$

Скласти статистичний ряд і визначити розмах вибірки.

Розв'язання

Обсяг вибірки $n = 15$. Упорядковуючи елементи вибірки за величиною, отримуємо варіаційний (ранжированый) ряд

$$\{2\ \ 2\ \ 2\ \ 3\ \ 4\ \ 4\ \ 5\ \ 5\ \ 5\ \ 7\ \ 7\ \ 7\ \ 7\ \ 10\ \ 10\}$$

Розмах вибірки складає $W = 10 - 2 = 8$.

Різними в заданій вибірці є елементи

$$x_1 = 2,\ \ x_2 = 3,\ \ x_3 = 4,\ \ x_4 = 5,\ \ x_5 = 7,\ \ x_6 = 10.$$

Їх частоти відповідно дорівнюють:

$$n_1 = 3,\ \ n_2 = 1,\ \ n_3 = 2,\ \ n_4 = 3,\ \ n_5 = 4,\ \ n_6 = 2.$$

Отже, статистичний ряд початкової вибірки можна записати у вигляді наступної таблиці:

x_i	2	3	4	5	7	10
n_i	3	1	2	3	4	2

Для контролю правильності знаходимо $\sum_i n_i = 15$.

Приклад 1.2

Представити вибірку з n спостережень у вигляді таблиці частот, використовуючи 7 інтервалів угрупування. Вибірка:

20,3	15,4	17,2	19,2	23,3	18,1	21,9	15,3	16,8	13,2	20,4
16,5	19,7	20,5	14,0	20,1	16,8	14,7	20,8	19,5	15,3	19,3
17,8	16,2	15,7	22,8	21,9	12,5	10,1	21,1	18,3	14,7	14,5
18,1	18,4	13,9	19,1	18,5	20,2	23,6	16,7	20,4	19,5	17,2
18,6	19,1	21,3	17,8	11,8	17,8	17,8	13,5	19,6	17,5	19,4

Розв'язання

Розмах вибірки $w = 23,6 - 10,1 = 13,5$. Довжина інтервалу угрупування $h = 13,5/7 \approx 2$. Як перший інтервал зручно взяти 10–12 і так далі.

Визначимо:

i – номер інтервалу;

Ω_i – межі i-го інтервалу;

x_i – середини інтервалів;

n_i – частоти;

$m_i = \sum_{j=1}^{i} n_j$ – накопичені частоти;

$p_i = n_i/n$ – відносні частоти (частості);

$F_i = \sum_{j=1}^{i} p_j$ – накопичені частості.

У результаті отримуємо наступну таблицю:

i	Ω_i	x_i	n_i	$quadm_i$	p_i	F_i
1	10–12	11	2	2	0,0364	0,0364
2	12–14	13	4	6	0,0727	0,1091
3	14–16	15	8	14	0,1455	0,2545
4	16–18	17	12	26	0,2182	0,4727
5	18–20	19	15	41	0,2727	0,7455
6	20–22	21	11	52	0,2000	0,9455
7	22–24	23	3	55	0,0545	1,0000

Приклад 1.3

Для визначення міцності нитки проведено серію з $n = 1000$ випробувань. Отримані наступні результати:

Міцність нитки, $г$	180–190	190–200	200–210	210–220
Кількість випробувань, n_i	50	90	150	280
Міцність нитки, $г$	220–230	230–240	240–250	
Кількість випробувань, n_i	220	120	90	

Побудувати кумулянтний ряд.

Розв'язання

Спочатку знаходимо накопичені частоти для кожного з інтервалів даного інтервального варіаційного ряду:

$$m_1 = n_1 = 50; \;\; m_2 = n_1 + n_2 = 50 + 90 = 140;$$

$$m_3 = 290;\ \ m_4 = 570;\ \ m_5 = 790;\ \ m_6 = 910;\ \ m_7 = 1000.$$

Таким чином, після нормування на 1000 кумулянтний ряд для даного розподілу має вигляд :

Міцність нитки, $г$	180–190	190–200	200–210	210–220
Кількість випробувань, n_i	50	90	150	280
$F_i = m_i/n$	0,050	0,140	0,290	0,570
Міцність нитки, $г$	220–230	230–240	240–250	
Кількість випробувань, n_i	220	120	90	
$F_i = m_i/n$	0,790	0,910	1,000	

Приклад 1.4

Побудувати дискретний варіаційний ряд для наступного розподілу 45 пар чоловічого взуття, проданого в магазині за день :

39 41 40 42 41 40 42 44 40 43 42 41 43 43 38
39 42 41 42 39 41 37 43 41 38 43 42 41 39 41
40 41 38 44 40 39 41 40 42 40 41 42 40 41 42

Розв'язання

Для побудови варіаційного ряду різні значення ознаки розташовуємо в порядку їхнього зростання. Остаточно варіаційний ряд приймає вигляд :

Розмір	37	38	39	40	41	42	43	44
Частота	1	3	5	8	12	9	5	2

Приклад 1.5

Для визначення середнього відсотка сирого білка в зернах пшениці було відібрано 626 зерен, обстеження яких показало, що вибіркове середнє дорівнює 16,8, а вибіркова дисперсія становила 4.

Чому дорівнює ймовірність p того, що середній відсоток сирого білка відрізняється від 16,8 за абсолютною величиною менш ніж на 0,2%?

Розв'язання

Обсяг n генеральної сукупності є невідомим. Тому помилку вибірки знаходимо за формулою $\mu \approx \sigma/\sqrt{n-1}$.

У нашому випадку $n = 626$, $\sigma^2 = 4$ і, отже, $\mu \approx 2/\sqrt{625} = 2/25 = 0{,}08$.

Оскільки $\Pr(|\overline{\mathrm{X}} - \mathrm{X}| < \varepsilon) \approx 2\Phi\,(\varepsilon/\mu)$, то для шуканої ймовірності отримуємо

$$p \;=\; \Pr(|\overline{\mathrm{X}} - 16{,}8| < 0{,}2) \;\approx\; 2\Phi\,(0{,}2/0{,}08) \;=\; 2\Phi(2{,}5) \;=\; 0{,}98758.$$

Приклад 1.6

На кожну сотню виготовлених деталей в середньому бувають дві, що не задовольняють стандарт (брак). Було перевірено 10 партій по 100 виробів у кожній. Відхилення кількості виявлених бракованих виробів від середнього наведено у таблиці :

Номер партії	1	2	3	4	5	6	7	8	9	10
Відхилення від середнього	–1	0	1	1	–1	1	0	–2	2	1

Побудувати варіаційний ряд. Знайти вибіркове середнє відхилення кількості бракованих виробів від встановленого і його вибіркову дисперсію.

Розв'язання

Вибірковий ряд $\{-2, -1, -1, 0, 0, 1, 1, 1, 1, 2\}$ містить п'ять різних значень: $\{-2, -1, 0, 1, 2\}$. Їхні частості відповідно дорівнюють $\{0{,}1, 0{,}2, 0{,}2, 0{,}4, 0{,}1\}$. Таблиця, що надає властивості вибіркової випадкової величини, наступна:

Номер, i	1	2	3	4	5
Значення, x_i	–2	–1	0	1	2
Частість, f_i	0,1	0,2	0,2	0,4	0,1

Вибіркове середнє складає $\overline{\mathrm{X}} = 0,2$.

Вибіркова дисперсія дорівнює $\sigma^{*2} = 1,36$.

Приклад 1.7

З метою визначення середньої суми внесків в ощадній касі, що має $N = 2200$ вкладників, проведене вибіркове обстеження (безповторний відбір) $n = 111$ внесків, яке дало такі результати:

Сума внеску	10–30	30–50	50–70	70–90	90–110	110–130
Кількість внесків	1	3	10	30	60	7

Користуючись цими даними, знайти довірчі межі для генерального середнього, яке можна було б гарантувати з імовірністю $p = 0,96$.

Розв'язання

Крок інтервалу складає $h = 20$. Визначимо середину i-го інтервалу через x_i, $i = 1, \ldots, 6$. Для спрощення розрахунків введемо допоміжну випадкову величину $\mathrm{U} = (\mathrm{X} - x_4)/h = (\mathrm{X} - 80)/20$. Спочатку знайдемо вибіркове середнє і вибіркову дисперсію, для чого зведемо обчислення в таблицю:

i	Межа інтервалу	x_i	m_i	$u_i = (x_i - 80)/20$	$m_i u_i$	$m_i u_i^2$
1	10–30	20	1	–3	–3	9
2	30–50	40	3	–2	–6	12
3	50–70	60	10	–1	–10	10
4	70–90	80	30	0	0	0
5	90–110	100	60	1	60	60
6	110–130	120	7	2	14	28
Сума			111		55	119

З таблиці маємо

$$\overline{\mathrm{U}} = \frac{\sum_i m_i u_i}{\sum_i m_i} = \frac{55}{111} = 0,4955; \qquad \overline{\mathrm{U}^2} = \frac{\sum_i m_i u_i^2}{\sum_i m_i} = \frac{119}{111} = 1,0721.$$

Звідси

$$\sigma_{\mathrm{U}}^2 = \overline{\mathrm{U}^2} - \left(\overline{\mathrm{U}}\right)^2 = 1,0721 - 0,2455 = 0,8266,$$

що дає

$$\overline{\mathrm{X}} = 80 + 20 \cdot 0,4955 = 89,91; \qquad \sigma_{\mathrm{X}}^2 = 20^2 \cdot 0,8266 = 330,64.$$

Тому помилка визначення середнього μ складає

$$\mu \approx \sqrt{\frac{\sigma_{\text{x}}^2}{n-1}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{330,64}{111-1}\left(1-\frac{111}{2200}\right)} = 1,69.$$

Оскільки $\Phi(t) = 0,96$ при $t = 2,05$ (див. таблиці функції Лапласа), то межова похибка $\varepsilon = t\mu = 2,05 \cdot 1,69 = 3,46$. Отже, з імовірністю $p = 0,96$ довірчими межами для генерального середнього будуть:

$$\overline{\text{X}} - \varepsilon = 89,91 - 3,46 = 86,45 \quad \text{та} \quad \overline{\text{X}} + \varepsilon = 89,91 + 3,46 = 93,37.$$

Приклад 1.8

З партії деталей було відібрано 400, розподіл яких за розміром подано в наступній таблиці:

Розмір деталі, *мм*	7,95–8,00	8,00–8,05	8,05–8,10	8,10–8,15	8,15–8,20	8,20–8,25
Кількість деталей	12	28	132	150	62	16

Знайти помилку вибірки μ при визначенні середнього.

Розв'язання

Спочатку знаходимо вибіркову дисперсію σ^{*2}: $\sigma^{*2} = 0,00272$. Це дає для помилки вибірки при визначенні середнього

$$\mu = \sigma_0/\sqrt{n-1} = \sqrt{0,00272/(400-1)} = 0,002611.$$

Оскільки $\sqrt{0,00272/400} = 0,002608$, то для таких великих значень n можна помилку вибірки μ знаходити практично за формулою $s = \sigma_0/\sqrt{n}$.

Приклад 1.9

Для групованої вибірки даних, що подана в таблиці, обчислити середнє $\overline{\text{X}}$ і дисперсію $\sigma_{\text{x}}^{*\,2}$.

Номер інтервалу i	1	2	3
Межа інтервалу Ω_i	134–138	138–142	142–146
Частота n_i	1	3	15
Номер інтервалу i	4	5	6
Межа інтервалу Ω_i	146–150	150–154	154–158
Частота n_i	18	14	2

Розв'язання

У випадкової величини X, що задана вибіркою, довжина кожного з інтервалів угрупування складає $h = 4$, а значення середини інтервалу, що зустрічається з найбільшою частотою, дорівнює $d^* = 148$. Визначимо набір $\{z_i\}$ – значення середин i-х інтервалів і перетворемо групову вибірку таким чином:

$$u_i = \frac{x_i - d^*}{h} = \frac{x_i - 148}{4}, \qquad i = 1, 2, \ldots, 6,$$

тобто використаємо приведену випадкову величину U.

Обчислення зведемо в таблицю:

i	$quadz_i$	u_i	n_i	$n_i u_i$	$n_i u_i^2$
1	136	–3	1	–3	9
2	140	–2	3	–6	12
3	144	–1	15	–15	15
4	148	0	18	0	0
5	152	1	14	14	14
6	156	2	2	4	8
Сума	–	–	53	–6	58

Тепер знаходимо

$$\overline{\mathrm{X}} \;=\; h\overline{\mathrm{U}} + d^* \;=\; 4 \cdot \frac{-6}{53} + 148 \;\approx\; 147,55,$$

$$\sigma^{*\,2}_{\mathrm{X}} \;=\; h^2 \sigma^{*\,2}_{\mathrm{U}} \;=\; 4^2 \cdot \frac{58 - (-6)^2/53}{53} \;\approx\; 17,30.$$

Приклад 1.10

Значення n незалежних випадкових величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$, що мають одну й ту саму функцію розподілу $G(x)$, розташовані в порядку зростання.

Знайти розподіл ν-го значення знизу U та μ-го значення V зверху в цьому ранжированому ряду.

Розв'язання

Для того, щоб величина U потрапила на інтервал $[u,\, u + du)$, необхідно і достатньо, щоб які-небудь $\nu - 1$ з n величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ набули значення, менші ніж u, одне значення влучило в інтервал $[u,\, u + du)$, а інші $n - \nu$ величин набули значення, не менші ніж $u + du$.

Таким чином, ми приходимо до схеми повторення дослідів з трьома несумісними подіями, що створюють повну групу:

$$A_1 \;=\; \{\mathrm{X} < u\},$$

$$A_2 \;=\; \{\mathrm{X} \geq u + du\},$$

$$A_3 \;=\; \{u \leq \mathrm{X} < u + du\}.$$

Якщо визначити густину розподілу $g(u) = dG(u)/du$, то ймовірності наведених подій при одному досліді дорівнюють відповідно (з точністю до нескінченно малих вищих порядків)

$$\Pr(A_1) = G(u), \qquad \Pr(A_2) = 1 - G(u), \qquad \Pr(A_3) = g(u)du.$$

Тоді отримаємо

$$f_{\mathrm{U}}(u) \;=\; \frac{n\,!}{(\nu-1)\,!\,(n-\nu)}\, G^{\nu-1}(u)\,[1 - G(u)]^{n-\nu}\, g(u)$$

– ця формула визначає густину розподілу ν-го значення ряду.

Аналогічно знаходиться густина μ-го значення V:

$$f_{\rm V}(v) \;=\; \frac{n\,!}{(\mu-1)\,!\,(n-\mu)}\,G^{\mu-1}(v)\,[1-G(v)]^{n-\mu}\,g(v).$$

Для ранжированого ряду розглянемо випадкові величини:

найменше значення ряду $\mathrm{U}=\min\limits_{1\le i\le n}\{\mathrm{X}_i\}$;

найбільше значення ряду $\mathrm{V}=\max\limits_{1\le i\le n}\{\mathrm{X}_i\}$.

При $\nu = 1$ та $\mu = 1$ отримуємо для густини розподілу найменшого U і найбільшого V значення ряду, що розглядається,

$$f_{\rm U}(u) \;=\; n[1-G(u)]^{n-1}\,g(u),$$

$$f_{\rm V}(v) \;=\; nG^{n-1}(u)\,g(v).$$

Наведемо також вирази для густини розподілу різниці між ν-м значенням зверху і ν-м значенням знизу $\mathrm{R} = \mathrm{V} - \mathrm{U}$:

$$f_{\rm R}(r) \;=\; \frac{n!}{[(\nu-1)!]^2\,(n-2\nu)!}\;\times$$

$$\times\int\limits_{-\infty}^{\infty} g(x)g(r+x)G^{\nu-1}(x)[G(r+x)-G(x)]^{n-2\nu}\,[1-G(r+x)]^{\nu-1}\,dx.$$

В окремому випадку при $\nu = 1$ звідси отримуємо густину розподілу *широти розкиду*

$$\mathrm{S} \;=\; \max_{1\le i\le n}\{\mathrm{X}_i\} - \min_{1\le i\le n}\{\mathrm{X}_i\}$$

у вибірці з n незалежних випадкових величин (*розмах вибірки*):

$$f_{\rm S}(S) \;=\; n(n-1)\int\limits_{-\infty}^{\infty}[G(x+s)-G(x)]^{n-2}g(x)g(s+x)\,dx.$$

Приклад 1.11

Для перевірки роботи програми побудови гістограми був опрацьований масив $\{x_n\}$, $n = 1, 2, \ldots, N$, де обсяг вибірки $N = 10000$. Значення аргументу x були здобуті за правилом $x_n = n\pi/N$, а значення функції $y(x)$ – за правилом $y_n = \sin(x_n)$. Таким чином, вихідний масив є вибіркою значень *невипадкової функції*.

Побудувати гістограму масива $\{y_n\}$, обравши кількість каналів аналізу $M = 24$.

Розв'язання

Значення функції обмежені, очевидно, інтервалом $0 \le y \le 1$. Тому в цьому інтервалі й будемо задавати канали аналізу гістограми шириною $h = 1/M$ кожний.

Гістограма наведена на рис. 1.9. Такий її вигляд відповідає графіку густини розподілу деякої випадкової величини, яка підпорядковується закону арксинуса.

Зауваження. Якщо б змінна x була *випадковою величиною*, що рівномірно розподілена на інтервалі $(0 \le x \le \pi)$, то тип гістограми був би таким же.

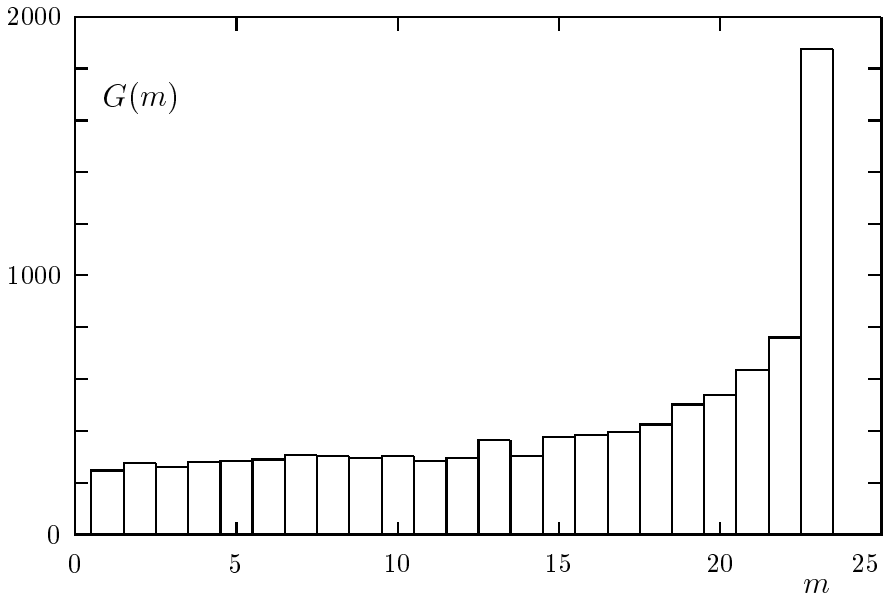


Рисунок 1.9 — Гістограма $G(m)$ функції $y = \sin(x)$ (невипадкова величина x рівномірно заповнює інтервал $[0,\ \pi/2]$; обсяг вибірки $N = 10000$; кількість каналів аналізу гістограми $M = 24$; значення обмежені величиною $\max y = 1$)

1.8. Задачі для розв'язання

Задача 1.1

Спостереження за товщиною (в *мм*) слюдяних зразків дали наступні результати:

0,021	0,030	0,039	0,031	0,042	0,034	0,036	0,030	0,028
0,030	0,033	0,024	0,031	0,040	0,031	0,033	0,031	0,027
0,031	0,045	0,031	0,034	0,027	0,030	0,048	0,030	0,028
0,030	0,033	0,046	0,043	0,030	0,033	0,028	0,031	0,027
0,031	0,036	0,051	0,034	0,031	0,036	0,034	0,037	0,028
0,030	0,039	0,031	0,042	0,037	0,031	0,036	0,028	0,026

Побудувати за цими даними інтервальний варіаційний ряд з рівними інтервалами (перший інтервал 0,020–0,024, другий 0,024–0,028 і т.д.) і накреслити гістограму.

Задача 1.2

Спостереження за товщиною (в *мм*) мідних зразків дали наступні результати:

0,031	0,040	0,049	0,041	0,052	0,044	0,046	0,040	0,038	0,040
0,043	0,034	0,041	0,050	0,041	0,043	0,041	0,037	0,041	0,055
0,041	0,044	0,047	0,040	0,058	0,040	0,038	0,040	0,043	0,056
0,053	0,040	0,043	0,038	0,041	0,037	0,041	0,046	0,061	0,044
0,041	0,046	0,044	0,047	0,038	0,040	0,049	0,041	0,042	0,037

Побудувати кумулянтний ряд, накреслити кумуляту і огиву статистичного розподілу наведених даних.

Задача 1.3

Побудувати дискретний варіаційний ряд і накреслити полігон розподілу групи абітурієнтів за числом балів, отриманих ними на вступних іспитах:

20	19	22	24	21	18	23	17	20	16	15	23	21	24	21	18	23	21	19	20
24	21	20	18	17	22	20	16	22	18	20	17	21	17	19	20	16	20	21	18
20	19	21	24	22	23	21	23	22	21	19	20	23	22	25	22	23	21	25	22
19	23	21	23	21	21	19	20	23	26	25	22	23	21	25	22	21	21	20	19

Задача 1.4

Дані про урожайність на різних ділянках поля наведені в наступній таблиці:

Урожайність, *ц/га*	9–12	12–15	15–18	18–21	21–24	24–27
Частка дільниці від загальної посівної площі, %	6	12	33	22	19	8

Побудувати кумулянтний ряд, накреслити кумуляту і огиву.

Задача 1.5

Середня температура повітря у вересні в двох містах (X) і (Y) вимірювалася протягом 40 років. Дані наведені у таблиці.

X	Y	X	Y	X	Y	X	Y	X	Y
12,0	10,8	13,9	10,1	14,9	13,0	16,0	16,0	18,0	14,0
12,0	11,3	14,2	10,0	14,9	14,2	16,9	12,9	18,0	14,9
12,0	12,0	14,0	10,0	15,1	13,8	17,2	13,9	18,1	16,0
12,0	13,0	14,0	12,0	15,0	16,0	16,9	14,8	18,4	17,8
12,8	10,9	13,9	12,4	15,5	13,9	16,9	15,0	19,2	15,0
13,8	10,0	15,0	11,0	15,9	14,7	17,0	16,0	19,3	16,1
13,1	13,0	14,0	14,8	16,0	13,0	16,8	17,0	20,0	17,0
13,0	13,0	14,0	15,2	15,9	15,0	17,5	16,0	20,1	17,7

Знайти вибіркові середньомісячні температури в обох населених пунктах і їх середньоквадратичні відхилення.

Задача 1.6

Через кожну годину вимірювалася напруга струму в електромережі. При цьому були отримані наступні значення (в V):

227	219	215	230	232	223	220	222	218	219	222	221	227	226	226	209
211	215	218	220	216	220	220	221	225	224	212	217	219	220	220	222
209	210	216	224	216	222	223	222	227	222	215	227	216	222	220	224

Побудувати статистичний розподіл і накреслити полігон.

Задача 1.7

Випробовувалася чутливість 40 приймачів. Дані наведені в таблиці, в якій в першому рядку дані інтервали чутливості в мікровольтах, у другій – середні точки

цих інтервалів $f_{\text{ср}}$, в третій – число приймачів n_i, чутливість яких виявилася в цьому інтервалі.

Інтервал	25–75	75–125	125–175	175–225	225–275
$f_{\text{ср}}$	50	100	150	200	250
n_i	0	0	1	5	8
Інтервал	275–325	325–375	375–425	425–475	475–525
$f_{\text{ср}}$	300	350	400	450	500
n_i	6	8	6	2	2
Інтервал	525–575	575–625	625–675	675–725	725–775
$f_{\text{ср}}$	550	600	650	700	750
n_i	0	1	1	0	0

Побудувати емпіричну функцію розподілу і гістограму вибірки, знайти середню чутливість приймачів з цієї партії та її середнє вибіркове відхилення.

Задача 1.8

Обстеження дало такий розподіл за зростом групи юнаків:

Зріст, *см*	Кількість юнаків
143–146	1
146–149	2
149–152	8
152–155	26
155–158	65
158–161	120
161–164	181
164–167	201

Зріст, *см*	Кількість юнаків
167–170	170
170–173	120
173–176	64
176–179	28
179–182	10
182–185	3
185–188	1
188–191	1

Знайти моду, медіану і середнє арифметичне цього розподілу. Побудувати кумулянтний ряд, накреслити кумуляту і огиву.

Задача 1.9

Спостереження за відсотком жиру в молоці 44 корів дали такі результати:

3,86 4,06 3,67 3,97 3,76 3,61 3,96 4,04 3,84 3,94 3,98
3,99 3,69 3,76 3,71 3,94 3,82 3,71 3,81 4,02 4,17 3,72
3,73 3,52 3,89 3,92 4,18 4,26 4,16 3,76 4,00 3,46 4,08
3,57 3,87 4,07 4,03 4,14 3,72 4,33 3,82 4,03 3,62 3,91

Побудувати за цими даними інтервальний варіаційний ряд з рівними інтервалами (перший інтервал 3,45–3,55 %, другий інтервал 3,55–3,65 % і т.д.) і зобразити його графічно.

Задача 1.10

Випадкова величина X рівномірно розподілена на інтервалі $(0;\ 2\pi)$, а випадкова величина Y зв'язана з X співвідношенням $Y = \text{tg}(X)$. Отримана вибірка обсягом n значень випадкової величини Y.

Навести диференціальний та інтегральний закони розподілу випадкової величини Y. Побудувати гістограму та кумуляту вибірки.

Задача 1.11

Група абітурієнтів отримала наступні бали на вступних іспитах:

20	19	22	24	21	18	23	17	20	16	15	23	20	21	24	21
18	23	21	19	20	24	21	20	18	19	17	22	20	16	22	18
20	17	21	17	19	20	16	20	21	18	22	23	21	25	22	20
19	21	24	22	23	21	29	22	21	19	20	23	22	25	21	16
21	20	20	24	21	19	18	23	18	19	19	28	22	16	22	21

Знайти моду, медіану і середнє арифметичне цього розподілу.

Задача 1.12

На молочній фермі зареєстрували відомості про величину удою корів:

Удій, *кг*	Кількість корів
400–600	1
600–800	3
800–1000	6
1000–1200	11
1200–1400	15
1400–1600	20

Удій, *кг*	Кількість корів
1600–1800	14
1800–2000	12
2000–2200	10
2200–2400	6
2400–2600	2
2600 та більш	2

Знайти дисперсію, коефіцієнт варіації і розмах варіації розподілу.

Задача 1.13

Обстеження показало наступний розподіл за зростом групи дівчат:

Зріст, *см*	Кількість дівчат
133–136	1
136–139	2
139–142	8
142–145	26
145–148	65
148–151	120
151–154	181
154–157	201

Зріст, *см*	Кількість дівчат
157–160	170
160–163	120
163–166	64
166–169	28
169–172	10
172–175	3
175–178	1
178–181	1

Знайти моду, медіану і середнє арифметичне цього розподілу. Побудувати кумулянтний ряд і накреслити кумуляту і огиву.

Задача 1.14

Випадкова величина X рівномірно розподілена на інтервалі $(0;\ 2\pi)$, а випадкова величина Y зв'язана з X співвідношенням $Y = \sin(X)$. Отримана вибірка обсягом n значень випадкової величини Y.

Навести диференціальний та інтегральний закони розподілу випадкової величини Y. Побудувати гістограму та кумуляту вибірки.

1.9. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

Напишить програму, яка генерує вибірку обсягом N випадкових величин, підпорядкованих заданому закону розподілу. В кажному з законів розподілу передбачте можливість варіювання параметрів. У випадку використання математичних пакетів користуватися вбудованими функціямі можна лише для порівняння. Результат роботи програми – масив, якій містить значення вибірки даних. Необхідно передбачити візуалізацію даних (побудувати програмно гістограму та кумуляту). Результати оформіть графічно.

Варіант 1

Рівномірний розподіл.
Вхідні дані для програми:
n_1 – ліва межа можливих значень;
n_2 – права межа можливих значень;
N – обсяг вибірки;
x_n – послідовність вибіркових значень, $1 \leq n \leq N$;
M – кількість каналів анализу гістограми та кумуляти.

Варіант 2

Нормальний закон Гаусса.
Вхідні дані для програми:
m_{x} – математичне сподівання;
σ^2_{x} – дисперсія;
N – обсяг вибірки;
x_n – послідовність вибіркових значень, $1 \leq n \leq N$;
M – кількість каналів анализу гістограми та кумуляти.

Варіант 3

Розподіл Пуассона.
Вхідні дані для програми:
m_{x} – математичне сподівання;
N – обсяг вибірки;
x_n – послідовність вибіркових значень, $1 \leq n \leq N$;
M – кількість каналів анализу гістограми та кумуляти.

1.10. Завдання для перевірки

1. Які задачі розглядає математична статистика?
2. Що називається генеральною сукупністю, вибірковою сукупністю?
3. У чому полягає сутність вибіркового методу?

4. Що називається статистичним законом розподілу випадкової величини?

5. Що називається емпіричною функцією розподілу випадкової величини?

6. У чому полягає відмінність між емпіричною функцією розподілу і теоретичною (інтегральною) функцією розподілу випадкової величини?

7. Які властивості має емпірична функція розподілу випадкової величини?

8. Назвіть основні види графіків, що слугують для зображення статистичних рядів.

9. Назвіть імовірнісні аналоги полігона, гістограми і кумуляти.

2. Спеціальні закони розподілу математичної статистики

Розподіли основних статистик, що обчислюються за вибіркою з нормально розподіленої генеральної сукупності, пов'язані з законом Гаусса (нормальним законом), гамма-розподілом, розподілом χ^2, розподілом Стьюдента і розподілом Фішера-Снедекора. При їх розгляді ми також зустрінемося з гамма-функцією Ейлера $\Gamma(\alpha)$ і розподілом Колмогорова $K(\lambda)$. В задачах математичної статистики також часто використовуються розподіли Колмогорова та Колмогорова-Смирнова $K(\lambda)$, розподіли Бартлетта, Крамера та інші.

2.1. Нормальний закон

Випадкова величина X має *нормальний розподіл* з параметрами m та σ^2, якщо її густина розподілу наступна:

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \tag{2.1}$$

областю визначення X є вся числова вісь, тобто $-\infty < \mathrm{X} < \infty$.

Про таку випадкову величину ще кажуть, що *вона розподілена згідно із законом Гаусса*. Ця випадкова величина найбільш широко застосовується при побудові статистичних моделей.

На практиці скорочено нормальну випадкову величину X з параметрами m і σ^2 часто позначають як $\mathcal{N}(m;\sigma)$.

Крива розподілу, що описує густина (2.1), симетрична відносно точки m, в якій густина досягає максимуму. З цієї симетрії безпосередньо випливає, що математичне сподівання випадкової величини X

$$\mathrm{M}[\mathrm{X}] \;=\; m. \tag{2.2}$$

Зі зміною значення математичного сподівання крива, як ціле, зміщується в бік зміни, як це показано на рис. 2.1–2.2. У точці $x = m$ досягається максимум, що дорівнює $\left(\sqrt{2\pi}\sigma\right)^{-1} = 0,3989/\sigma$.

Дисперсія нормальної випадкової величини X

$$\mathrm{D}[\mathrm{X}] \;=\; \sigma^2. \tag{2.3}$$

Таким чином, параметр σ можна інтерпретувати як міру розсіювання випадкової величини X навколо свого математичного сподівання (рис. 2.3–2.4). З визначення

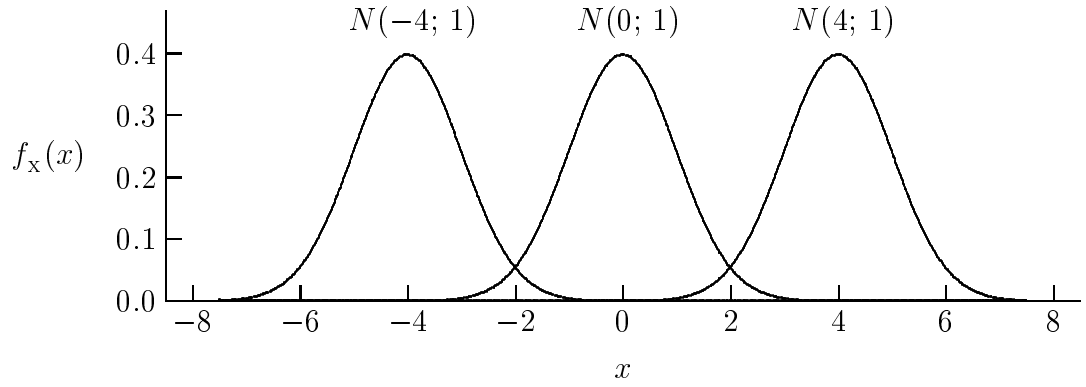


Рисунок 2.1 — Густина $f_X(x)$ нормального розподілу Гаусса для випадків $N(-4;\ 1)$, $N(0;\ 1)$ та $N(4;\ 1)$ з параметрами $\sigma = 1$ та $m = -4,\ 0$ й 4 відповідно

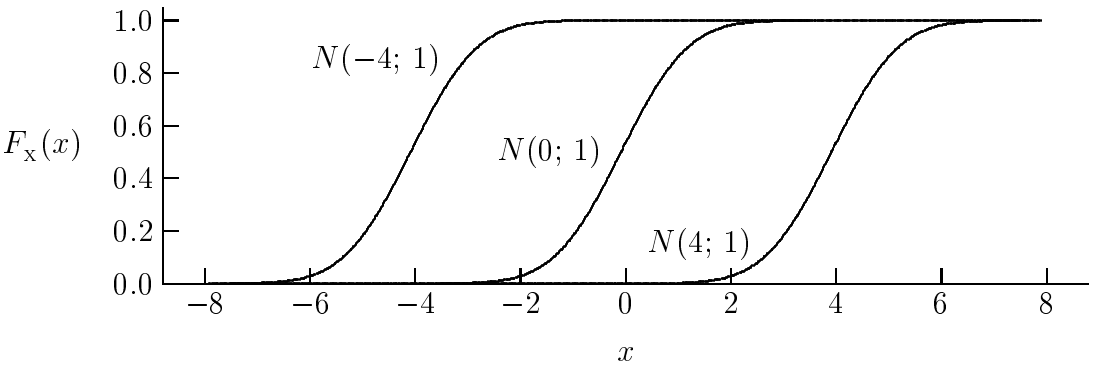


Рисунок 2.2 — Інтегральний закон $F_X(x)$ розподілу Гаусса для випадків $N(-4;\ 1)$, $N(0;\ 1)$ й $N(4;\ 1)$ з параметрами $\sigma = 1$ та $m = -4,\ 0$ й 4 відповідно

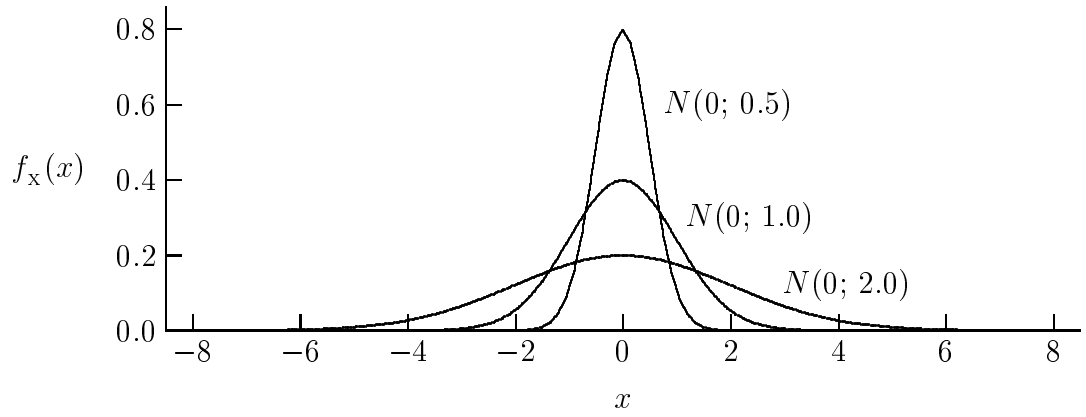


Рисунок 2.3 — Густина $f_X(x)$ нормального розподілу Гаусса для випадків $N(0;\ 0,5)$, $N(0;\ 1,0)$, $N(0;\ 2,0)$ з параметрами $m = 0$ та $\sigma = 0{,}5,\ 1{,}0$ й $2{,}0$ відповідно

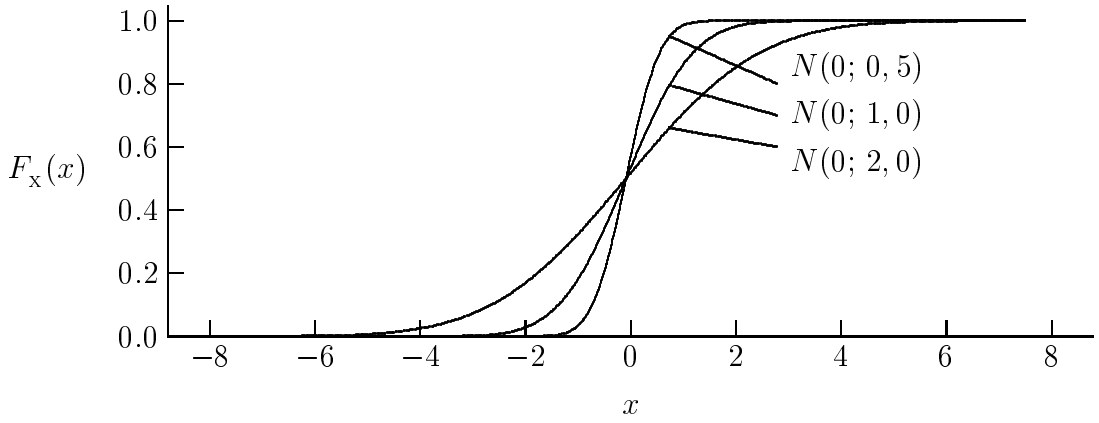


Рисунок 2.4 — Інтегральний закон $F_{\text{X}}(x)$ розподілу Гаусса для випадків $N(0;\,0,5)$, $N(0;\,1,0)$, $N(0;\,2,0)$ з параметрами $m = 0$ та $\sigma =$ 0,5, 1,0 й 2,0 відповідно

(2.1) випливає, що нормальний розподіл повністю визначається своїми двома параметрами — математичним сподіванням m та дисперсією σ^2.

Інтегральна функція розподілу $F_{\text{X}}(x)$ нормальної випадкової величини має вигляд

$$F_{\text{X}}(x) \;=\; \Pr(\text{X} < x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma} \int\limits_{-\infty}^{x} \exp\left(-\frac{(x'-m)^2}{2\sigma^2}\right) dx'. \tag{2.4}$$

За допомогою функції $F_{\text{X}}(x)$ може бути визначена ймовірність влучення нормальної випадкової величини в заданий інтервал $(\alpha,\,\beta)$:

$$\Pr(\alpha \le \text{X} < \beta) \;=\; F_{\text{X}}(\beta) - F_{\text{X}}(\alpha). \tag{2.5}$$

Характеристична функція $q(\lambda)$ нормальної випадкової величини X наступна $(\sqrt{-1})$:

$$q(\lambda) \;=\; \text{M}[\exp(i\lambda\text{X})] \;=\; \exp\left(im\lambda - \frac{\lambda^2\sigma^2}{2}\right). \tag{2.6}$$

Однією з найважливіших властивостей нормальної випадкової величини є її стійкість при композиції. А саме, нехай X_1 і X_2 — нормальні випадкові величини з параметрами m_1, m_2 і σ_1^2, σ_2^2. Тоді їх сума $\text{X}= \text{X}_1 + \text{X}_2$ буде також нормальною випадковою величиною з параметрами $m = m_1 + m_2$ і $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Для довільних значень параметрів m і σ табулювати функції $f_{\text{X}}(x)$ і $F_{\text{X}}(x)$ досить складно, тому користуються *стандартною нормальною величиною* (*стандартом*)

$$\text{Z} \;=\; \frac{\text{X} - m}{\sigma}, \tag{2.7}$$

для якої густина розподілу $f_{\text{Z}}(z)$ наступна:

$$f_{\text{Z}}(z) \;=\; \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \tag{2.8}$$

тобто її параметри дорівнюють $m_{\text{Z}} = 0$, $\sigma_{\text{Z}} = 1$ і, таким чином, $\text{Z} \to \mathcal{N}(0;\,1)$. Графік $f_{\text{Z}}(z)$, наведений на рис. 2.1, називають *кривою Гаусса*.

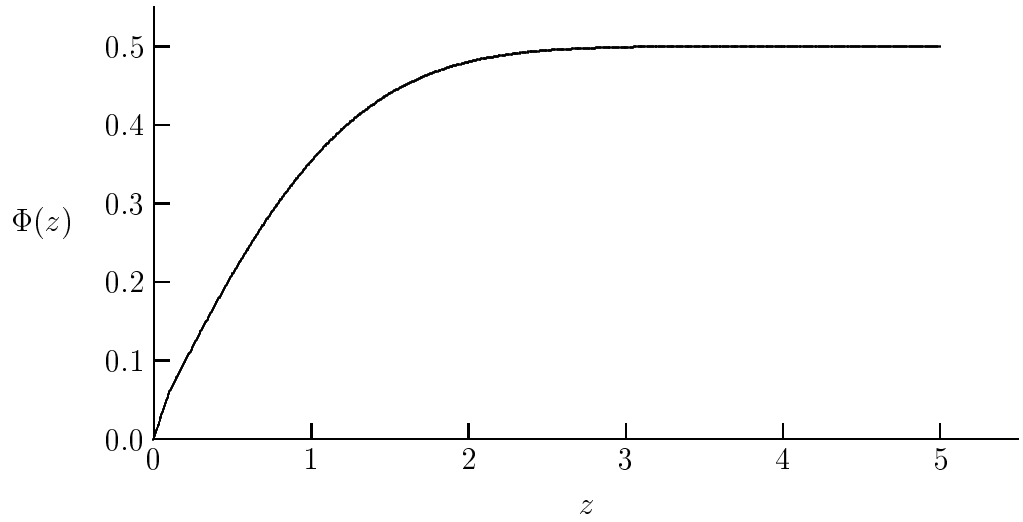


Рисунок 2.5 — Функція Лапласа $\Phi(z)$

Інтегральна функція розподілу $F_{\text{z}}(z)$ для стандартної нормальної величини визначається з виразу

$$F_{\text{z}}(z) \;=\; \Pr(\text{Z} < z) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{z} \exp\left(-\frac{u^2}{2}\right) du. \tag{2.9}$$

Внаслідок симетрії події $\text{Z} < 0$ та $\text{Z} > 0$ рівноймовірні, тому $F_{\text{z}}(0) = 0,5$ і

$$F_{\text{z}}(z) \;=\; 0,5 + \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{z} \exp\left(-\frac{u^2}{2}\right) du.$$

Нормальний розподіл часто використовується в статистичних обчисленнях.

Функція розподілу $F_{\text{z}}(z)$ залежить тільки від однієї змінної z. Оскільки інтеграл в (2.9) не виражається через елементарні функції, в практичних застосуваннях користуються *функцією Лапласа*

$$\Phi(z) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{z} \exp\left(-\frac{u^2}{2}\right) du, \tag{2.10}$$

і тому

$$F_{\text{z}}(z) \;=\; 0,5 + \Phi(z). \tag{2.11}$$

Функція Лапласа є непарною, $\Phi(-z) = -\Phi(z)$, тому при табулюванні наводять тільки невід'ємні значення аргументу z.

Для нормальної випадкової величини X можна записати при $x_1 \le x_2$

$$\Pr(x_1 \le \text{X} < x_2) \;=\; \Phi\left(\frac{x_2 - m}{\sigma}\right) - \Phi\left(\frac{x_1 - m}{\sigma}\right). \tag{2.12}$$

У тому випадку, коли межі x_1 та x_2 розташовуються симетрично відносно m, тобто $x_1 = m - \varepsilon$ та $x_2 = m + \varepsilon$, вираз (2.12) можна записати в більш компактній формі

$$\Pr(|\text{X} - m| < \varepsilon) \;=\; 2\Phi\left(\frac{\varepsilon}{\sigma}\right), \tag{2.13a}$$

або, якщо підставити $\varepsilon = z\sigma$,

$$\Pr(|\mathrm{X} - m| < z\sigma) \;=\; 2\Phi(z). \tag{2.13b}$$

Вважаючи в наведеній формулі $z = 1, 2, 3, 4$, з таблиць функції Лапласа знайдемо

$$\begin{aligned} \Pr(|\mathrm{X} - m| < \sigma) \;&=\; 2\Phi(1) \;=\; 0,683; \\ \Pr(|\mathrm{X} - m| < 2\sigma) \;&=\; 2\Phi(2) \;=\; 0,954; \\ \Pr(|\mathrm{X} - m| < 3\sigma) \;&=\; 2\Phi(3) \;=\; 0,9973; \\ \Pr(|\mathrm{X} - m| < 4\sigma) \;&=\; 2\Phi(4) \;=\; 0,99994. \end{aligned} \tag{2.14}$$

Цю рівність, звичайно, для $z = 3$ формулюють у вигляді *правила трьох сигм*: практично достовірно (тобто з імовірністю 0,9973), що відхилення нормально розподіленої випадкової величини від її математичного сподівання не перевищить за абсолютною величиною трьох дисперсій. Імовірність протилежної події $\Pr(|\mathrm{X} - m| > 3\sigma)$ при цьому становитиме 0,0027.

Визначення. *Квантилем* $u_{1-\alpha/2}$ стандартного нормального розподілу з густиною $f_{\mathrm{Z}}(z)$, що відповідає заданому рівню значущості α, називається таке значення, при якому виконується рівність

$$\Pr\left(|\mathrm{Z}| < u_{1-\alpha/2}\right) \;=\; \int_{-u_{1-\alpha/2}}^{u_{1-\alpha/2}} f_{\mathrm{Z}}(z)\,dz \;=$$

$$=\; F_{\mathrm{Z}}(u_{1-\alpha/2}) - F_{\mathrm{Z}}(-u_{1-\alpha/2}) \;=\; 2\Phi(u_{1-\alpha/2}) \;=\; 1 - \alpha. \tag{2.15}$$

З геометричної точки зору знаходження квантиля $u_{1-\alpha/2}$ полягає в такому виборі двох граничних значень z, при яких площа, обмежена зверху кривою густини $f_{\mathrm{Z}}(z)$, віссю абсцис знизу і вертикальними лініями, що проходять через точки $z = -u_{1-\alpha/2}$ та $z = u_{1-\alpha/2}$, дорівнювала б α. Іншими словами, для знаходження квантиля $u_{1-\alpha/2}$ при заданому рівні α необхідно розв'язати рівняння

$$2\Phi(u_{1-\alpha/2}) \;=\; 1 - \alpha. \tag{2.16}$$

Такий квантиль називається *двостороннім*. Можливо також використання *лівостороннього* та *правостороннього* квантилів. Лівосторонній квантиль $-u_{1-\alpha}$ шукають з умови

$$\Pr\left(\mathrm{Z} < -u_{1-\alpha}\right) \;=\; \int_{-\infty}^{-u_{1-\alpha}} f_{\mathrm{Z}}(z)\,dz \;=\; 0,5 - \Phi(u_{1-\alpha}) \;=\; \alpha, \tag{2.17a}$$

відповідно правосторонній квантиль $u_{1-\alpha}$ – з умови

$$\Pr\left(\mathrm{Z} > u_{1-\alpha}\right) \;=\; \int_{u_{1-\alpha}}^{\infty} f_{\mathrm{Z}}(z)\,dz \;=\; 0,5 + \Phi(u_{1-\alpha}) \;=\; 1 - \alpha. \tag{2.17b}$$

У випадку нормального розподілу лівосторонній і правосторонній квантилі рівні за модулем, це ж стосується обох двосторонніх квантилів. Конкретні їх значення можна знайти на основі формул (2.16) і (2.17) з таблиць функції Лапласа (див. додаток).

Детальніше техніка застосування нормальної випадкової величини при розв'язанні задач математичної статистики викладена нижче.

2.2. Системи нормальних випадкових величин

Нормальний закон розподілу для системи з двох випадкових величин (X,Y) (нормальний закон на площині) має густину вигляду

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-r_{\mathrm{XY}}^2}}\exp\left\{-\;\frac{1}{2(1-r_{\mathrm{XY}}^2)}\,Q(x,y)\right\}, \tag{2.18}$$

де

$$Q(x,y) \;=\; \frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}-\;2r_{\mathrm{XY}}\frac{(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}+\frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2},$$

а $m_{\mathrm{X}}, m_{\mathrm{Y}}$ – математичні сподівання випадкових величин X й Y; $\sigma_{\mathrm{X}}, \sigma_{\mathrm{Y}}$ – їх середньоквадратичні відхилення; r_{XY} – їх коефіцієнт кореляції.

Для випадкових величин, розподілених згідно з нормальним законом, некорельованість рівнозначна незалежності.

Якщо випадкові величини X, Y некорельовані (незалежні), то $r_{\mathrm{XY}}=0$ і

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}\exp\left\{-\,\frac{1}{2}\left(\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}\;+\;\frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right\}. \tag{2.19}$$

У цьому разі осі Ox, Oy називаються *головними осями розсіювання.*

Якщо при цьому $m_{\mathrm{X}} = m_{\mathrm{Y}} = 0$, то нормальний закон розподілу приймає *канонічний вигляд*:

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}\exp\left\{-\;\frac{x^2}{2\,\sigma_{\mathrm{X}}^2}\;-\;\frac{y^2}{2\,\sigma_{\mathrm{Y}}^2}\right\}. \tag{2.20}$$

Ймовірність влучення випадкової точки, розподіленої згідно з нормальним законом, в прямокутник R, що має сторони, паралельні головним осям розсіювання, виражається формулою

$$\Pr\{(\mathrm{X};\,\mathrm{Y})\in R\} \;= \tag{2.21}$$

$$=\;\left[\Phi\left(\frac{\beta-m_{\mathrm{X}}}{\sigma_{\mathrm{X}}}\right)-\Phi\left(\frac{\alpha-m_{\mathrm{X}}}{\sigma_{\mathrm{X}}}\right)\right]\left[\Phi\left(\frac{\delta-m_{\mathrm{Y}}}{\sigma_{\mathrm{Y}}}\right)-\Phi\left(\frac{\gamma-m_{\mathrm{Y}}}{\sigma_{\mathrm{Y}}}\right)\right],$$

де $x\in[\alpha,\beta),\;\; y\in[\gamma,\delta)$.

Еліпсом рівної густини (еліпсом розсіювання) називається еліпс, у всіх точках якого спільна густина $f(x,y)$ нормального закону постійна: $f(x,y)=const$. Півосі еліпса пропорційні σ_{X}, σ_{Y}: $a=k\sigma_{\mathrm{X}}$, $b=k\sigma_{\mathrm{Y}}$.

Ймовірність влучення випадкової точки (X,Y), розподіленої за нормальним законом, в область E_k, яка обмежена еліпсом розсіювання з півосями a і b, дорівнює

$$\Pr\{(\mathrm{X},\mathrm{Y})\in \mathrm{E}_k\} \;=\; 1-\exp(-k^2/2), \tag{2.22}$$

де k – розміри півосі еліпса, виражені в середніх квадратичних відхиленнях: $k=a/\sigma_{\mathrm{X}}=b/\sigma_{\mathrm{Y}}$.

Якщо $\sigma_{\mathrm{X}}=\sigma_{\mathrm{Y}}=\sigma$, розсіювання за нормальним законом називається *круговим.*

При круговому нормальному розсіюванні з $m_{\rm X} = m_{\rm Y} = 0$ відстань R від точки (X;Y) до початку координат (центра розсіювання) розподілена *згідно з законом Релея*

$$f(r) \;=\; \left(r/\sigma^2\right) \exp\left(-r^2/2\sigma^2\right), \qquad r \geq 0. \tag{2.23}$$

Розподілу Релея підкоряється модуль вектора на площині, якщо його ортогональні складові (проекції на координатні осі) незалежні і розподілені нормально з нульовими математичними сподіваннями і рівними дисперсіями.

Узагальненням закону розподілу Релея є *розподіл Релея-Райса*. При заданій константі s густина цього закону наступна:

$$f(r) \;=\; \left(r/\sigma^2\right) \exp\left(-r^2/2\sigma^2\right)\, I_0\left(rs/\sigma^2\right), \qquad r \geq 0, \tag{2.24}$$

де

$$I_0(x) \;=\; (2\pi)^{-1} \int\limits_0^{2\pi} \exp[x\cos(\varphi)]\, d\varphi \tag{2.24a}$$

– *модифікована функція Бесселя нульового індексу.*

Нормальний закон у просторі трьох вимірювань для незалежних випадкових величин X, Y, Z

$$f(x,y,z) \;= \tag{2.25}$$

$$= \;\frac{1}{(2\,\pi)^{3/2}\,\sigma_{\rm X}\,\sigma_{\rm Y}\,\sigma_{\rm Z}} \exp\left(-\,\frac{(x-m_{\rm X})^2}{2\,\sigma_{\rm X}^2} - \,\frac{(y-m_{\rm Y})^2}{2\,\sigma_{\rm Y}^2} - \,\frac{(z-m_{\rm Z})^2}{2\,\sigma_{\rm Z}^2}\right).$$

Ймовірність влучення випадкової точки (X, Y, Z) в тривимірну область E_k, обмежену еліпсоїдом рівної густини з півосями $a = k\sigma_{\rm X}$, $b = k\sigma_{\rm Y}$, $c = k\sigma_{\rm Z}$, дорівнює

$$\Pr\{({\rm X},{\rm Y},{\rm Z}) \in E_k\} \;=\; 2\Phi(k) - 1 - \sqrt{2/\pi}k\exp(-k^2/2). \tag{2.26}$$

Нормальний закон розподілу для системи X з n випадкових величин $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n)$ (нормальний закон в евклідовому просторі розмірності n) має густину вигляду

$$f(X) \;=\; \frac{1}{\sqrt{(2\pi)^n\,\det(K)}} \exp\left(-\,\frac{1}{2}(X - X_c)^{\rm T} K^{-1} (X - X_c)\right). \tag{2.27}$$

Вектор X, утворений з n-компонентного набору, розподілений згідно з багатовимірним нормальним законом з кореляційною матрицею K і вектором математичних сподівань $X_c = (X_{c1}, X_{c2}, \ldots, X_{cn})$.

З визначення (2.27) випливає, що математичне сподівання і дисперсія D[X] випадкового вектора X наступні:

$$\mathrm{M[X]} = X_c, \qquad \mathrm{D[X]} = \mathrm{M}[(\mathrm{X} - X_c)^{\rm T}(\mathrm{X} - X_c)] = \mathrm{Sp}(K). \tag{2.28}$$

Таким чином, дисперсія випадкової векторної величини X дорівнює сумі діагональних елементів $\mathrm{Sp}(K)$ кореляційної матриці K.

2.3. Гамма-функція та її властивості. Гамма-розподіл

I. Гамма-функцією або інтегралом Ейлера називається функція вигляду

$$\Gamma(\alpha) \;=\; \int\limits_0^\infty x^{\alpha-1}\exp(-x)\,dx. \tag{2.29}$$

Областю визначення гамма-функції є вся числова вісь, $-\infty < \alpha < \infty$, крім нуля і від'ємних цілих чисел, $\alpha = 0, -1, -2, \ldots$.

Гамма-функція (2.29) є інтегралом, що залежить від параметра α. Вона задовольняє наступним властивостям:

1) $\Gamma(1) \;=\; \Gamma(2) \;=\; 1$.

Дійсно, вважаючи в інтегралі (2.29) $\alpha = 1$ та інтегруючи, маємо

$$\Gamma(1) \;=\; \int\limits_0^\infty \mathrm{e}^{-x}\,dx \;=\; 1.$$

2) $\Gamma(\alpha+1) \;=\; \alpha\Gamma(\alpha)$ при $\alpha > 0$.

Дійсно, інтегруючи по частинах, знайдемо

$$\Gamma(\alpha+1) \;=\; \int\limits_0^\infty x^\alpha \mathrm{e}^{-x}\,dx \;=\; -x^\alpha \mathrm{e}^{-1}\Big|_0^\infty + \alpha\int\limits_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\,dx \;=\; \alpha\,\Gamma(\alpha).$$

Отже, якщо α – додатне ціле число, $\alpha = n$, то

$$\Gamma(n) \;=\; (n-1)! \qquad \text{або} \qquad \Gamma(n+1) \;=\; n! \tag{2.30}$$

Таким чином, гамма-функція може розглядатися як узагальнення факторіала. Через цю властивість гамма-функцію іноді називають факторіальною функцією.

Якщо ж аргумент α дорівнює нулю або від'ємному цілому числу, то значення гамма-функції розбігається.

3) $\Gamma\left(\frac{1}{2}\right) \;=\; \sqrt{\pi}$.

Дійсно,

$$\Gamma\left(\frac{1}{2}\right) \;=\; \int\limits_0^\infty x^{-1/2}\mathrm{e}^{-x}\,dx.$$

Скористаємося заміною $x = z^2$, тоді $dx = 2z\,dz$, що дає

$$\Gamma\left(\frac{1}{2}\right) \;=\; 2\int\limits_0^\infty z^{-1}\exp(-z^2)\,z\,dz \;=\; 2\int\limits_0^\infty \exp(-z^2)\,dz \;=\; \sqrt{\pi}. \tag{2.31}$$

Отже, якщо аргумент α пропорційний $\frac{1}{2}$, то $\Gamma(\alpha)$ може бути легко обчислена. Наприклад, $\Gamma(-1/2) = -2\sqrt{\pi}$.

При великих значеннях аргументу α значення гамма-функції обчислюється за формулою

$$\Gamma(\alpha) \;=\; (\alpha-1)\Gamma(\alpha-1) \;=\; (\alpha-1)(\alpha-2)\Gamma(\alpha-2) \;=\; \ldots \tag{2.32}$$

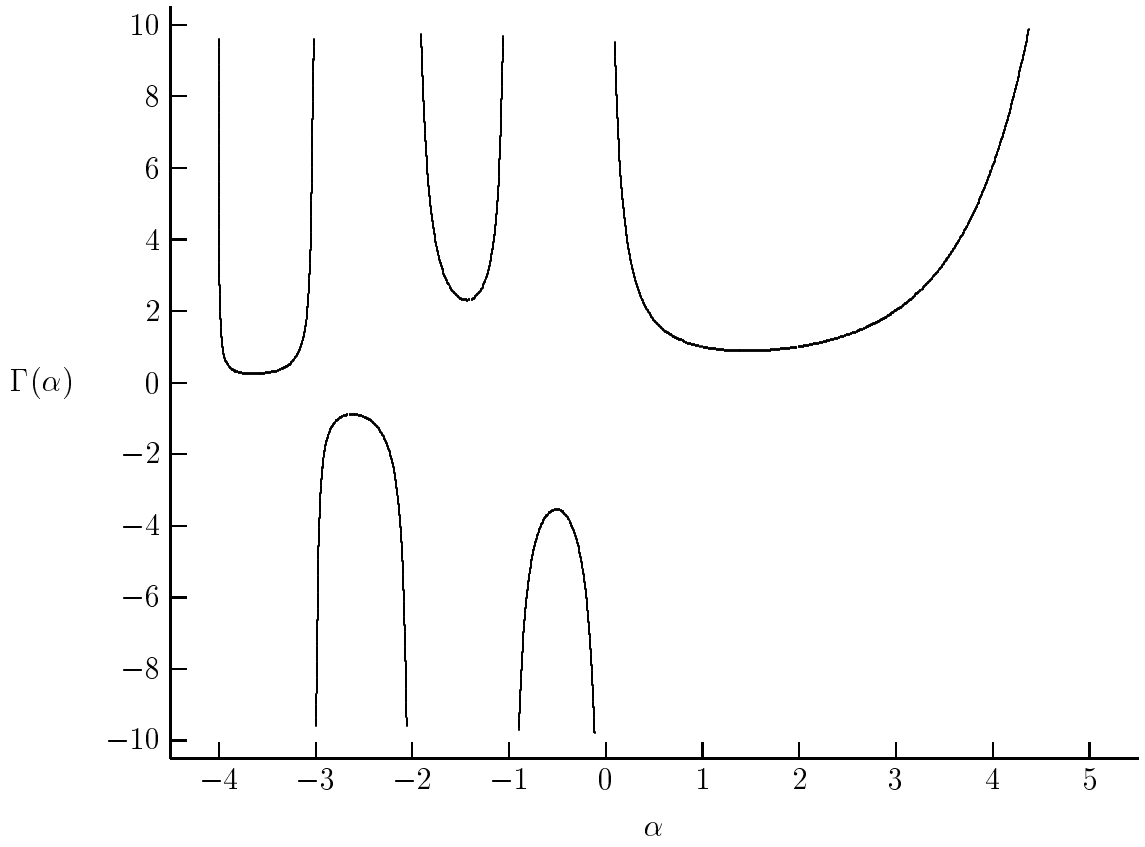


Рисунок 2.6 — Графік гамма-функції Ейлера $\Gamma(\alpha)$

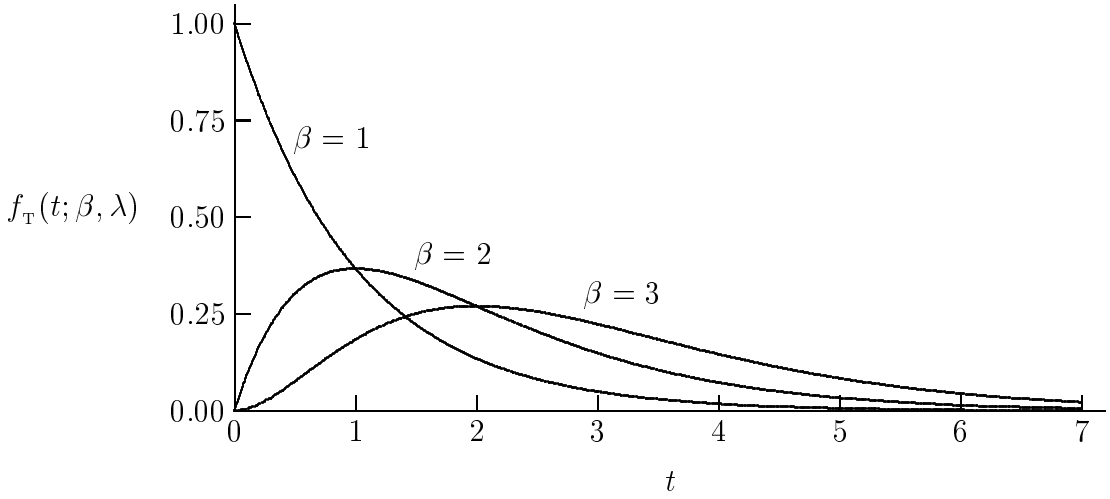


Рисунок 2.7 — Густина розподілу ймовірностей $f_т(t;\beta,\lambda)$ гамма-розподілу для $\lambda = 1{,}0$ та трьох значень параметра β

і так далі, доки аргумент гамма-функції не виявиться в інтервалі $(1;2)$. У цьому інтервалі гамма-функцію детально табулюють. Графік гамма-функції $\Gamma(\alpha)$ зображений на рис. 2.6.

II. Розповсюдженою практичною моделлю утворення *гамма-розподілу* є потік подій зі сталою інтенсивністю λ. При цьому випадкова величина T розглядається як час, необхідний для появи даної кількості β подій. Таким чином, як і в показниковому законі, величина T змінюється в інтервалі $(0,\infty)$. Можна показати, що цей розподіл буде задаватися густиною розподілу ймовірностей

$$f_{\mathrm{T}}(t;\beta,\lambda) \;=\; \frac{\lambda^{\beta}}{\Gamma(\beta)}\, t^{\beta-1}\mathrm{e}^{-\lambda t}, \qquad t \geq 0, \tag{2.33}$$

або функцією розподілу

$$\Pr(\mathrm{T} \leq t) \;=\; F_{\mathrm{T}}(t;\beta,\lambda) \;=\; \frac{\lambda^{\beta}}{\Gamma(\beta)} \int\limits_{0}^{t} u^{\beta-1}\mathrm{e}^{-\lambda u}\, du. \tag{2.34}$$

Розподіл (2.33) або (2.34) називається *гамма-розподілом* (рис. 2.7). Він залежить від свого аргументу t і параметрів β, λ.

Моменти першого і другого порядків визначаються з рівності

$$\mathrm{M}[\mathrm{T}] \;=\; \frac{\beta}{\lambda}, \qquad \mathrm{D}[\mathrm{T}] \;=\; \frac{\beta}{\lambda^2}. \tag{2.35}$$

З цих двох рівнянь можна отримати оцінки параметрів гамма-розподілу за здобутими даними:

$$\lambda^{*} \;=\; \mathrm{M}[\mathrm{T}]^{*} \,/\, \mathrm{D}[\mathrm{T}]^{*}; \tag{2.36a}$$

$$\beta^{*} \;=\; \mathrm{M}[\mathrm{T}]^{*2} \,/\, \mathrm{D}[\mathrm{T}]^{*} \;=\; \lambda^{*}\mathrm{M}[\mathrm{T}]^{*}. \tag{2.36b}$$

При певних значеннях параметра β отримаємо окремі випадки гамма-розподілу. Так, при $\beta = 1$ гамма-розподіл збігається з показниковим. При β – позитивному цілому числі отримуємо *розподіл Ерланга*, що широко використовується в теорії масового обслуговування.

Характеристична функція гамма-розподілу наступна:

$$q_{\mathrm{T}}(z) \;=\; \mathrm{M}\left[\exp(iz\mathrm{T})\right] \;=\; (1 - iz/\lambda)^{-\beta}\,. \tag{2.37}$$

2.4. Розподіл χ^2 (хі-квадрат)

Цей розподіл пов'язаний з нормальним розподілом Гаусса і широко використовується при розв'язанні різних задач статистичного аналізу. В основі утворення цього розподілу полягає розгляд вибірки з нормальної сукупності.

Розглянемо випадкову величину Y, яка розподілена згідно з нормальним законом з математичним сподіванням $\mathrm{M}[\mathrm{Y}] = a$ і середнім квадратичним відхиленням σ, або, більш стисло, нехай $\mathrm{Y} \to \mathcal{N}(a;\,\sigma)$.

Тоді випадкова величина $\mathrm{U} = (\mathrm{Y} - a)/\sigma$, яку називають *стандартизованою випадковою величиною*, розподілена згідно з нормальним законом з параметрами $\mathrm{M}[\mathrm{U}] = 0$ та $\sigma_{\mathrm{U}} = 1$, тобто $\mathrm{U} \to \mathcal{N}(0;\, 1)$.

Квадрат стандартизованої випадкової величини

$$\mathrm{U}^2 \;=\; \left(\frac{\mathrm{Y} - a}{\sigma}\right)^2 \;\equiv\; \chi^2 \tag{2.38}$$

називається *випадковою величиною χ^2 (хі-квадрат) з одним ступенем вільності.*

Розглянемо тепер n незалежних випадкових величин $\mathrm{Y}_1,\, \mathrm{Y}_2,\, \ldots,\, \mathrm{Y}_n$, розподілених згідно з нормальним законом з математичними сподіваннями $a_1, a_2, \ldots, a_n$ і середніми квадратичними відхиленнями $\sigma_1, \sigma_2, \ldots, \sigma_n$. Утворимо для кожної з цих випадкових величин *стандартизовану* нормальну випадкову величину

$$\mathrm{U}_i \;=\; \frac{\mathrm{Y}_i - a_i}{\sigma_i}, \qquad\qquad i = 1, 2, \ldots, n.$$

Сума квадратів стандартизованих змінних

$$\chi^2 \;=\; \mathrm{U}_1^2 + \mathrm{U}_2^2 \ldots + \mathrm{U}_n^2 \;=\; \left(\frac{\mathrm{Y}_1 - a_1}{\sigma_1}\right)^2 + \left(\frac{\mathrm{Y}_2 - a_2}{\sigma_2}\right)^2 + \ldots + \left(\frac{\mathrm{Y}_n - a_n}{\sigma_n}\right)^2 \tag{2.39}$$

називається *випадковою величиною χ^2 з $\nu = n$ ступенями вільності.* У статистичних таблицях і при виконанні розрахунків кількість ступенів вільності прийнято позначати літерою ν.

Густина розподілу випадкової величини χ^2 має вигляд

$$f(t) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1}\, t^{\nu/2-1}\, \exp\left(-t/2\right), \tag{2.40}$$

при цьому $t \geq 0$.

Іноді зручно густину розподілу вказувати безпосередньо в термінах χ^2:

$$f(\chi^2)\, d(\chi^2) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1}\, (\chi^2)^{\nu/2-1}\, \exp\left(-\chi^2/2\right)\, d(\chi^2). \tag{2.40$'$}$$

Порівнюючи густину розподілу (2.40) з раніше наведеною густиною (2.33), приходимо до висновку, що розподіл χ^2 являє собою окремий випадок гамма-розподілу при $\beta = \dfrac{n}{2}$ і $\lambda = \dfrac{1}{2}$.

Інтегральна функція χ^2-розподілу має вигляд

$$F(\chi^2) \;=\; \Pr(\mathrm{X} \leq \chi^2) \;=\; \left[2^{\nu/2}\,\Gamma(\nu/2)\right]^{-1} \int_0^{\chi^2} x^{\nu/2-1}\, \exp\left(-x/2\right)\, dx. \tag{2.41}$$

Таким чином, розподіл χ^2 залежить від одного параметра ν – кількості ступенів вільності.

На рис. 2.8–2.9 зображені приклади графіків густини розподілу ймовірностей $f(\chi^2)$ та інтегральної функції χ^2-розподілу $F(\chi^2)$ відповідно. Як це видно з графіків для $f(\chi^2)$, густина χ^2-розподілу асиметрична, вона має правий подовжений "хвіст" (в

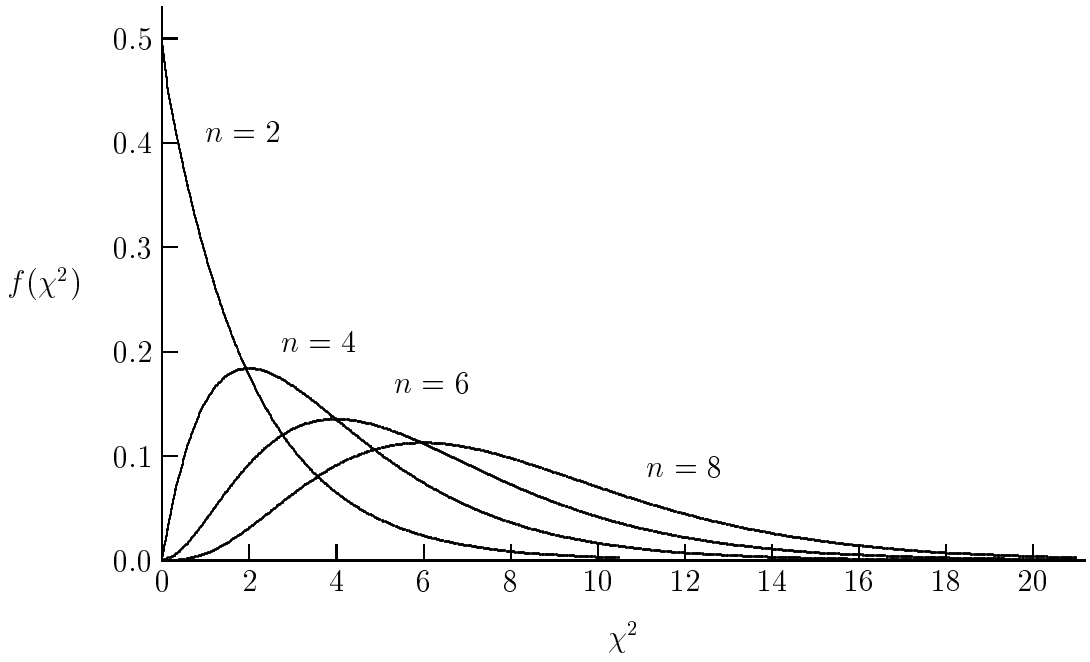


Рисунок 2.8 — Густина розподілу ймовірностей χ^2 для кількості ступенів вільності $n = 2,\ 4,\ 6,\ 8$

периферійній області великих відхилень χ^2). Зі зростанням n (або ν) асиметричність густини зменшується, при цьому закон розподілу прямує до нормального.

Розрахувавши перші два моменти χ^2, отримаємо

$$\mathrm{M}[\chi^2] \;=\; \nu, \qquad \mathrm{D}[\chi^2] \;=\; 2\nu. \tag{2.42}$$

Розподіл χ^2 часто використовується у статистичних обчисленнях, зокрема, у зв'язку з наступною теоремою.

Т е о р е м а. *Нехай $x_1, x_2, \ldots, x_n$ – задана вибірка з нормально розподіленої генеральної сукупності $\mathcal{N}(m, \sigma)$ обсягу n, та при цьому $x^* = \frac{1}{n}\sum_{i=1}^{n} x_i$ й $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x^*)^2$ – відповідно вибіркове арифметичне середнє і вибіркова дисперсія. Тоді статистики X^* та S^2 – незалежні випадкові величини, причому статистика $(n-1)\sigma^{-2}\mathrm{S}^2$ має розподіл χ^2_{n-1}.*

Характеристична функція розподілу χ^2 з кількістю ступенів вільності $n = \nu$ наступна:

$$q_\nu(z) \;=\; \mathrm{M}\left[\exp(iz\chi^2_\nu)\right] \;=\; (1-2iz)^{-\nu/2}. \tag{2.43}$$

Нехай розглядається випадкова величина χ^2_n з n ступенями вільності та випадкова величина χ^2_m з m ступенями вільності, а також адитивна випадкова величина $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ з $n+m$ ступенями вільності. З вигляду характеристичної функції (2.43) випливає, що

$$q_{n+m}(z) \;=\; q_n(z)\, q_m(z) \;=\; (1-2iz)^{-(n+m)/2}, \tag{2.44}$$

тому густина розподілу ймовірностей композиції $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ така:

$$f_{n+m}(t) \;=\; \left[2^{(n+m)/2}\,\Gamma\left(\frac{n+m}{2}\right)\right]^{-1} t^{(n+m)/2-1}\,\exp(-t/2). \tag{2.45}$$

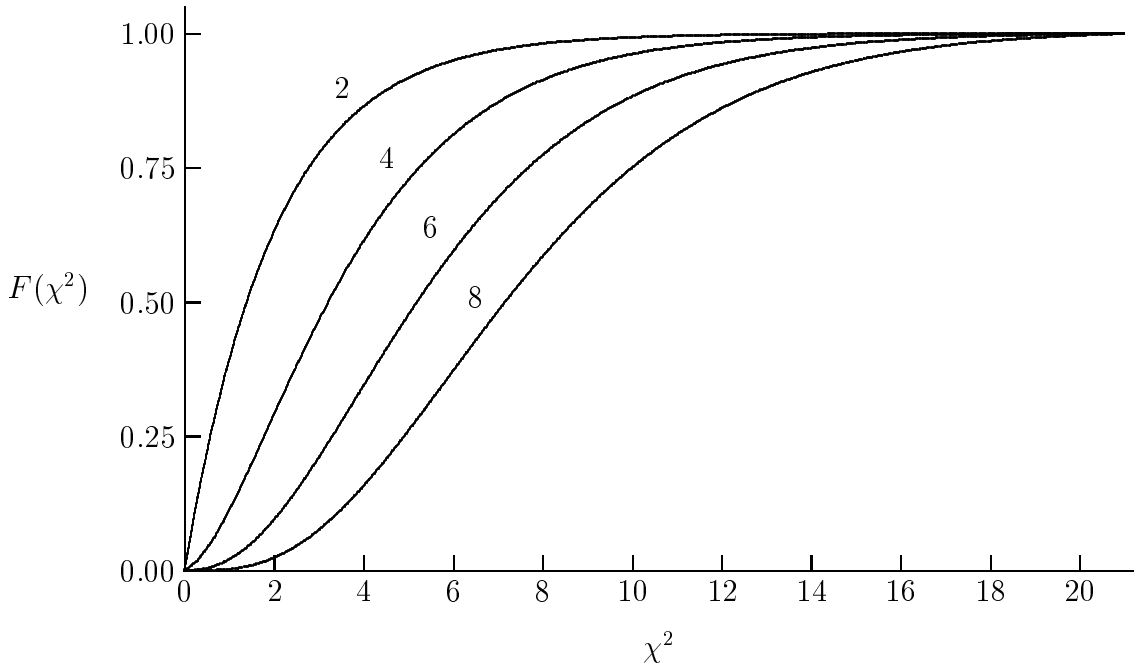


Рисунок 2.9 — Інтегральний закон розподілу χ^2 для кількості ступенів вільності $n = 2,\ 4,\ 6,\ 8$ (вказані цифрами на кривих)

Таким чином, адитивна випадкова величина $\chi^2_{n+m} = \chi^2_n + \chi^2_m$ також описується законом χ^2 з $n + m$ ступенями вільності, тобто випадкова величина χ^2 володіє властивістю стійкості при композиції.

Нехай є n квадратів стандартизованих змінних $\mathrm{U}^2_1,\ \mathrm{U}^2_2,\ \ldots,\ \mathrm{U}^2_n,$ на які, в свою чергу, накладені s лінійних залежностей (зв'язків). Тоді розподіл їх суми буде підкорюватися закону (2.40), але з кількістю ступенів вільності, що дорівнює $\nu = n - s$.

На практиці використовується не сама густина розподілу ймовірностей або функція розподілу, а *квантилі* χ^2-*розподілу*, що відповідають *рівню значущості* α, які при заданому ν означають $\chi^2_{\alpha,\nu}$.

Визначення. *Квантилем* $\chi^2_{\alpha,\nu}$, що відповідає заданому рівню значущості α, називається таке значення $\chi^2 = \chi^2_{\alpha,\nu}$, при якому виконується рівність

$$\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) \;=\; 1 - F(\chi^2_{\alpha,\nu}) \;=\; \int_{\chi^2_{\alpha,\nu}}^{\infty} f(\chi^2)\, d(\chi^2) \;=\; \alpha. \qquad (2.46)$$

З геометричної точки зору знаходження квантиля $\chi^2_{\alpha,\nu}$ полягає в такому виборі значення $\chi^2 = \chi^2_{\alpha,\nu}$, при якому площа, обмежена зверху кривою густини $f(\chi^2)$, віссю абсцис знизу і вертикальною лінією, що проходить через точку $\chi^2 = \chi^2_{\alpha,\nu}$, дорівнювала б α. Іншими словами, для знаходження квантиля $\chi^2_{\alpha,\nu}$ при заданих α і ν необхідно розв'язати рівняння (2.46).

На рис. 2.10 (зверху), на якому наведена густина $f(\chi^2)$ для випадку $\alpha = 0,20$ і $\nu = 10$, ця площа чисельно дорівнює рівню значущості α і заштрихована вертикальними лініями.

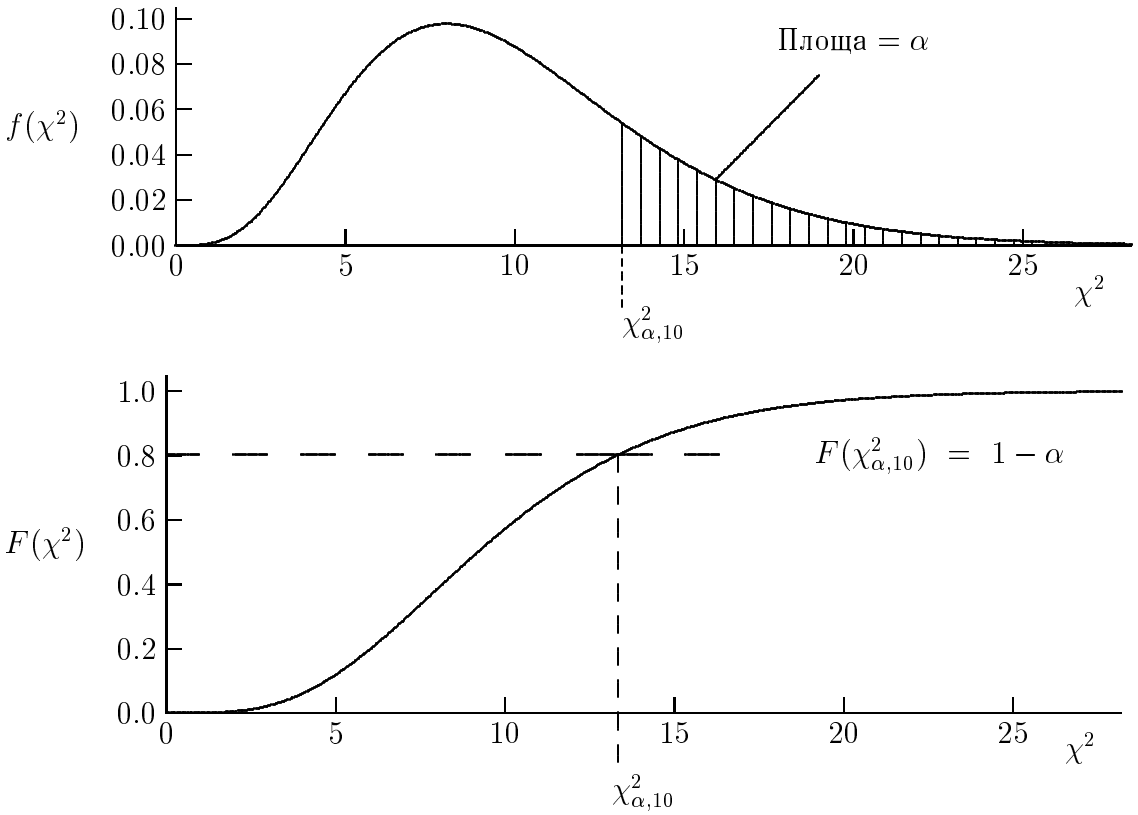


Рисунок 2.10 — Густина розподілу ймовірностей $f(\chi^2)$ (зверху) і інтегральний закон розподілу $F(\chi^2)$ (знизу) для кількості ступенів вільності $\nu = 10$ ($\alpha = 0,2$; квантиль $\chi^2_{\alpha,\nu} = \chi^2_{0,2,10}$ дорівнює 13,20)

На тому ж рисунку (знизу) наведена інтегральна функція розподілу $F(\chi^2)$, розглянутий той же випадок ($\alpha = 0,20$ і $\nu = 10$). На цьому рисунку пунктиром відмічений рівень значущості α, що відповідає квантилю $\chi^2_{\alpha,\nu}$.

У додатку наведена таблиця значень квантилів $\chi^2_{\alpha,\nu}$ для різних значень кількості ν ступенів вільності і рівня значущості α.

2.5. Розподіл Стьюдента

Розподіл Стьюдента (t-розподіл) має велике значення при статистичних обчисленнях, пов'язаних з нормальним законом, а саме тоді, коли середнє квадратичне відхилення σ невідоме і ще підлягає визначенню за дослідними даними (рис. 2.11 та 2.12).

Нехай Y і Y_1, Y_2, ..., Y_n – незалежні випадкові величини, що мають нормальний розподіл з параметрами

$$M[Y] = M[Y_1] = M[Y_2] = \ldots = M[Y_n] = 0,$$

$$\sigma[Y] = \sigma[Y_1] = \sigma[Y_2] = \ldots = \sigma[Y_n] = 1.$$

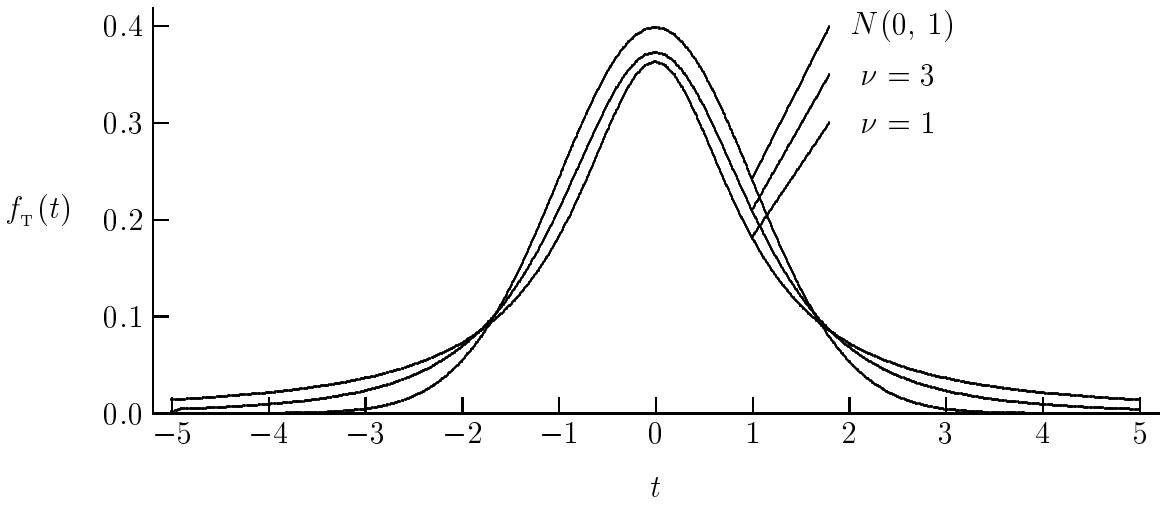


Рисунок 2.11 — Густина розподілу Стьюдента $f_{\mathrm{T}}(t)$ при кількостях ступенів вільності $\nu = 1$, $\nu = 3$ і $\nu = \infty$ – крива нормального розподілу $N(0;\,1)$

Випадкова величина

$$\mathrm{T} \;=\; \mathrm{Y}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Y}_i^2\right)^{-1/2} \;=\; \mathrm{Y}\left(\frac{1}{n}\chi_n^2\right)^{-1/2}, \tag{2.47}$$

що є функцією нормально розподілених випадкових величин, називається *безрозмірним дробом Стьюдента*.

У курсі теорії ймовірностей говориться, що густина розподілу ймовірностей випадкової величини T має вигляд (див. приклади до розділу)

$$f_{\mathrm{T}}(t) \;=\; S(t,\nu) \;=\; \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+t^2/\nu\right)^{-(\nu+1)/2}, \qquad -\infty < t < \infty. \tag{2.48}$$

Цей розподіл непараметричний, тобто він не залежить від параметрів початкових випадкових величин, а залежить лише від їх кількості.

У формулі (2.48) літерою ν позначена кількість доданків у підкореневому виразі дробу Стьюдента, тобто $\nu = n$. Таке позначення кількості ступенів вільності загальноприйняте в математичній статистиці, оскільки полегшує користування статистичними таблицями.

З формули (2.48) випливає, що розподіл випадкової величини T не залежить від параметрів розподілу незалежних випадкових величин Y і $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n$, а залежить тільки від одного параметра – кількості ступенів вільності ν, які дорівнюють кількості доданків у підкореневому виразі дробу Стьюдента (2.47).

Математичне сподівання і дисперсія випадкової величини T відповідно дорівнюють

$$\mathrm{M}[\mathrm{T}] \;=\; 0, \qquad \mathrm{D}[\mathrm{T}] \;=\; \frac{\nu}{\nu-2}, \qquad \nu > 2. \tag{2.49}$$

При необмеженому збільшенні кількості ступенів вільності $n \gg 1$ розподіл Стьюдента асимптотично переходить у нормальний розподіл Гаусса з параметрами $\mathrm{M}[\mathrm{T}] = 0$, $\mathrm{D}[\mathrm{T}] = 1$.

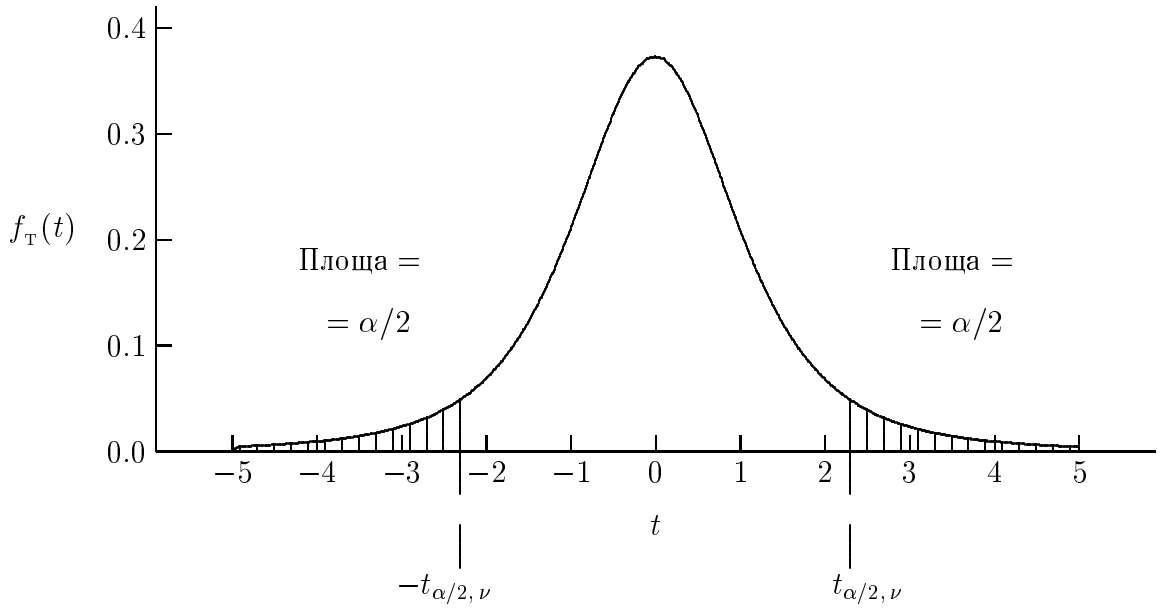


Рисунок 2.12 — Густина ймовірностей $f_{\mathrm{T}}(t)$ розподілу Стьюдента для кількості ступенів вільності $\nu = 3$ (заштрихована площа дорівнює α)

На рис. 2.11 зображений графік густини розподілу Стьюдента для різної кількості ступенів вільності. Аналізуючи цей графік, помічаємо, що зі збільшенням кількості ступенів вільності ν він наближується до кривої Гаусса. Якщо ж кількість ступенів вільності ν мала, то ймовірності великих відхилень трохи більше в порівнянні з нормальним законом (при $\mathrm{T} > 2$ крива t–розподілу розташовується вище нормальної кривої).

У математичній статистиці досить часто використовуються квантилі $t_{\alpha/2;\,\nu}$ розподілу Стьюдента залежно від кількості ν ступенів вільності і заданого рівня ймовірності α.

Значення квантилів розподілу Стьюдента $t_{\alpha/2;\,\nu}$ можна знайти з розв'язку рівняння

$$\Pr\left(|\mathrm{T}| > t_{\alpha/2;\,\nu}\right) \;=\; 2\int\limits_{t_{\alpha/2;\,\nu}}^{\infty} f_{\mathrm{T}}(t')\,dt' \;=\; \alpha. \qquad (2.50)$$

З геометричної точки зору знаходження квантилів розподілу Стьюдента $t_{\alpha/2;\,\nu}$ складається в такому виборі значення $t_{\alpha/2;\,\nu}$, при якому сумарна площа під кривою густини $f_{\mathrm{T}}(t)$ на дільницях $\left(-\infty,\, -t_{\alpha/2;\,\nu}\right)$ та $\left(t_{\alpha/2;\,\nu}, \infty\right)$ дорівнювала б величині рівня значущості α. На рисунку 2.12 сумарна площа заштрихованих двох дільниць складає α.

2.6. Розподіл Фішера

Розподіл Фішера (F-розподіл) використовується при порівнянні дисперсій нормальних розподілів, обчислених на основі дослідних даних. (Цей закон ще часто називають розподілом Фішера-Снедекора).

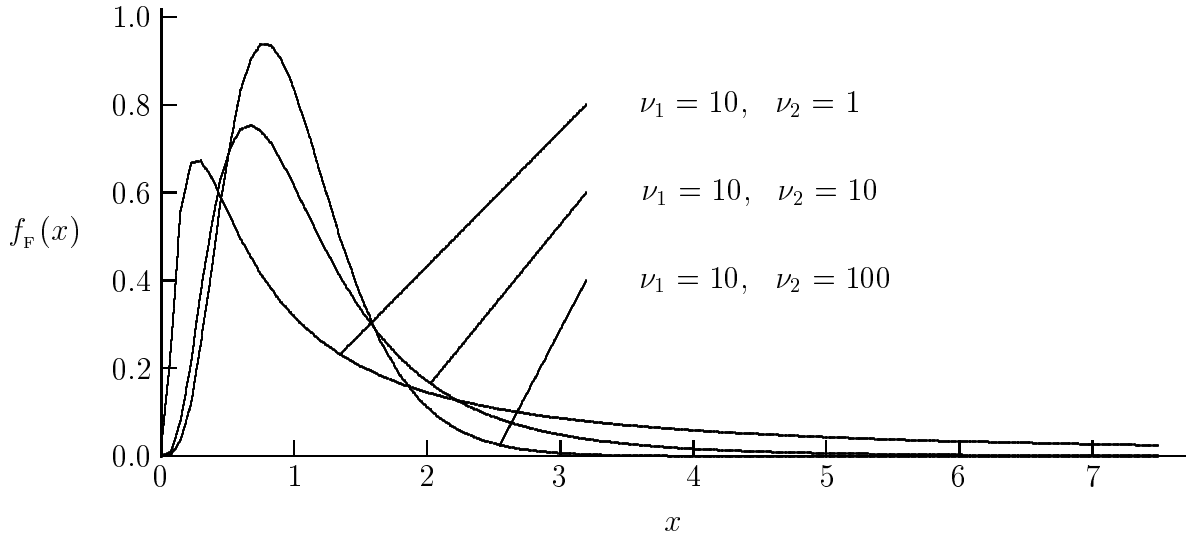


Рисунок 2.13 — Густина розподілу Фішера $f_{\mathrm{F}}(x)$ для різних кількостей ступенів вільності ν_1 та ν_2

Нехай випадкові величини X_1, X_2, ..., X_m й Y_1, Y_2, ..., Y_n незалежні і мають нормальний розподіл з параметрами:

$$\mathrm{M}[\mathrm{X}_i] = 0; \qquad \mathrm{D}[\mathrm{X}_i] = 1; \qquad i = 1, 2, \ldots, m;$$
$$\mathrm{M}[\mathrm{Y}_j] = 0; \qquad \mathrm{D}[\mathrm{Y}_j] = 1; \qquad j = 1, 2, \ldots, n.$$

Безрозмірна випадкова величина

$$\mathrm{F} \;=\; \left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{X}_i^2\right)\left(\frac{1}{n}\sum_{j=1}^{n}\mathrm{Y}_j^2\right)^{-1} \tag{2.51}$$

розподілена відповідно до закону Фішера (рис 2.13), тобто має густину розподілу ймовірностей $(x \geq 0)$

$$f_{\mathrm{F}}(x) \;=\; \frac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{-1+\nu_1/2}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-(\nu_1+\nu_2)/2}, \tag{2.52}$$

де $\nu_1 = m$ – кількість ступенів вільності чисельника; $\nu_2 = n$ – кількість ступенів вільності знаменника.

З формули (2.52) випливає, що розподіл випадкової величини F залежить від двох параметрів – *кількості ступенів вільності* $\nu_1 = m$ і $\nu_2 = n$. Для зручності користувачів статистичними таблицями кількості ступенів вільності часто позначаються літерами $\nu_1 = m$ і $\nu_2 = n$. Графік густини ймовірностей F-розподілу зображений на рисунку 2.13.

У математичній статистиці досить часто використовуються квантилі F-розподілу $f_{\alpha;\nu_1,\nu_2}$ залежно від кількості ступенів вільності ν_1 і ν_2 та заданого рівня ймовірності α.

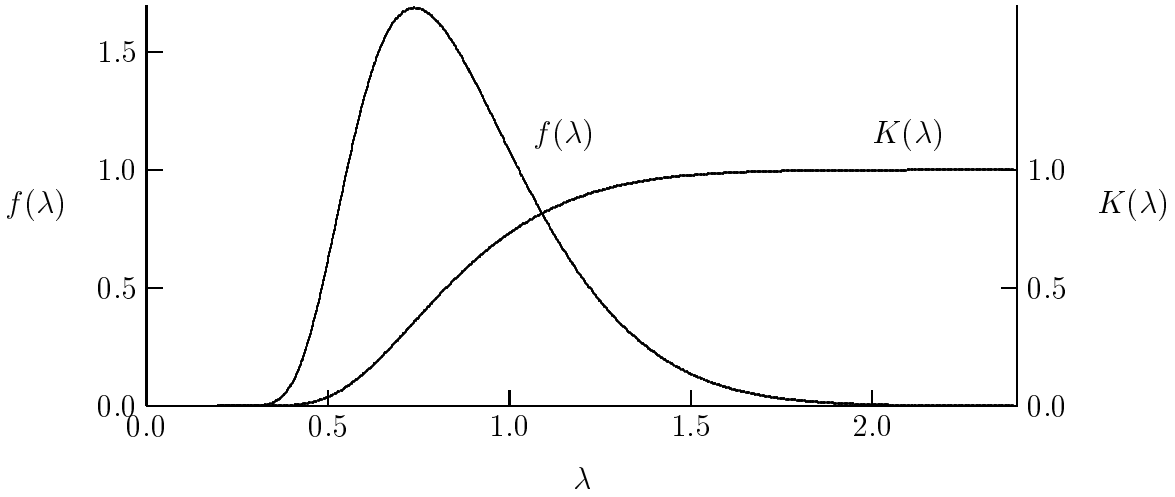


Рисунок 2.14 — Густина розподілу ймовірностей $f(\lambda)$ та інтегральний закон розподілу $K(\lambda)$ випадкової величини, що розподілена за законом Колмогорова

З розв'язку рівняння

$$\Pr\left(|\mathrm{F}| > f_{\alpha;\nu_1,\nu_2}\right) = 2\int\limits_{f_{\alpha;\nu_1,\nu_2}}^{\infty} f_{\mathrm{F}}(x)\,dx = \alpha \tag{2.53}$$

можуть бути знайдені значення квантилів $f_{\alpha;\nu_1,\nu_2}$ розподілу Фішера.

2.7. Розподіл Колмогорова

У практичних задачах перевірки гіпотез про згоду даних вибірки з конкретним теоретичним законом розподілу для будь-якої неперервної випадкової величини застосовується λ-критерій Колмогорова (рис. 2.14) та критерій Смірнова-Колмогорова (рис. 2.15).

Виникаюча при перевірці такої гіпотези випадкова величина (вибіркова статистика) Λ має інтегральну функцію розподілу $K(\lambda)$ вигляду

$$K(\lambda) = \Pr\left(\Lambda < \lambda\right) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2\lambda^2) \tag{2.54}$$

і густину розподілу ймовірностей (рис. 2.14)

$$f(\lambda) = \frac{d}{d\lambda}K(\lambda) = 4\lambda \sum_{k=-\infty}^{\infty} (-1)^k k^2 \exp(-2k^2\lambda^2). \tag{2.55}$$

Також, як і розподіл Стьюдента, цей розподіл непараметричний, тобто він не залежить від будь-якого параметра.

В односторонніх критеріях згоди вибіркової статистики Λ із заданим теоретичним розподілом також використовується розподіл Смірнова-Колмогорова, у якого інтегральна функція розподілу $K(\lambda)$ наступна:

$$K(\lambda) = \Pr\left(\Lambda < \lambda\right) = 1 - \exp(-2\lambda^2), \tag{2.56}$$

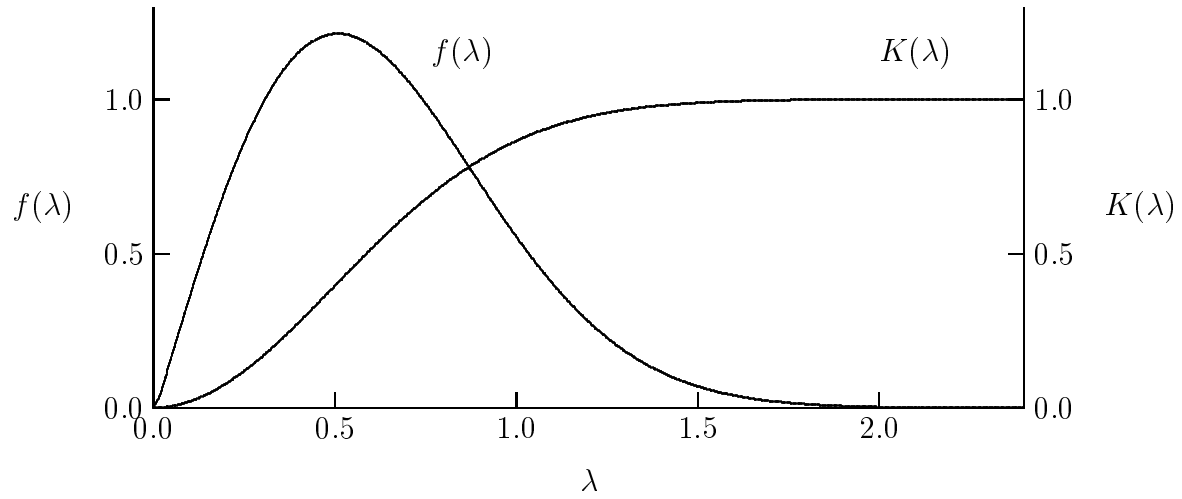


Рисунок 2.15 — Густина розподілу ймовірностей $f(\lambda)$ та інтегральний закон розподілу $K(\lambda)$ випадкової величини, розподіленої згідно із законом Смірнова-Колмогорова

з відповідною їй густиною розподілу ймовірностей (рис. 2.15)

$$f(\lambda) \;=\; \frac{d}{d\lambda}K(\lambda) \;=\; 4\lambda\exp(-2\lambda^2). \tag{2.57}$$

Цей розподіл також застосовується для перевірки гіпотези про те, що дві вибірки витягнуті з однієї і тієї ж генеральної сукупності.

У багатьох задачах математичної статистики (при оцінюванні або перевірці статистичних гіпотез) використовуються також і інші закони розподілу.

2.8. Розподіл Бернуллі

Випадкова величина X називається *розподіленою щодо біноміального закону Бернуллі*, якщо вона є кількістю m появ випадкової події у n випробуваннях, за умови, що в одному (будь-якому) досліді ця подія настає з заданою ймовірністю p.

Закон розподілу випадкової величини X:

$$\mathrm{P}_{m,n} \;=\; \mathrm{P}\{\mathrm{X}=m\} \;=\; C_n^m p^m q^{n-m}, \tag{2.58a}$$

або

$$\mathrm{P}_{m,n} \;=\; \frac{n\,!}{m\,!\,(n-m)\,!}\,p^m q^{n-m}, \tag{2.58b}$$

де $0<p<1,\ q=1-p;\ m=0,1,\ldots,n$. Розподіл (2.58) залежить від двох параметрів: n (рис. 2.16) та p (рис. 2.17).

З теореми про повторення дослідів випливає, що *кількість* X *появ події при* n *незалежних дослідах має біноміальний розподіл.* Таким чином, імовірність однієї складної події, яка полягає в тому, що у n випробуваннях елементарна подія A настане m разів і не настане $n-m$ разів, дорівнює $\mathrm{P}_{m,n}$.

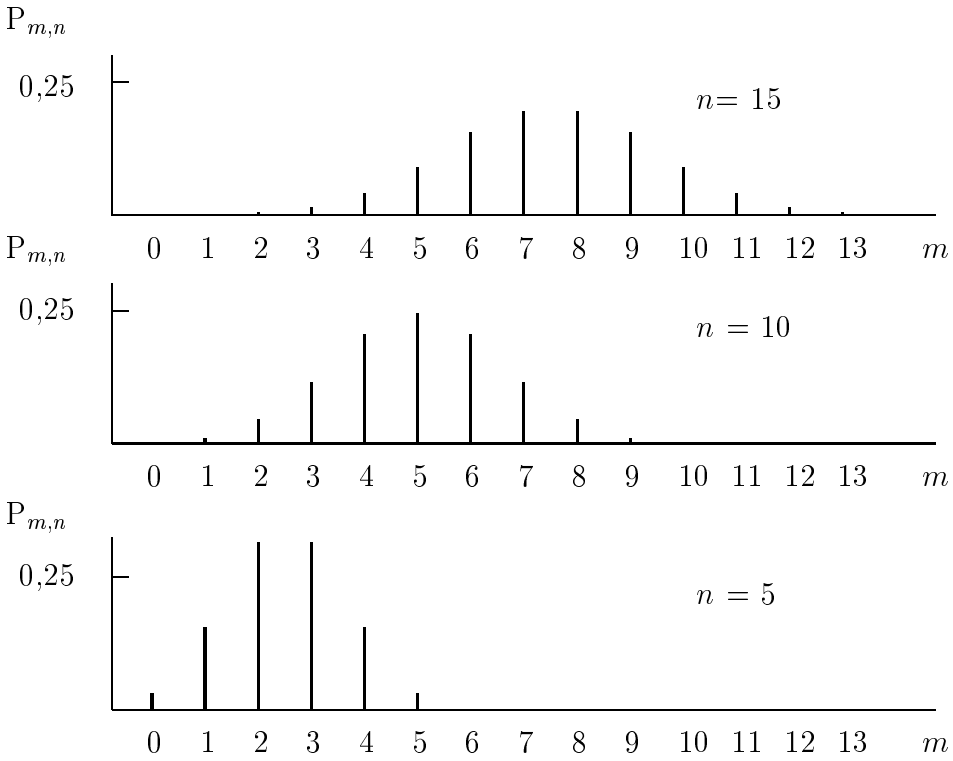


Рисунок 2.16 — Розподіл Бернуллі; $n = 5; 10; 15$; $p = 0{,}5$

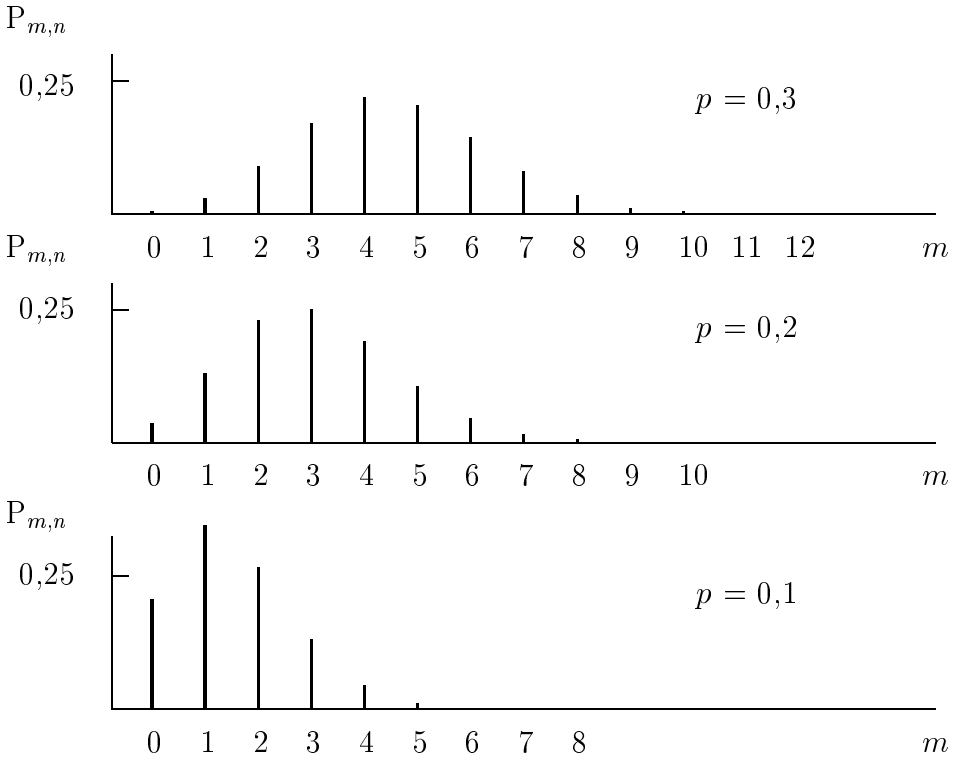


Рисунок 2.17 — Розподіл Бернуллі; $p = 0{,}1; 0{,}2; 0{,}3$; $n = 15$

Для випадкової величини X, що має біноміальний розподіл з параметрами p й n,

$$\mathrm{M[X]} \;=\; np, \qquad \mathrm{D[X]} \;=\; npq, \tag{2.59}$$

де $q = 1 - p$.

Користуватися формулою Бернуллі (2.58) за великих значень параметра n достатньо важко, бо вона вимагає виконання дій над великими числами.

Мають місце асимптотичні формули, що дозволяють наблизити ймовірність появи події, що дорівнює рівно m разів у n випробуваннях, якщо кількість випробувань достатньо велика.

Відповідні вирази формулюються у вигляді локальної та інтегральної теорем Муавра-Лапласа.

Локальна форма теореми Муавра-Лапласа:

Якщо ймовірність p появи події A *у кожному випробуванні стала та відмінна від нуля та одиниці, то ймовірність* $\mathrm{P}_n(k)$ *того, що подія* A *відбудеться у n випробуваннях рівно k разів, буде приблизно*

$$\mathrm{P}_n(k) \;=\; \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right), \tag{2.60}$$

тим точніше, чим більше кількість випробувань n.

Інтегральна форма теореми Муавра-Лапласа:

Якщо ймовірність p появи події A *у кожному випробуванні стала та відмінна від нуля та одиниці, то ймовірність* $\mathrm{P}_n(k_1, k_2)$ *того, що подія* A *настане у n випробуваннях у діапазоні від k_1 до k_2 разів, буде приблизно*

$$\mathrm{P}_n(k_1, k_2) \;=\; \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} \exp(-z^2/2)\, dz, \tag{2.61}$$

де $\alpha = (k_1 - np)/\sqrt{npq}$ *и* $\beta = (k_2 - np)/\sqrt{npq}$.

2.9. Розподіл Пуассона

Дискретна випадкова величина X називається *розподіленою щодо закону Пуассона*, якщо її можливі значення 0, 1, 2, ..., m, ..., а ймовірність випадкової події $\{\mathrm{X} = m\}$ виражається формулою

$$\mathrm{P}_m \;=\; \Pr\{\mathrm{X} = m\} \;=\; \frac{\lambda^m}{m\,!} \exp(-\lambda), \tag{2.62}$$

де $\lambda > 0$.

Розподіл Пуассона залежить від одного параметра λ (рис. 2.18).

Для випадкової величини X, розподіленої за законом Пуассона, середнє значення та дисперсія такі:

$$\mathrm{M[X]} \;=\; \lambda, \qquad \mathrm{D[X]} \;=\; \lambda. \tag{2.63}$$

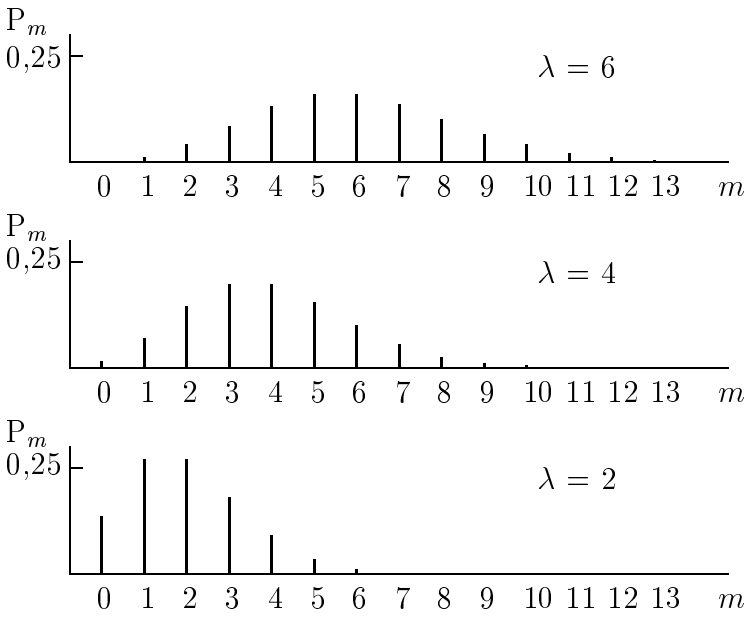


Рисунок 2.18 — Розподіл Пуассона; $\lambda = 2;\ 4;\ 6$

Пуассонівський розподіл є граничним для біноміального розподілу (2.58) при $p \to 0$, $n \to \infty$, за умови, що добуток $np = \lambda = const$. Цим розподілом можна користуватися приблизно в тому випадку, коли виконується велика кількість незалежних дослідів, у кожному з яких подія A відбувається з малою ймовірністю.

Пуассонівському закону розподілу підкоряється також кількість точок, що влучають в деяку задану область простору (одновимірного, двовимірного або тривимірного), якщо випадкове розміщення точок у цьому просторі задовольняє деяким обмеженням.

Одновимірний варіант зустрічається при розгляді "потоків подій". *Потоком подій* називається послідовність однорідних подій, що відбуваються одна за одною у випадкові моменти часу.

Середня кількість подій λ, що відбуваються за одиницю часу, називається *інтенсивністю потоку*. Величина λ може бути як сталою, так і змінною: $\lambda = \lambda(t)$.

Потік подій називається *потоком без наслідків*, якщо ймовірність влучення того чи іншого числа подій на якусь ділянку часу не залежить від того, скільки подій влучило на будь-яку іншу ділянку, що не перетинається з нею.

Потік подій називається *ординарним*, якщо ймовірність появи на елементарній ділянці двох або більше подій зневажливо мала у порівнянні з імовірністю появи однієї події.

Ординарний потік подій без наслідків називається *пуассонівським*. Якщо події утворюють пуассонівський потік, то кількість X подій, що влучають на будь-яку ділянку часу $(t,\ t+\tau)$, розподілена за законом Пуассона:

$$\mathrm{P}_m = \frac{a^m}{m!}\exp(-a), \tag{2.64}$$

де a – математичне сподівання числа точок, що влучають на ділянку:

$$a \;=\; \int\limits_{t}^{t+\tau} \lambda(t)\,dt. \tag{2.65}$$

Якщо $\lambda(t) \;=\; \lambda \;=\; const$, пуассонівський потік називається стаціонарним пуассонівським або *найпростішим*. Для найпростішого потоку кількість подій, що влучають на будь-яку ділянку часу тривалістю τ, розподілена за законом Пуассона з параметром $a = \lambda\tau$.

Випадковим полем точок називається сукупність точок, випадковим чином розкиданих на площині (або у просторі).

Інтенсивністю (або густиною) поля λ називається середня кількість точок, що влучають в одиницю площі (об'єму).

Поле точок називається *пуассонівським*, якщо воно має такі властивості:

1) ймовірність влучення тієї чи іншої кількості точок у будь-яку область площини (простору) не залежить від того, скільки їх влучило у будь-яку область, що не перетинається з даною;

2) ймовірність влучення в елементарну область $\Delta x\,\Delta y$ двох або більше точок зневажливо мала у порівнянні з імовірністю влучення однієї точки (якість ординарності).

Кількість X точок пуассонівського поля, що влучають у будь-яку область S площини (простору), розподілена за законом Пуассона:

$$\mathrm{P}_m \;=\; \Pr\{\mathrm{X} = m\} \;=\; \frac{a^m}{m\,!}\,\exp(-a), \qquad\qquad m = 0, 1, 2, \ldots, \tag{2.66}$$

де a – математичне сподівання кількості точок, що влучають в область S.

Якщо інтенсивність поля $\lambda(x, y) = \lambda = const$, то воно називається *однорідним* (властивість, аналогічна стаціонарності потоку подій). При однорідному полі з інтенсивністю λ маємо $a = L\lambda$, де L – розмір ділянки, або $a = S\lambda$, де S – площа області, або $a = V\lambda$, де V – об'єм області.

Якщо поле неоднорідне, то

$$a \;=\; \int\limits_{L} \lambda(x, y)\,dx \qquad\qquad -\ \text{для ділянки;} \tag{2.67a}$$

$$a \;=\; \iint\limits_{(S)} \lambda(x, y)\,dx\,dy \qquad\qquad -\ \text{для площини;} \tag{2.67b}$$

$$a \;=\; \iiint\limits_{(V)} \lambda(x, y, z)\,dx\,dy\,dz \qquad\qquad -\ \text{для об'єму.} \tag{2.67c}$$

У багатьох задачах математичної статистики (при оцінюванні або перевірці статистичних гіпотез) використовуються також і інші закони розподілу.

2.10. Приклади

Приклад 2.1

Знайти математичні сподівання, дисперсію і коефіцієнт кореляції двовимірної нормальної випадкової величини.

Розв'язання

Запишемо вираз для густини розподілу $f_{\mathrm{XY}}(x,y)$ нормальної системи з двох випадкових величин (X,Y)

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-r_{\mathrm{XY}}^2}}\exp\left\{-\;\frac{1}{2(1-r_{\mathrm{XY}}^2)}\,Q(x,y)\right\}, \qquad (*)$$

$$Q(x,y) \;=\; \frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} \;-\; 2r_{\mathrm{XY}}\frac{(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} + \frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2},$$

де m_{X}, m_{Y}, σ_{X}, σ_{Y}, r_{XY} — параметри закону.

1) Очевидно, що

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{X}] \;=\; \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x\,f_{\mathrm{XY}}(x,y)\,dx\,dy \;=\; \int\limits_{-\infty}^{\infty} x\,f_{\mathrm{X}}(x)\,dx \;=\; \mathrm{M}_{\mathrm{X}}[\mathrm{X}],$$

де $f_{\mathrm{X}}(x)$ – парціальна густина X-компоненти:

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{X}}}\exp\left\{-\,\frac{1}{2}\,\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}\right\},$$

що випливає з виразу $(*)$ інтегруванням по змінній y.

Інтегруючи, отримаємо для перших двох моментів X-компоненти

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{X}] \;=\; m_{\mathrm{X}},$$

$$\mathrm{D}_{\mathrm{XY}}[\mathrm{X}] \;=\; \mathrm{D}_{\mathrm{X}}[\mathrm{X}] \;=\; \sigma_{\mathrm{X}}^2.$$

2) Аналогічно діючи для Y-компоненти, знайдемо

$$\mathrm{M}_{\mathrm{XY}}[\mathrm{Y}] \;=\; m_{\mathrm{Y}},$$

$$\mathrm{D}_{\mathrm{XY}}[\mathrm{Y}] \;=\; \mathrm{D}_{\mathrm{Y}}[\mathrm{Y}] \;=\; \sigma_{\mathrm{Y}}^2.$$

3) Запишемо вираз для змішаного другого моменту

$$k_{\mathrm{XY}} \;=\; \mathrm{M}_{\mathrm{XY}}[(\mathrm{X}-m_{\mathrm{X}})(\mathrm{Y}-m_{\mathrm{Y}})] \;=\; \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})\,f_{\mathrm{XY}}(x,y)\,dx\,dy.$$

У подвійному інтегралі, що виник, перейдемо до нових змінних інтегрування (u,v) за правилом: $u=(x-m_{\mathrm{X}})/\sigma_{\mathrm{X}}$, $v=(y-m_{\mathrm{Y}})/\sigma_{\mathrm{Y}}$, що дає

$$k_{\mathrm{XY}} \;=\; \frac{1}{2\pi\,\sqrt{1-r_{\mathrm{XY}}^2}}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} uv\,\exp\left\{-\,\frac{u^2-2r_{\mathrm{XY}}uv+v^2}{2(1-r_{\mathrm{XY}}^2)}\right\}du\,dv.$$

Тепер зручно замість змінної u ввести ще одну змінну інтегрування за правилом $t = u - r_{\text{XY}}v$:

$$k_{\text{XY}} = \frac{1}{2\pi\sqrt{1-r_{\text{XY}}^2}} \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} (t + r_{\text{XY}}v)v \exp\left\{-\frac{t^2 + (1-r_{\text{XY}}^2)v^2}{2(1-r_{\text{XY}}^2)}\right\} dtdv.$$

У цьому інтегралі внесок від першого доданку в круглих дужках під інтегралом тотожно дорівнює нулю з-за непарності підінтегрального виразу і парності меж інтегрування відносно змінної t.

Подвійний інтеграл, що залишився, стає добутком двох однократних інтегралів:

$$k_{\text{XY}} = \frac{1}{\sqrt{2\pi(1-r_{\text{XY}}^2)}} \int\limits_{-\infty}^{\infty} \exp\left\{-\frac{t^2}{2(1-r_{\text{XY}}^2)}\right\} dt \, \frac{r_{\text{XY}}}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} v^2 \exp\left\{-\frac{v^2}{2}\right\} dv.$$

Перший з цих інтегралів-співмножників тотожньо дорівнює 1. Другий інтеграл (без співмножника r_{XY}) збігається зі значенням дисперсії змінної v, яка також дорівнює 1. Отже,

$$k_{\text{XY}} = r_{\text{XY}}.$$

Таким чином, закон двовимірного розподілу (*) повністю визначається завданням його характеристик m_{X}, m_{Y}, σ_{X}, σ_{Y} і r_{XY}.

Приклад 2.2

Нехай на виробництві витрата матеріалу носить випадковий характер із середньою інтенсивністю 20 одиниць на день. Для покриття витрати виконується щомісячне постачання обсягом в 640 одиниць.

Визначити:

а) ймовірність утворення дефіциту (недостачі) матеріалів;

б) обсяг постачання, при якому ймовірність дефіциту не перевищить 0,01.

Розв'язання

Позначимо через T проміжок часу, в терміні якого сумарна витрата буде дорівнювати обсягу постачання β. Величина T є випадковою, що підкоряється гамма–розподілу з параметрами β і λ.

Дефіцит утвориться, якщо T виявиться менше заданого інтервалу між постачанням, тобто при $\text{T} \le 30$.

З визначення функції розподілу маємо $\Pr(\text{T} \le 30) = F_{\text{T}}(30;\, \beta,\, \lambda)$.

а) Задаючись в (2.34) $\beta = 640$, $\lambda = 20$ і $t = 30$, отримаємо

$$\Pr(\text{T} \le 30) = F_{\text{T}}(30;\, 640,\, 20) = \frac{20^{640}}{\Gamma(640)} \int\limits_0^{30} t^{639}\text{e}^{-20t}\, dt = 0,0569.$$

Отже, дефіцит буде мати місце з імовірністю 0,0569.

б) За заданою величиною $p = \Pr(\text{T} \le 30) = 0,01 = F_{\text{T}}(30;\, \beta,\, 20)$ знаходимо $\beta = 660$ (одиниць). При такому обсязі постачання ймовірність дефіциту не перевищить 0,01. Відповідно, з імовірністю 0,99 за розглянутий тимчасовий інтервал дефіциту не буде.

Приклад 2.3

Дискретна випадкова величина X підкоряється біноміальному розподілу Бернуллі з параметрами n та p.

Знайти характеристичну функцію $q(\lambda)$ випадкової величини X.

Розв'язання

Скористаємося визначенням характеристичної функції

$$q(\lambda) \;=\; \mathrm{M}[\,\exp(i\lambda \mathrm{X})].$$

Для випадкової величини, що підпорядковується закону Бернуллі, маємо

$$\mathrm{P}_m \;=\; \Pr(\mathrm{X}=m) \;=\; C_n^m p^m (1-p)^{n-m}, \qquad m=0,1,\ldots,n.$$

Тоді отримаємо

$$q(\lambda) \;=\; \sum_{m=0}^{n} \exp(i\lambda m)\,\mathrm{P}_m \;=\; \left(p\mathrm{e}^{i\lambda}+1-p\right)^n.$$

Приклад 2.4

Дискретна випадкова величина X підкоряється розподілу Пуассона з параметром a.

Знайти характеристичну функцію $q(\lambda)$ випадкової величини X.

Розв'язання

Для випадкової величини, що підпорядковується закону Пуассона, маємо

$$\mathrm{P}_m \;=\; \Pr(\mathrm{X}=m) \;=\; \frac{a^m}{m!}\exp(-a), \qquad m=0,1,\ldots.$$

З визначення характеристичної функції $q(\lambda)=\mathrm{M}[\,\exp(i\lambda \mathrm{X})]$ випливає у випадку розподілу Пуассона

$$q(\lambda) \;=\; \exp\left[a\left(\mathrm{e}^{i\lambda}-1\right)\right].$$

Приклад 2.5 (Розподіл χ^2 з n ступенями вільності)

Заданий набір незалежних нормальних величин $\{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n\}$ з параметрами $\{m_1=0,\, m_2=0,\, \ldots,\, m_n=0\}$ й $\{\sigma_1=1,\, \sigma_2=1,\, \ldots,\, \sigma_n=1\}$. Розглянемо випадкову величину

$$\mathrm{Y} \;=\; \sum_{i=1}^{n} \mathrm{X}_i^2.$$

Отримати закон розподілу випадкової величини Y.

Розв'язання

Нехай спочатку $n=1$. У цьому випадку характеристична функція $q_{\mathrm{Y}_1}(\lambda)$ така:

$$q_{\mathrm{Y}_1}(\lambda) \;=\; \mathrm{M}[\,\exp(i\lambda \mathrm{Y}_1)] \;=\; \mathrm{M}[\,\exp(i\lambda \mathrm{X}_1^2)] \;=\; \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-x_1^2/2+i\lambda x_1^2\right)\,dx_1.$$

Інтегруючи, знайдемо

$$q_{\mathrm{Y}_1}(\lambda) \;=\; (1-2i\lambda)^{-1/2}.$$

Оскільки за умовою компоненти випадкової величини (ВВ) Y незалежні, то у випадку, коли вона утворена з n доданків, отримаємо

$$q_{\mathrm{Y}}(\lambda) \;=\; \prod_{i=1}^{n} q_{\mathrm{Y}_i}(\lambda) \;=\; \left[q_{\mathrm{Y}_1}(\lambda)\right]^n \;=\; (1-2i\lambda)^{-n/2}.$$

Густину розподілу ВВ Y можна визначити за допомогою зворотного перетворення Фур'є, що дає

$$f_{\mathrm{Y}}(t) \;=\; Ct^{\frac{n}{2}-1}\mathrm{e}^{-t/2}, \qquad\qquad t \geq 0,$$

де C – стала.

Цю величину визначимо з умови $\int\limits_0^\infty f_{\mathrm{Y}}(t)\,dt = 1$, звідси $C^{-1} = 2^{n/2}\,\Gamma(n/2)$.

Отже,

$$f_{\mathrm{Y}}(t) \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)}\,t^{n/2-1}\mathrm{e}^{-t/2}.$$

Знайдена густина відповідає випадковій величині, розподіленій згідно із законом χ^2 з n ступенями вільності.

Приклад 2.6 (Розподіл Стьюдента)

Заданий набір з n незалежних нормальних величин $\{\mathrm{X}_1,\ \mathrm{X}_2,\ \ldots,\ \mathrm{X}_n\}$ з параметрами $\{m_1 = 0,\ m_2 = 0,\ \ldots,\ m_n = 0\}$ та $\{\sigma_1 = 1,\ \sigma_2 = 1,\ \ldots,\sigma_n = 1\}$. Крім того, є ще одна незалежна нормальна величина X_0 з параметрами $m_0 = 0$ та $\sigma_0 = 1$.

Розглянемо випадкову величину

$$\mathrm{T} \;=\; \mathrm{X}_0 \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{X}_i^2\right)^{-1/2}.$$

Отримати закон розподілу випадкової величини T.

Розв'язання

Випадкова величина T, що розглядається, є стандартизованим відношенням Стьюдента

$$\mathrm{T} \;=\; \frac{\mathrm{X}_0}{\sqrt{\chi^2/n}}.$$

Для знаходження густини розподілу $f_{\mathrm{T}}(t)$ скористаємося технікою δ-функції Дірака

$$f_{\mathrm{T}}(t) \;=\; \mathrm{M}\left[\delta\left(\frac{\mathrm{X}_0}{\sqrt{\chi^2/n}} - t\right)\right].$$

Тут математичне сподівання шукають відносно випадкових величин χ^2 та X_0 з густинами розподілів

$$f_{\mathrm{X}_0}(x_0)\,dx_0 \;=\; \frac{1}{2\pi}\,\exp\left(-x_0^2/2\right)\,dx_0,$$

$$f_{\chi^2}(\chi^2)\,d\chi^2 \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)}\,(\chi^2)^{\frac{n}{2}-1}\mathrm{e}^{-\chi^2/2}\chi^2\,d\chi^2.$$

Підставимо ці густини у вираз для $f_{\mathrm{T}}(t)$ і врахуємо фільтруючу властивість δ-функції, тоді отримаємо

$$f_{\mathrm{T}}(t) \;=\; \frac{1}{2^{n/2}\,\Gamma(n/2)}\,\frac{1}{n}\int\limits_0^\infty (\chi^2)^{n-1}\exp\left(-\frac{1}{2}\,\frac{t^2\chi^2}{n}-\frac{\chi^2}{2}\right)d\chi^2.$$

Використовуючи тепер допоміжну змінну інтегрування u за правилом

$$u \;=\; \frac{1}{2}\,\frac{t^2\chi^2}{n}+\frac{\chi^2}{2},$$

отримаємо

$$f_{\mathrm{T}}(t) \;=\; \left[2^{n/2}\,\sqrt{n}\,\left(\frac{1+t^2/n}{2}\right)\right]^{-1}\int\limits_0^\infty u^{(n-1)/2}\mathrm{e}^{-u}\,du.$$

Користуючись визначенням Γ–функції, остаточно знайдемо

$$f_{\mathrm{T}}(t) \;=\; \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\,\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}.$$

Знайдена густина відповідає випадковій величині, розподіленій згідно з законом Стьюдента.

Приклад 2.7

В задачах оцінювання характеристик за обмеженою кількістю спостережень n, ідентифікації процесів та інших використовується статистика

$$\mathrm{R} \;=\; \mathrm{X}^{(n)}_{\max}-\mathrm{X}^{(n)}_{\min},$$

яка має назву *розмах* або *широта вибірки* (тут $\mathrm{X}^{(n)}_{\max}$ та $\mathrm{X}^{(n)}_{\min}$ – крайні члени варіаційного ряду). У загальному вигляді розподіл розмаху визначається таким чином:

$$\Pr\left((X^{(n)}_{\max}-X^{(n)}_{\min})<t\right) \;=\; n\int\limits_{-\infty}^{\infty}[F_{\mathrm{X}}(x+t)-F_{\mathrm{X}}(x)]^{n-1}\,dF_{\mathrm{X}}(x).$$

Використовуючи цей результат, наведемо (без доказу) декілька тверджень, які виявляються корисними в теорії оцінювання.

Твердження 1. Нехай з генеральної сукупності з рівномірною густиною

$$f_{\mathrm{X}}(x) \;=\; \frac{1}{b-a}, \qquad\qquad x\in(a,\,b),$$

витягнута вибірка обсягом n. Тоді густина розподілу розмаху R цієї вибірки має вигляд

$$f_{\mathrm{R}}(t) \;=\; n(n-1)\,\frac{t^{n-2}}{(b-a)^{n-1}}.$$

Твердження 2. Нехай випадкова величина Y є центрованою та рівномірно розподіленою, тобто $\mathrm{Y} \in [-a,\, a]$. Тоді густина розподілу її розмаху R має вигляд

$$f_{\mathrm{R}}(t) \;=\; \frac{1}{2a}.$$

Твердження 3. Нехай випадкова величина X має нормальний розподіл з нульовим математичним сподіванням та дисперсією σ_{X}^2, а друга випадкова величина Y є центрованою та рівномірно розподіленою випадковою величиною, тобто $\mathrm{Y} \in [-a,\, a]$. Введемо в розгляд статистику $\mathrm{T} = \mathrm{X}/\mathrm{Y}$. Тоді густина розподілу цієї статистики T така:

$$f_{\mathrm{T}}(t) \;=\; \frac{4\sigma_{\mathrm{X}}^2}{a\sqrt{\pi}}\, t^{-2} \left[1 - \exp\left(-\frac{a^2 t^2}{4\sigma_{\mathrm{X}}^2}\right)\right].$$

Твердження 4. Нехай є вибірка обсягом n з нормальної генеральної сукупності з параметрами m_{X} та σ_{X}. Тоді закон розподілу розмаху R цієї вибірки визначається таким чином:

$$F_{\mathrm{R}}(t) \;=\; \Pr\left((X_{\max}^{(n)} - X_{\min}^{(n)}) < t\right) \;=\; n \int\limits_{-\infty}^{\infty} [F(x+t) - F(x)]^{n-1}\, dF(x),$$

де $F(x)$ – інтегральний закон розподілу нормальної величини.

Диференціюючи під знаком інтеграла, для густини розподілу розмаху цієї вибірки отримаємо

$$f_{\mathrm{R}}(t) \;=\; \frac{n(n-1)}{2\pi\sigma_{\mathrm{X}}^2} \int\limits_{-\infty}^{\infty} [F(x+t) - F(x)]^{n-2} \exp\left(-\frac{(x - m_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2}\right) dx.$$

Приклад 2.8 (Розподіл Фішера–Снедекора)

Нехай задані дві випадкові величини S і T, що підкоряються закону χ^2 відповідно з n_1 і n_2 ступенями вільності та мають один і той же параметр σ. Отримати закон розподілу випадкової величини $\mathrm{U} = \mathrm{S}/\mathrm{T}$.

Розв'язання

Закон розподілу початкових величин наступний (тут C_1 і C_2 – нормувальні константи):

$$f_{\mathrm{S}}(s)\, ds \;=\; C_1 \left(\frac{s}{\sigma}\right)^{n_1 - 1} \exp\left(-\frac{s^2}{2\sigma^2}\right) \frac{ds}{\sigma}, \qquad s \geq 0,$$

$$f_{\mathrm{T}}(t)\, dt \;=\; C_2 \left(\frac{t}{\sigma}\right)^{n_2 - 1} \exp\left(-\frac{t^2}{2\sigma^2}\right) \frac{dt}{\sigma}, \qquad t \geq 0.$$

Щоб визначити форму закону величини U, розглянемо розподіл пари S і T. Використаємо випадкові величини R, U і зробимо заміну змінних $s^2 + t^2 \;=\; r^2$ та $t/s \;=\; u,\;\; u, r \geq 0,$ що дає $s \;=\; r/\sqrt{1+u^2}$ та $t \;=\; ru/\sqrt{1+u^2}$.

Звідси

$$\left|\frac{D(s,t)}{D(r,u)}\right| \;=\; \frac{r}{\sqrt{1+u^2}}.$$

Тоді елемент диференціальної ймовірності $f_{\mathrm{R,U}}(r,u)\,dr\,du$ пари величин (R,U) запишеться у вигляді (C_3 – нормувальна константа)

$$f_{\mathrm{R,U}}(r,u)\,dr\,du \;=\; C_3\,\exp\left(-\frac{r^2}{2\sigma^2}\right)\left(\frac{r}{\sigma}\right)^{n_1+n_2-1}\,\frac{dr}{\sigma}\,\frac{u^{n_2-1}\,du}{(1+u^2)^{(n_1+n_2)/2}}.$$

Звідси випливає, що величини R і U – незалежні, при цьому величина U підкоряється закону χ^2 з n_1+n_2 кількостями ступенів вільності.

Проінтегрувавши останній вираз за r, отримаємо елемент диференціальної ймовірності ВВ U у формі

$$f_{\mathrm{U}}(u)\,du \;=\; C_4\,\frac{u^{n_2-1}}{(1+u^2)^{(n_1+n_2)/2}}\,du,$$

де C_4 – нормувальна константа. Цю сталу знайдемо з умови нормування $\int\limits_0^\infty f_{\mathrm{U}}(u)\,du = 1$.

Отже, густина розподілу випадкової величини U така:

$$f_{\mathrm{U}}(u) \;=\; \frac{2\,\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\,\Gamma\left(\frac{n_2}{2}\right)}\,\frac{u^{n_2-1}}{(1+u^2)^{(n_1+n_2)/2}}.$$

Знайдена густина відповідає випадковій величині, розподіленій згідно із законом Фішера–Снедекора.

2.11. Задачі для розв'язання

Задача 2.1

Партія виробів вважається придатною до випуску, якщо брак у ній не перевищує 3 %. З партії у 2000 виробів було відібрано і перевірено 400. При цьому бракованих виробів виявилося 6.

Яка імовірність того, що вся партія задовольняє технічні умови і може бути прийнята?

Задача 2.2

Система випадкових величин (X,Y) подпорядкована нормальному закону розподілу

$$f_{\mathrm{XY}}(x,y) \;=\; \frac{1}{2\pi\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma_{\mathrm{X}}^2}-2\rho\,\frac{xy}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}+\frac{y^2}{\sigma_{\mathrm{Y}}^2}\right]\right).$$

Визначити закон розподілу випадкової величини $\mathrm{Z}=\mathrm{X}/\mathrm{Y}$. Розглянути випадок, коли $\rho=0$.

Задача 2.3

У деякому районі в приватному володінні громадян знаходиться 3500 корів. У випадковій вибірці обстежували 800 корів і встановили, що у цієї групи середній річний удій дорівнює 2800 *кг*, а середнє квадратичне відхилення $\sigma = 250$ *кг*.

Визначити ймовірність, яка може гарантувати, що середньорічний удій всіх корів відрізняється від 2800 *кг* за абсолютною величиною менше ніж 10 *кг*.

Задача 2.4

Скільки осіб у віці від 20 до 25 років треба опитати вибірково, щоб встановити серед них відсоток студентів з точністю до 0,5 %, що гарантується з імовірністю 0,999?

Задача 2.5

Знайти математичне сподівання та дисперсію випадкової величини $\mathrm{Z} = |\mathrm{X}|$, якщо ВВ X підпорядкована нормальному закону з параметрами m, σ^2.

Відповідь:

$$\mathrm{M[Z]} = \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{m^2}{2\sigma^2}\right) + m\Phi\left(\frac{m}{\sigma}\right), \quad \mathrm{D[Z]} = \sigma^2 + m^2 - (\mathrm{M[Z]})^2.$$

Задача 2.6

Випадкова крапка (X, Y) розподілена рівномірно в колі радиусом 1 з центром в точці $(0; 0)$.

Знайти закони розподілу випадкових величин U=X/Y та V=Y/X.

Відповідь: $f_{\mathrm{U}}(u) = [\pi(1+u^2)]^{-1}$, $f_{\mathrm{V}}(v) = [\pi(1+v^2)]^{-1}$.

Задача 2.7

Неперервна випадкова величина X має закон розподілу $F_{\mathrm{X}}(x)$, а пов'язана з нею випадкова величина Y визначається законом розподілу

$$F_{\mathrm{Y}}(y) \;=\; \Pr(\mathrm{Y} < y) \;=\; n \int\limits_{-\infty}^{\infty} \left[F_{\mathrm{X}}(x+y) - F_{\mathrm{X}}(x)\right]^{n-1} f_{\mathrm{X}}(x)\, dx,$$

де $n = 2, 3, \ldots$ – параметр.

Для різних випадкових величин X знайти математичне сподівання та дисперсію СВ Y.

Задача 2.8

Система випадкових величин (X, Y) подпорядкована нормальному закону розподілу

$$f_{\mathrm{XY}}(x, y) \;=\; \frac{1}{2\pi\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \exp\left(-\frac{x^2}{2\sigma_{\mathrm{X}}^2} - \frac{y^2}{2\sigma_{\mathrm{Y}}^2}\right).$$

Знайти закон розподілу випадкових величин $\mathrm{Z} = \mathrm{X}–\mathrm{Y}$, $\mathrm{Z}_1 = \mathrm{X}+\mathrm{Y}$ та $\mathrm{Z}_2 = \mathrm{X}-\mathrm{Y}$.

Відповідь: Випадкові величини Z_1 та Z_2 мають однаковий закон розподілу з густиною

$$f_{\mathrm{Z}}(z) \;=\; \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{Z}}} \exp\left(-\frac{z^2}{2\sigma_{\mathrm{Z}}^2}\right), \qquad \sigma_{\mathrm{Z}}^2 = \sigma_{\mathrm{X}}^2 + \sigma_{\mathrm{Y}}^2.$$

Задача 2.9

Отримані дві вибірки:

0,79 0,75 0,86 0,05 1,29 0,42 1,12 0,70 1,54 1,43

та

2,05 1,38 1,45 0,35 0,64 0,58 1,03 0,12 1,30 1,09 0,33

з двох нормальних розподілів з середніми a й b та однаковою дисперсією σ^2, всі параметри невідомі.

При рівні довіри $p = 0{,}95$ побудувати довірчий інтервал для різності середніх $\gamma = a - b$.

2.12. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

Побудуйте програму візуалізації густини розподілу та інтегрального закону розподілу для указаних випадків. Результати оформіть графічно.

Варіант 1

Рівномірний закон.

Вхідні дані для програми:

a, b – ліва та права межі можливих значень;

b – права межа можливих значень;

$f_{\text{X}}(x)$ – густина розподілу ймовірностей;

$F_{\text{X}}(x)$ – інтегральний закон розподілу ймовірностей.

Результат роботи програми – масив, який містить значення густини розподілу ймовірностей $f_{\text{X}}(x)$ та інтегрального закону розподілу ймовірностей $F_{\text{X}}(x)$ в обраних вузлах аргумента x випадкової величини X.

Варіант 2

Нормальний закон Гаусса.

Вхідні дані для програми:

m_{X} – математичне сподівання;

σ^2_{X} – дисперсія;

$f_{\text{X}}(x)$ – густина розподілу ймовірностей;

$F_{\text{X}}(x)$ – інтегральний закон розподілу ймовірностей.

Результат роботи програми — массив, який містить значення густини розподілу ймовірностей $f_{\text{X}}(x)$ та інтегрального закону розподілу ймовірностей $F_{\text{X}}(x)$ в обраних вузлах аргумента x випадкової величини X.

Варіант 3

Розподіл χ^2.

Вхідні дані для програми:

n – кількість ступенів вільності;

$f(\chi^2)$ – густина розподілу ймовірностей;

$F(\chi^2)$ – інтегральний закон розподілу ймовірностей;

Результат роботи програми – массив, який містить значення густини розподілу ймовірностей $f(\chi^2)$ та інтегрального закону розподілу ймовірностей $F(\chi^2)$ в обраних вузлах аргумента x випадкової величини χ^2_n.

Завдаанння 2

Побудуйте програму візуалізації последовательности амплитуд розподілу ймовірностей P_m для указаних випадків.

Результат роботи програми – масив, який містить значення шуканої функції. Результати оформіть графічно.

Варіант 1

Рівномірний дискретний закон.

Вхідні дані для програми:

n_1, n_2 – ліва та права межі можливих значень;

P_m – послідовність амплітуд розподілу ймовірностей.

Результат роботи програми – масив, який містить значення послідовності амплітуд рівномірного розподілу ймовірностей P_m в вузлах дискретного аргументу $n_1 \le m \le n_2$ випадкової величини X.

Варіант 2

Розподіл Бернуллі.

Вхідні дані для програми:

n – загальна кількість іспитів;

p – ймовірність появи сприятливої події за один іспит.

P_m – послідовність амплітуд розподілу ймовірностей.

Результат роботи програми – масив, який містить значення послідовності амплітуд розподілу ймовірностей Бернуллі P_m в вузлах дискретного аргументу $0 \le m \le n$ випадкової величини X.

Варіант 3

Розподіл Пуассона.

Вхідні дані для програми:

a – математичне сподівання.

P_m – послідовність амплітуд розподілу ймовірностей.

Результат роботи програми – масив, який містить значення послідовності амплітуд розподілу ймовірностей Пуассона P_m в вузлах дискретного аргументу $m \ge 0$ випадкової величини X.

2.13. Завдання для перевірки

1. Назвіть основні характеристики одновимірного нормального закону, двовимірного нормального закону, багатовимірного нормального закону.

2. Яка функція називається гамма-функцією?

3. Назвіть основні властивості гамма-функції.

4. Згідно з яким законом розподілена сума квадратів випадкових величин, кожна з яких має стандартизований нормальний розподіл?

5. Що означає кількість ступенів вільності випадкової величини χ^2?

6. Яка випадкова величина називається безрозмірним дробом Стьюдента?

7. Який закон розподілу має відношення суми квадратів нормальних стандартизованих випадкових величин?

8. Яке значення має кількість ступенів вільності t-розподілу? Розкрийте значення кількості ступенів вільності випадкової величини, що має розподіл Фішера.

9. Розкрийте значення квантиля χ^2-розподілу, квантиля t-розподілу, квантиля F-розподілу.

10. Який розподіл є граничним (при $n \to \infty$) для χ^2-розподілу? Для t-розподілу? Для F-розподілу?

11. Яке значення має параметр кореляції двовимірного нормального закону?

12. Який вигляд має характеристична функція одновимірного нормального закону? Двовимірного нормального закону? Багатовимірного нормального закону?

13. Наведіть вирази для характеристичної функції гамма-розподілу, χ^2-розподілу, t-розподілу Стьюдента.

3. Статистична теорія оцінювання параметрів розподілу

3.1. Постановка задачі оцінювання

Припустимо, що для оцінки закону розподілу випадкової величини X, що досліджується, з генеральної сукупності з невідомою функцією розподілу $F_{\text{x}}(x)$ вилучена вибірка $x_1, x_2, \ldots, x_n$. Припустимо також, що експериментатор візуально на вигляд гістограми або полігона частостей, або на основі будь-яких інших міркувань вибрав клас Ω функцій визначеного вигляду (нормальних, показникових, біноміальних і т.д.), до якого, на його думку, може належати функція розподілу випадкової величини X, що досліджується.

Найчастіше при дослідженнях неперервних випадкових величин експериментатор намагається вибрати клас нормальних функцій, тобто побудувати нормальну модель сукупності, тому що ця модель найбільш опрацьована в аналітичному відношенні. Крім того, нормальний закон є стійким при композиції і є граничним законом багатьох законів розподілу. Тому клас нормальних функцій часто можна приймати за наближену модель генеральної сукупності.

Після того, як клас функцій Ω вибраний, проводиться *оцінка* (підгонка) параметрів вибраного класу функцій. Наприклад, якщо вибраний нормальний клас функцій для опису випадкової величини X, що досліджується, то за вибіркою $x_1, x_2, \ldots, x_n$ потрібно оцінити два параметри – математичне сподівання m_{x} і середнє квадратичне відхилення σ_{x}, від яких залежить нормальний розподіл.

Якщо вибраний клас функцій Пуассона, то на основі вибірки $x_1, x_2, \ldots, x_n$ потрібно оцінити тільки один параметр, яким визначається закон Пуассона – інтенсивність λ.

Нехай з генеральної сукупності з функцією розподілу $F_{\text{x}}(x;\theta)$, де θ – невідомий параметр, зроблена вибірка обсягом n і отримані результати x_1, x_2, ..., x_n. Взагалі, за результатами вибірки, якого б великого обсягу вона не була, не можна визначити точне значення невідомого параметра θ, а можна лише знайти його наближене значення $\hat{\theta}$, яке і називається оцінкою.

Для знаходження наближених значень (оцінки) невідомого параметра θ будемо розглядати функції вигляду

$$\hat{\theta} \;=\; u\left(x_1, x_2, \ldots, x_n\right), \tag{3.1}$$

які називаються *вибірковими функціями* або *статистиками*. Задача оцінки невідомого параметра θ зводиться до знаходження таких вибіркових функцій $\hat{\theta} = u\left(x_1, x_2, \ldots, x_n\right)$, які можна використати як оцінку невідомого параметра θ.

Будь-яка вибірка є обмеженою та випадковою. Отже, всі вибіркові функції $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ також є випадковими. Наприклад, оцінкою математичного сподівання є середнє арифметичне

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

Якщо повторити дослід k разів, вибираючи з генеральної сукупності випадкові вибірки одного і того ж обсягу n, то за їх даними знайдемо ряд значень $\{\overline{x}_k\}$, які, взагалі, будуть відрізнятися одне від іншого.

Таким чином, будемо розглядати оцінку $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ невідомого параметра θ як випадкову величину, а її значення, обчислені на основі даної вибірки обсягом n, – як одну реалізацію випадкової величини, тобто як одну з безлічі можливих значень цієї випадкової величини.

Можливі оцінки параметрів розподілу поділяються на точкові та інтервальні. *Точкова оцінка* параметра θ визначається одним числом $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$. *Інтервальною оцінкою* параметра θ називають оцінку, яка визначається двома числами $\hat{\theta}_1$ і $\hat{\theta}_2$ – кінцями інтервалу, який накриває параметр, що оцінюється.

3.2. Непараметричне і параметричне оцінювання. Статистичні оцінки та їх властивості

З наведеного вище випливає, що оцінкою невідомого параметра θ є функція $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, що залежить від спостережених значень випадкової величини, які можуть бути використані для знаходження наближеного значення невідомого параметра θ. Таким чином, далі будемо розглядати тільки певні класи функцій, близьких в певному значенні до параметра θ, що оцінюється. В математичній статистиці є спеціальний розділ – *теорія оцінювання*, в якому займаються побудовою правил конструювання функції $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ для знаходження точкових оцінок невідомих параметрів.

Перейдемо до основних властивостей, які повинні мати "добрі" оцінки невідомого параметра $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$.

Передусім, з точки зору точності та надійності оцінок бажано, щоб знайдені на основі вибіркових функцій $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, що використовуються, оцінки невідомих параметрів по можливості були тісно сконцентровані біля значень параметрів, що оцінюються, іншими словами, щоб розсіювання випадкової величини $\hat{\theta}$ біля θ було по можливості найменшим.

Визначення 1. Оцінка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ називається *спроможною*, якщо при збільшенні числа вимірювань оцінка сходиться по ймовірності до параметра, що оцінюється, тобто якщо

$$\lim_{n\to\infty} \Pr\left(|\theta - \hat{\theta}| < \varepsilon\right) = 1. \tag{3.2}$$

Вимога спроможності допомагає уникнути серйозних помилок ε у визначенні $\hat{\theta}$ при досить великих n.

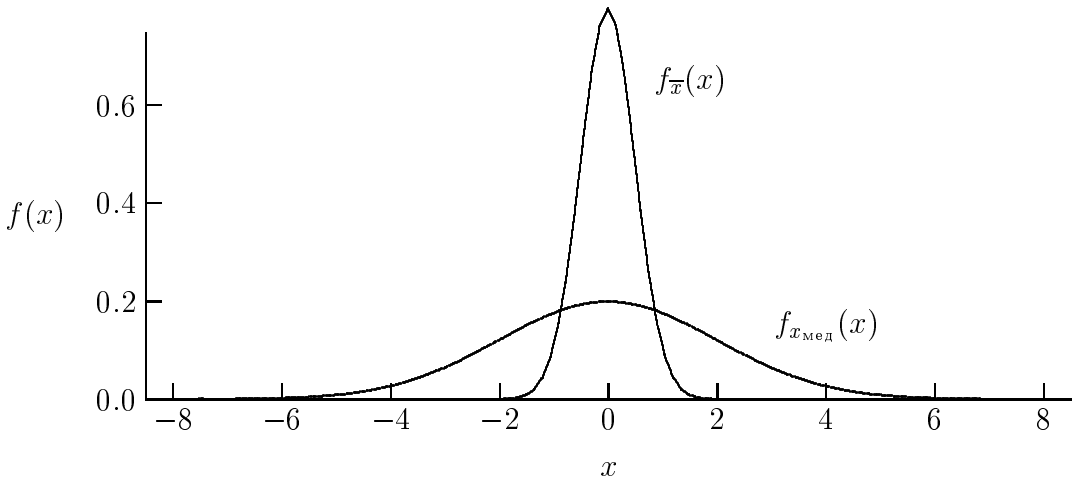


Рисунок 3.1 — Густини розподілу $f_{\overline{x}}(x)$ і $f_{x_{\text{мед}}}(x)$

Визначення 2. Оцінка $\hat{\theta} = u\,(x_1, x_2, \ldots, x_n)$ називається *незсуненою* (оцінкою без систематичної помилки), якщо її математичне сподівання дорівнює параметру, що оцінюється, тобто якщо

$$\mathrm{M}[\hat{\theta}] \;=\; \theta. \tag{3.3}$$

Якщо умова (3.3) не виконується, то оцінка $\hat{\theta}$ називається *зсуненою* (що містить систематичну помилку). Часто поруч з незсуненими оцінками $\hat{\theta} = u\,(x_1, x_2, \ldots, x_n)$ застосовуються *асимптотично незсунені оцінки*, тобто такі оцінки, для яких виконується $\mathrm{M}[\hat{\theta}] \to \theta$ при збільшенні обсягу вибірки.

Спроможні, незсунені або асимптотично незсунені оцінки можуть бути отримані різними методами. Наприклад, дві оцінки математичного сподівання – середнє арифметичне $\overline{s}$ і вибіркова медіана $x_{\text{мед}}$ – є незсуненими і спроможними оцінками.

Приклади розподілів цих оцінок зображено на рисунку 3.1.

З цих двох оцінок доцільніше вибрати $\overline{s}$, оскільки дисперсія цієї оцінки менша, ніж дисперсія вибіркової медіани. У строгих курсах математичної статистики доводиться, що дисперсія будь-якої незсуненої оцінки одного параметра θ задовольняє *нерівності Крамера-Рао*

$$\mathrm{D}[\hat{\theta}] \;\geq\; \frac{1}{I_n(\theta)}, \tag{3.4}$$

де $I_n(\theta)$ – *інформація Фішера*, що міститься у вибірці обсягом n відносно невідомого параметра θ. Для неперервної випадкової величини X із густиною розподілу $f_{\mathrm{X}}(x;\theta)$ справедлива нерівність

$$\mathrm{D}[\hat{\theta}] \;\geq\; \frac{1}{n\,I_1}, \qquad I_1 \;=\; -\mathrm{M}\left[\,\frac{\partial^2}{\partial\theta^2}\ln f_{\mathrm{X}}(x;\theta)\right], \tag{3.5}$$

де I_1 – кількість інформації про параметр θ, що міститься в одному спостереженні, n – число виконаних випробувань, тобто оцінка параметра може бути отримана за кожним з n випробувань.

Отже, швидкість збіжності вибіркової дисперсії $\mathrm{D}[\hat{\theta}]$ до нуля не може бути більш швидкою ніж $1/n$.

Визначення 3. Незсунена оцінка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, для якої в нерівності Крамера-Рао (3.4) досягається знак рівності, називається *ефективною*.

У математичній статистиці застосовуються також асимптотично ефективні оцінки. Дисперсія асимптотично ефективних оцінок прямує до нижнього кордону нерівності Крамера-Рао при необмеженому збільшенні обсягу вибірки, тобто при $n \to \infty$.

Крім вище перерахованих трьох основних властивостей "добрих" оцінок (спроможність, незсуненість, ефективність), існує ще поняття достатньої оцінки.

Визначення 4. Оцінка $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$ називається *достатньою*, якщо вона використовує всю інформацію, що міститься у вибірці відносно параметра, що оцінюється. Достатні оцінки побудовані таким чином, що ніякі інші оцінки не можуть дати будь-якої додаткової інформації про параметри, що оцінюються.

Крім вище вказаних властивостей, які повинні мати "добрі" оцінки, є також інші. Наприклад, бажано, щоб оцінки параметрів мали простий лінійний вигляд. На жаль, не завжди можливо знайти при побудові оцінок функції $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, які мали б всі вказані властивості.

Нижче наведемо основні методи знаходження точкових оцінок невідомих параметрів і вкажемо властивості таких оцінок.

3.3. Метод моментів

Нехай за результатами вибірки $x_1, x_2, \ldots, x_n$ обсягом n, що витягнута з генеральної сукупності з функцією розподілу $F(x;\theta)$, потрібно оцінити невідомий параметр θ цього розподілу.

Аналогічно з моментами випадкової величини X введемо поняття емпіричних моментів. *Емпіричні* і відповідні їм *теоретичні початкові моменти* порядку k визначаються наступними формулами:

т е о р е т и ч н і :

$$\nu_k = \sum_{i=1}^{n} x_i^k \, p_i(x;\theta) \text{ – для дискретних ВВ X;} \tag{3.6}$$

$$\nu_k = \int_{-\infty}^{\infty} x^k f(x;\theta)\, dx \text{ – для неперервних ВВ X;} \tag{3.7}$$

е м п і р и ч н і :

$$\nu_k^* = \overline{x^k} = \frac{1}{n} \sum_{i=1}^{n} x_i^k. \tag{3.8}$$

Зокрема, позначимо, що початковий теоретичний момент першого порядку є математичним сподіванням ВВ X, а емпіричний початковий момент першого порядку є середнім арифметичним значенням спостережених значень ВВ X, тобто

$$\nu_1^* = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.9}$$

Аналогічно емпіричні і відповідні їм теоретичні *центральні моменти* порядку k визначаються такими формулами:

т е о р е т и ч н і :

$$\mu_k = \sum_{i=1}^{n} (x_i - m_x)^k \, p_i(x;\theta) \text{ – для дискретних ВВ X;} \tag{3.10a}$$

$$\mu_k = \int_{-\infty}^{\infty} (x - m_x)^k \, f(x;\theta)\, dx \text{ – для неперервних ВВ X;} \tag{3.10b}$$

е м п і р и ч н і :

$$\mu_k^* = \overline{(x - \overline{x})^k} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^k. \tag{3.11}$$

При цьому теоретичний центральний момент другого порядку є дисперсією випадкової величини X.

Потрібно мати на увазі, що емпіричні моменти є випадковими величинами, в той час як теоретичні моменти є фіксованими постійними величинами.

Найбільш простим є метод оцінювання, що називається *методом моментів*, запропонований англійським статистиком Карлом Пірсоном у 1894 р. Метод моментів засновується на тому, що емпіричні моменти (або їх функції) приймаються за оцінки відповідних теоретичних моментів (або їх функцій) і параметри виражаються через ці моменти.

Наприклад, для знаходження оцінок параметрів функції розподілу $F(x;\theta_1,\theta_2)$, що містять два невідомих параметри θ_1 та θ_2, складається система двох рівнянь

$$\begin{cases} \nu_1(\hat\theta_1,\hat\theta_2) = \nu_1^*(\hat\theta_1,\hat\theta_2), \\ \mu_2(\hat\theta_1,\hat\theta_2) = \mu_2^*(\hat\theta_1,\hat\theta_2). \end{cases} \tag{3.12}$$

Розв'язуючи цю систему, знаходять оцінки $\hat\theta_1$ та $\hat\theta_2$ функції розподілу $F(x;\hat\theta_1,\hat\theta_2)$.

Таким чином, порівнюючи теоретичний початковий момент першого порядку з емпіричним початковим моментом першого порядку, приходимо до висновку, що оцінкою математичного сподівання випадкової величини X, розподіленою згідно з будь-яким законом, є середнє арифметичне спостережених значень випадкової величини X

$$\mathrm{M}[\hat{\mathrm{X}}] = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.13}$$

Порівнюючи теоретичний і емпіричний центральні моменти другого порядку, приходимо до висновку, що оцінка дисперсії випадкової величини X, розподіленої згідно з будь-яким законом, знаходиться за формулою

$$\mathrm{D}[\hat{\mathrm{X}}] = \hat\sigma_x^2 = s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{3.14}$$

Діючи аналогічно, можна знайти оцінки теоретичних моментів будь-якого порядку.

Метод моментів відрізняється простотою, однак оцінки, знайдені цим методом, як правило, є зсуненими і малоефективними, тобто не є найкращими з можливих. Виключенням є лише нормальний розподіл, при якому метод моментів дає ефективні та спроможні оцінки $\overline{x}$ і s параметрів m_{X} і σ_{X}. З аналітичними дослідженнями властивостей оцінок, знайдених методом моментів, можна ознайомитись в курсах математичної статистики і теорії оцінювання.

3.4. Метод найбільшої правдоподібності

Метод найбільшої правдоподібності є широко розповсюдженим методом точкової оцінки. Він запропонований в 1912 р. англійським статистиком Робертом Фішером.

Нехай в дослідженні з генеральної сукупності з густиною розподілу ймовірностей $f(x;\theta)$ зроблена вибірка обсягом n, яка містить отримані результати x_1, x_2, ..., x_n. Припустимо спочатку, що X – дискретна випадкова величина, закон розподілу якої залежить від невідомого параметра θ. Наприклад, можна передбачити, що випадкова величина X розподілена згідно із законом Пуассона $\frac{1}{k!}\lambda^k \exp(-\lambda)$, де інтенсивність λ – невідомий параметр, який треба оцінити за даними вибірки. Будемо розглядати результати вибірки як реалізацію n-вимірної випадкової величини (X_1, X_2, ..., X_n). Передбачимо далі, що складові цієї випадкової величини незалежні та отримані в однорідних умовах.

У цьому випадку ймовірність (її називають *функцією правдоподібності*) того, що складові набудуть значень, які дорівнюють спостереженим значенням, така:

$$L \;=\; \Pr\left(\mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_n = x_n\right) \;= \tag{3.15}$$

$$= \Pr(x_1;\theta) \cdot \Pr(x_2;\theta) \cdot \ldots \cdot \Pr(x_n;\theta) \;=\; \prod_{i=1}^{n} \Pr(x_i;\theta)\,.$$

У випадку неперервної випадкової величини функція правдоподібності має вигляд

$$L \;=\; f(x_1, x_2, \ldots, x_n) \;=\; f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta) \;=\; \prod_{i=1}^{n} f(x_i;\theta)\,. \tag{3.16}$$

Формула (3.16) визначає густину розподілу ймовірностей неперервної випадкової величини (X_1, X_2, ..., X_n) (або густину розподілу вибірки).

Як оцінка невідомого параметра θ, знайдена за методом найбільшої правдоподібності, вибирається така функція $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, яка максимізує функцію правдоподібності. Отже, на основі відомих правил диференціального обчислення для знаходження оцінок найбільшої правдоподібності складається система m рівнянь (тут m – кількість параметрів, що оцінюються):

$$\frac{\partial}{\partial\theta_i} L \;=\; 0, \qquad\qquad i = 1, 2, \ldots, m, \tag{3.17}$$

і вибирається те розв'язання, яке перетворює функцію правдоподібності в максимум. Знайдені таким методом оцінки називають *оцінками максимальної (найбільшої) правдоподібності* або ОМП-оцінками. Оскільки екстремум функцій L і $\ln L$ досягається при одних і тих же значеннях вибіркової функції $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$, то іноді, для спрощення розрахунків, користуються *логарифмічною функцією правдоподібності*. У цьому випадку оцінки найбільшої правдоподібності знаходяться з системи рівнянь:

$$\frac{\partial}{\partial \theta_i} \ln L \;=\; 0, \qquad i = 1, 2, \ldots, m. \tag{3.18}$$

Метод найбільшої правдоподібності, що розглядається, має ряд переваг в порівнянні з методом моментів.

Наведемо деякі важливі властивості оцінок найбільшої правдоподібності.

1. Метод найбільшої правдоподібності дає спроможні оцінки.

2. Якщо існує ефективна оцінка, то метод найбільшої правдоподібності дає саме цю оцінку і іншої ОМП-оцінки не існує.

3. Оцінки найбільшої правдоподібності асимптотично ефективні.

4. Оцінки найбільшої правдоподібності мають асимптотично нормальний розподіл з параметрами:

$$\mathrm{M}[\hat{\theta}] \;=\; \theta, \qquad \mathrm{D}[\hat{\theta}] \;=\; -\frac{1}{n\,\mathrm{M}\left[\partial^2 \ln f(x;\theta)/\partial\theta^2\right]}. \tag{3.19}$$

5. Якщо існують достатні оцінки, то метод найбільшої правдоподібності дає ці оцінки.

Недоліком цього методу є те, що іноді оцінки найбільшої правдоподібності є зсуненими. Вони мають місце лише при виконанні певних умов регулярності. Зсув оцінок можна компенсувати введенням поправок (із зростанням n зсув меншає, тобто оцінки найбільшої правдоподібності асимптотично незсунені). Крім того, для знаходження оцінок методом найбільшої правдоподібності часто доводиться розв'язувати складні системи рівнянь.

3.5. Точкові оцінки невідомих параметрів розподілу

Крім наведених вище методу моментів і методу максимальної правдоподібності, оцінки середнього положення можуть знаходитися й іншими методами, наприклад, методом квантилів або методом найменших квадратів, які будуть розглянуті нижче. Крім того, в практиці застосовується і ряд інших оцінок, запропонованих інтуїтивно, без повного теоретичного обгрунтування.

Для оцінки математичного сподівання за вибіркою обсягом n застосовуються, наприклад:

а) середня гармонічна

$$\overline{x}_{\text{гарм}} \;=\; \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \;=\; n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}; \tag{3.20а}$$

б) середня геометрична

$$\overline{x}_{\text{геом}} \;=\; \left(\prod_{i=1}^{n} x_i\right)^{1/n}; \tag{3.20b}$$

в) степеневі середні зі сталою γ

$$\overline{x}_{\text{степ}} \;=\; \left(\prod_{i=1}^{n} x_i^{\gamma}\right)^{1/\gamma}; \tag{3.20c}$$

г) медіана $x_{\text{мед}}$ – значення випадкової величини X, що припадає на середину впорядкованого (ранжированого) статистичного ряду;

д) мода $x_{\text{мода}}$ – спостережене значення випадкової величини X, що зустрічається з найбільшою частотою.

Для оцінки дисперсії спостережених значень випадкової величини іноді застосовуються:

а) розмах

$$R \;=\; x_{\max} - x_{\min}, \tag{3.20d}$$

де $x_{\max}$ – найбільше зі спостережених значень випадкової величини X; $x_{\min}$ – найменше зі спостережених значень випадкової величини X;

б) коефіцієнт варіації

$$var \;=\; s/\overline{x}\;100\,(\%), \tag{3.20e}$$

де $s = (\mu_2^*)^{1/2}$.

В більшості практичних випадків ці оцінки можна вважати оцінками "гіршими за якістю", ніж оцінки, отримані методом найбільшої правдоподібності або методом моментів. Їх застосування пояснюється або простотою їх обчислення, або традицією, що історично склалася. Тому ці оцінки можна застосовувати як додаткові характеристики, що описують центр розподілу або розсіювання спостережених значень випадкової величини X.

3.6. Інтервальні оцінки параметрів. Точність знаходження оцінок

Точкова оцінка невідомого параметра θ, що знайдена за вибіркою обсягом n з генеральної сукупності з функцією розподілу $F(x;\theta)$, не дозволяє безпосередньо відповісти на питання, яку помилку ми здійснюємо, приймаючи замість точного значення параметра θ деяке його наближене значення (оцінку) $\hat{\theta} = u(x_1, x_2, \ldots, x_n)$. У зв'язку з цим в багатьох випадках більш вигідно користуватися інтервальною оцінкою, заснованою на визначенні деякого інтервалу, всередині якого з певною ймовірністю знаходиться значення параметра θ.

Нехай статистична характеристика $\hat{\theta} \;=\; u(x_1, x_2, \ldots, x_n)$, що знайдена за результатами вибірки обсягом n, є точковою оцінкою невідомого параметра θ. Чим

менше різниця $|\theta - \hat{\theta}|$, тим краще якість оцінки, тим точніше оцінка. Таким чином, додатне число ε характеризує точність оцінки

$$|\theta - \hat{\theta}| \ < \ \varepsilon. \tag{3.21}$$

Зрозуміло, що точність ε залежить від обсягу вибірки n. Який повинен бути обсяг вибірки n, щоб забезпечити задану точність ε, або як визначити точність ε при даному обсязі вибірки? На ці питання безпосередньо відповісти, використовуючи наведену нерівність, неможливо, оскільки статистичні методи не дозволяють категорично стверджувати, що оцінка задовольняє цій нерівності в значенні математичного аналізу, тобто з деякого обсягу вибірки n. Можна тільки говорити про ймовірність $\mathrm{P} = 1 - \alpha$, з якою вона виконується.

Визначення 1. *Довірчою ймовірністю* оцінки називають імовірність $\mathrm{P} = 1 - \alpha$ виконання нерівності $|\theta - \hat{\theta}| < \varepsilon$.

Звичайно, довірча ймовірність оцінки задається заздалегідь. Найчастіше вважають: $1 - \alpha = 0,95;\ 0,99;\ 0,9973;\ 0,999$.

Довірча ймовірність точкової оцінки показує, що при здобутті вибірок обсягом n з однієї і тієї ж генеральної сукупності з функцією розподілу $F(x;\theta)$ в $(1-\alpha)\cdot 100\%$ випадках параметр θ буде накриватися даним інтервалом. Якщо довірча ймовірність $(1-\alpha)$ вибрана досить близько до одиниці $(1-\alpha) \geq 0,90$, то число ε визначає межову похибку точкової оцінки невідомого параметра θ.

Нехай імовірність того, що $|\theta - \hat{\theta}| < \varepsilon$, дорівнює $1 - \alpha$:

$$\Pr\left(|\theta - \hat{\theta}| < \varepsilon\right) \ = \ 1 - \alpha. \tag{3.22}$$

Перетворюємо формулу:

$$\Pr\left(\hat{\theta} - \varepsilon \ < \ \theta \ < \ \hat{\theta} + \varepsilon\right) \ = \ 1 - \alpha. \tag{3.23}$$

Здобута формула (3.23) показує, що невідомий параметр θ міститься всередині інтервалу $[\hat{\theta} - \varepsilon;\, \hat{\theta} + \varepsilon]$ (рис. 3.2).

Визначення 2. *Довірчим інтервалом* називається інтервал $[\hat{\theta} - \varepsilon;\, \hat{\theta} + \varepsilon]$, який накриває невідомий параметр θ, що оцінюється, із заданою довірчою ймовірністю $\mathrm{P} = 1 - \alpha$.

На практиці важливу роль відіграє довжина довірчого інтервалу. Чим менше довжина довірчого інтервалу $[\hat{\theta} - \varepsilon;\, \hat{\theta} + \varepsilon]$, тим точніше оцінка. Якщо ж довжина довірчого інтервалу велика, то оцінка малопридатна для практики.

З формули (3.23) випливає, що довжина довірчого інтервалу дорівнює 2ε. Аналізуючи формулу (3.23), приходимо до висновку, що довжина довірчого інтервалу 2ε визначається двома величинами: довірчою ймовірністю $\mathrm{P} = 1 - \alpha$ і обсягом вибірки n. Таким чином, величини ε, $1 - \alpha$ і n тісно взаємопов'язані. Задаючи визначені значення двом з них, можна визначити величину третьої.

Загальна схема побудови довірчих інтервалів для параметрів нормального закону розподілу ймовірностей полягає в наступному.

1. З генеральної сукупності з функцією розподілу $F(x;\theta)$ витягується вибірка обсягом n. За результатами цієї вибірки методом найбільшої правдоподібності або методом моментів знаходиться точкова оцінка параметра θ, що оцінюється.

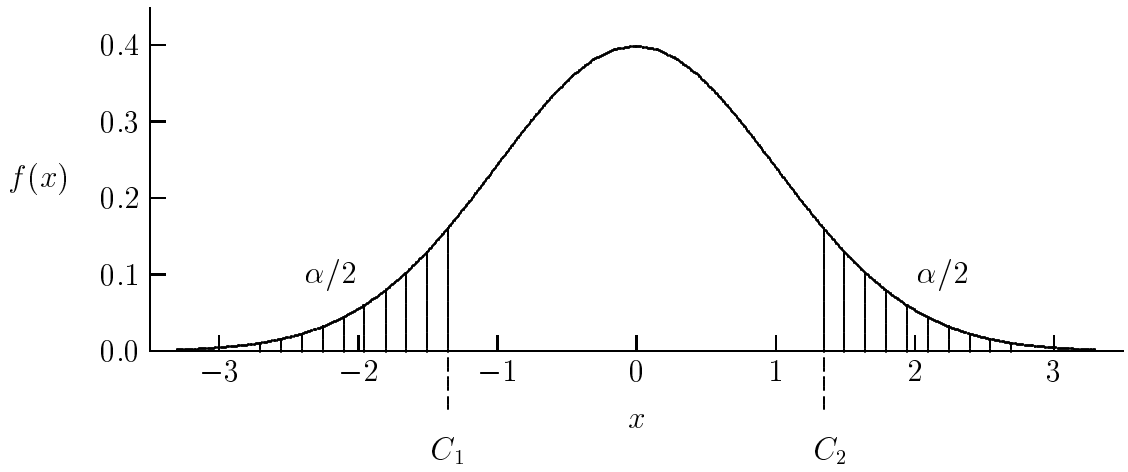


Рисунок 3.2 — До побудови довірчого інтервалу

2. Складається випадкова величина, наприклад $\mathrm{Y}(\theta)$, яка пов'язана з параметром θ і має відому густину розподілу ймовірностей $f_{\mathrm{Y}}(y;\theta)$.

3. Задають довірчу ймовірність $1-\alpha$. Звичайно, приймають довірчу ймовірність $1-\alpha$, яка дорівнює 0,90; 0,95; 0,99; 0,9973; 0,999.

4. Використовуючи густину розподілу імовірностей випадкової величини Y, знаходять такі два числа C_1 і C_2, що

$$\Pr\left(C_1 < \mathrm{Y} < C_2\right) \;=\; \int_{C_1}^{C_2} f_{\mathrm{Y}}(y;\theta)\,dy \;=\; 1-\alpha. \tag{3.24}$$

Значення C_1 і C_2 вибираються, як правило, за симетрічних умов:

$$\Pr\left(\mathrm{Y}(\theta) < C_1\right) \;=\; \frac{\alpha}{2}, \qquad \Pr\left(\mathrm{Y}(\theta) > C_2\right) \;=\; \frac{\alpha}{2}, \tag{3.25}$$

тобто, щоб сумарна площа фігур, які заштриховані на рис. 3.2, дорівнювала за величиною α.

5. Нерівність, що міститься в круглих дужках рівняння (3.24), перетворюється в рівносильну нерівність

$$\Pr\left(\hat{\theta}-\varepsilon \;<\; \theta \;<\; \hat{\theta}+\varepsilon\right) \;=\; 1-\alpha, \tag{3.26}$$

що накриває із заданою ймовірністю $1-\alpha$ невідомий параметр θ.

3.7. Довірчі інтервали для математичного сподівання нормальної випадкової величини з відомою дисперсією

Задачу аналіза довірчих інтервалів при оцінюванні математичного сподівання нормальної випадкової величини розглядають у двох варіантах:

– випадок, коли дисперсія відома;

– випадок, коли дисперсія невідома.

Використаємо вказану вище схему для знаходження довірчих інтервалів параметрів нормального закону a і σ. Розглянемо нормальну модель генеральної сукупності $\mathcal{N}(a;\,\sigma)$, в якій параметр σ будемо вважати фіксованим (відомим), а параметр a — невідомим. Для знаходження точкової оцінки параметра a з генеральної нормальної сукупності витягнута вибірка обсягом n. На основі цієї вибірки знайдена точкова оцінка математичного сподівання

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.27}$$

Раніше було показано, що якщо $\mathrm{X} \to \mathcal{N}(a;\,\sigma)$, то випадкова величина

$$\overline{\mathrm{X}} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i,$$

яка лінійно зв'язана з $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$, розподілена також згідно з нормальним законом, але з математичним сподіванням a і дисперсією σ^2/n, тобто $\overline{\mathrm{X}} \to \mathcal{N}(a;\,\sigma/\sqrt{n})$.

З метою побудувати довірчий інтервал для параметра a складемо стандартизовану випадкову величину (вибіркову статистику)

$$u \;=\; \frac{\overline{x} - a}{\sigma/\sqrt{n}}. \tag{3.28}$$

Випадкова величина u має стандартизований нормальний розподіл $u \to \mathcal{N}(0;\,1)$. Ймовірність того, що стандартизована випадкова величина u відхилиться від свого математичного сподівання на величину $u_{\alpha/2}$, знаходиться за формулою (рис. 3.3)

$$\Pr\left(-u_{\alpha/2} < \frac{\overline{x} - a}{\sigma/\sqrt{n}} < u_{\alpha/2}\right) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{-u_{\alpha/2}}^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt. \tag{3.29}$$

Задамо цю ймовірність, щоб дорівнювала $\mathrm{P} = 1 - \alpha$, тоді отримаємо

$$\Pr\left(-u_{\alpha/2} < \frac{\overline{x} - a}{\sigma/\sqrt{n}} < u_{\alpha/2}\right) \;=\; \frac{2}{\sqrt{2\pi}} \int\limits_{0}^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt \;=\; 1 - \alpha. \tag{3.30}$$

Розв'язуючи рівняння

$$\frac{2}{\sqrt{2\pi}} \int\limits_{0}^{u_{\alpha/2}} \exp\left(-t^2/2\right)\, dt \;=\; \Phi\left(u_{\alpha/2}\right) \;=\; 1 - \alpha, \tag{3.31}$$

знаходять *квантилі* стандартизованого нормального розподілу $u_{\alpha/2}$. Звичайно, квантилі нормального розподілу $u_{\alpha/2}$ знаходять у таблицях функції Лапласа (див. додаток) з умови $\Phi(u_{\alpha/2}) = 1 - \alpha$.

Для найбільш розповсюджених значень довірчої ймовірності $\mathrm{P} = 1 - \alpha$ квантилі стандартизованого нормального розподілу наведені в табл. 3.1.

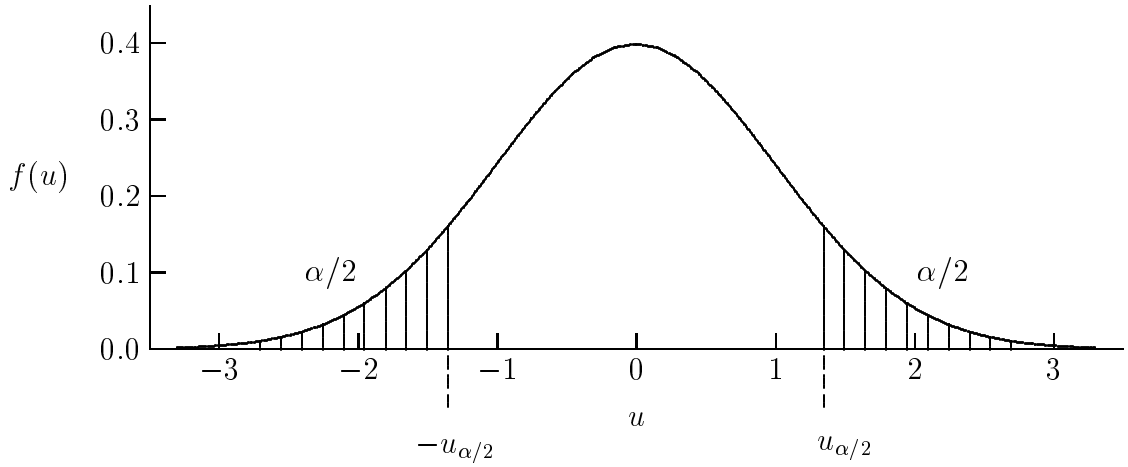


Рисунок 3.3 — До побудови довірчого інтервалу для оцінки математичного сподівання нормальної випадкової величини при відомому σ

Будемо вважати квантиль $u_{\alpha/2}$, що відповідає заданій довірчій ймовірності $\mathrm{P} = 1-\alpha$, відомим. Перетворюємо нерівність, що міститься в лівій частині рівняння (3.29),

$$\Pr\left(-\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} \;<\; (\overline{x}-a) \;<\; \frac{\sigma}{\sqrt{n}}\,u_{\alpha/2}\right) \;=\; 1-\alpha$$

або

$$\Pr\left(\overline{x}-\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} < a < \overline{x}+\frac{\sigma}{\sqrt{n}}\,u_{\alpha/2}\right) \;=\; 1-\alpha. \tag{3.32}$$

Отже, довірчий інтервал $[\overline{x}-\sigma\,u_{\alpha/2}/\sqrt{n};\ \overline{x}+\sigma\,u_{\alpha/2}/\sqrt{n}]$ накриває невідоме математичне сподівання a із заданою ймовірністю $1-\alpha$. Звідси випливає, що точність оцінки математичного сподівання (гранична похибка) така: $\varepsilon = \sigma\,u_{\alpha/2}/\sqrt{n}$.

Зауваження 1. Аналізуючи формулу (3.32), помічаємо, що:

а) збільшення обсягу вибірки n призводить до зменшення довжини довірчого інтервалу, тобто до поліпшення точності ε;

б) збільшення довірчої ймовірності $1-\alpha$ призводить до збільшення довжини довірчого інтервалу, тобто до зменшення точності ε;

в) якщо задати точність (межову похибку інтервальної оцінки) ε і довірчу імовірність $1-\alpha$, із співвідношення

$$\varepsilon \;=\; \frac{\sigma}{\sqrt{n}}\,u_{\alpha/2} \tag{3.33}$$

можна знайти такий мінімальний обсяг вибірки n, який забезпечить задану точність

$$n \;=\; \frac{1}{\varepsilon^2}\,u_{\alpha/2}^2\sigma^2. \tag{3.34}$$

Зауваження 2. Нехай ВВ X має довільну функцію розподілу ймовірностей $F(x)$. Нехай також для ВВ X виконується центральна гранична теорема теорії ймовірностей. Тоді, внаслідок граничної теореми, при досить великому обсязі

Таблиця 3.1 — Квантилі стандартизованого нормального розподілу

Довірча ймовірність $\mathrm{P} = 1 - \alpha$	Квантиль $u_{\alpha/2}$
0,90	1,64
0,95	1,96
0,99	2,58
0,9973	3,00
0,999	3,37

вибірки закон розподілу вибіркової статистики $u = (\overline{x} - \mathrm{M}[\overline{x}])/\sigma_{\overline{x}}$ буде приблизно нормальним. У цьому випадку довірчий інтервал для M[X] випадкової величини X, визначений за формулою (3.32), буде наближеним. З метою побудови точного довірчого інтервалу для M[X] випадкової величини X, що має довільний закон розподілу, необхідно знати або закон розподілу середньої арифметичної $\overline{x} = \frac{1}{n}\sum_i x_i$, або закон розподілу вибіркової статистики $u = (\overline{x} - \mathrm{M}[\overline{x}])/\sigma_{\overline{x}}$.

Закони розподілу цих величин залежать від закону розподілу випадкової величини X.

Приклад

Хлопчики та дівчатка народжуються не однаково часто. Цей факт був відомий дослідникам уже давно. Для побудови математичних моделей, які придатні відслідковувати процеси, що спостерігаються серед населення держави, необхідно мати великий статистичний матеріал.

З 1874 року до 1900 року в Швейцарії народилось 1359671 хлопчиків і 1285086 дівчаток, тобто разом народилося 2644757 дітей.

Яка ймовірність народження хлопчиків?

Розв'язання

Статистична частота народження хлопчиків за вказані роки дорівнює

$$\nu \;=\; 1359671/2644757 \;=\; 0,5141.$$

Можна прийти до висновку, що тут ми маємо справу з послідовними випробуваннями з двома можливими результатами в кожному, причому кількість іспитів n в досліді досить велика. Тому можна використати як теорему Бернуллі, так і теорему Муавра–Лапласа. Перша з них указує на те, що ймовірність появи хлопчика близька до 1/2 (як і ймовірність народження дівчинки). В математичній моделі Бернуллі частота народження хлопчика w_n, що спостерігається серед n народжених дітей, має дисперсію $\mathrm{D}[w_n] = p(1-p)/n$. Оскільки n дуже велике (воно дорівнює 2644757) та ймовірність p близька до 1/2, то з великою точністю маємо $\mathrm{D}[w_n] = (0,0003)^2$.

Якщо практично достовірною домовитися вважати подію з імовірністю не менш за 0,95, то з оцінки (3.32) випливає, що ймовірність p народження хлопчика міститься між такими межами:

$$p \;>\; \nu - 1,96\sqrt{\mathrm{D}[w_n]} \;=\; 0,5135,$$

$$p \;<\; \nu + 1,96\sqrt{\mathrm{D}[w_n]} \;=\; 0,5146.$$

Значення з такою високою отриманою точністю ймовірності p дозволяє прийняти практично достовірно, що протягом кількох років, які слідуватимуть після 1900 року, кількість хлопчиків, що народжуються на кожні 10000 дітей, складає від 5135 до 5146. Як показують численні спостереджения, що виконувались за тривалий час, ймовірність p є достатньо стійкою характеристикою.

3.8. Довірчі інтервали для математичного сподівання нормальної випадкової величини при невідомій дисперсії

Нехай випадкова величина $\mathrm{X} \to \mathcal{N}(a;\, \sigma)$, причому параметри a і σ є невідомими. Для знаходження точкових оцінок параметрів a і σ з генеральної нормальної сукупності витягнута вибірка обсягом n. На основі цієї вибірки знайдені точкові незсунені оцінки невідомих параметрів

$$\hat{a} \;=\; \overline{x}; \qquad \hat{\sigma} \;=\; s \;=\; \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.35}$$

З метою побудови довірчого інтервалу для математичного сподівання складемо допоміжну випадкову величину (вибіркову статистику)

$$t \;=\; \frac{\overline{x} - z}{s/\sqrt{n-1}}. \tag{3.36}$$

Перемножуючи чисельник і знаменник (3.36) на позитивну величину $\sqrt{n}/\sigma$, отримаємо

$$t \;=\; \frac{\overline{x} - a}{\sigma/\sqrt{n}} \cdot \left(\frac{1}{n-1} \cdot \frac{ns^2}{\sigma^2} \right)^{-1/2}. \tag{3.37}$$

У курсі теорії ймовірностей доводиться теорема про те, що величина ns^2/σ^2 розподілена згідно із законом χ^2 з $\nu = n-1$ ступенями вільності. Введемо позначення $\mathrm{Y} = (\overline{x} - a)\sqrt{n}/\sigma$, тоді

$$t \;=\; \mathrm{Y} \left(\frac{1}{n-1}\chi^2_{n-1} \right)^{-1/2}. \tag{3.38}$$

Неважко помітити, що $\mathrm{Y} \to \mathcal{N}(0;\, 1)$. Аналізуючи випадкову величину t, що визначається рівністю (3.38), приходимо до висновку, що ця випадкова величина підкоряється закону Стьюдента з $\nu = n-1$ ступенями вільності. Ймовірність того, що випадкова величина t попаде в інтервал $[-t_{\alpha/2,\,n-1};\, t_{\alpha/2,\,n-1}]$, знаходиться за формулою (рис. 3.4)

$$\Pr\left(-t_{\alpha/2,\,n-1} \;<\; \frac{\overline{x} - a}{s/\sqrt{n}} \;<\; t_{\alpha/2,\,n-1} \right) \;=\; 2 \int_0^{t_{\alpha/2,\,n-1}} f(t)\, dt. \tag{3.39}$$

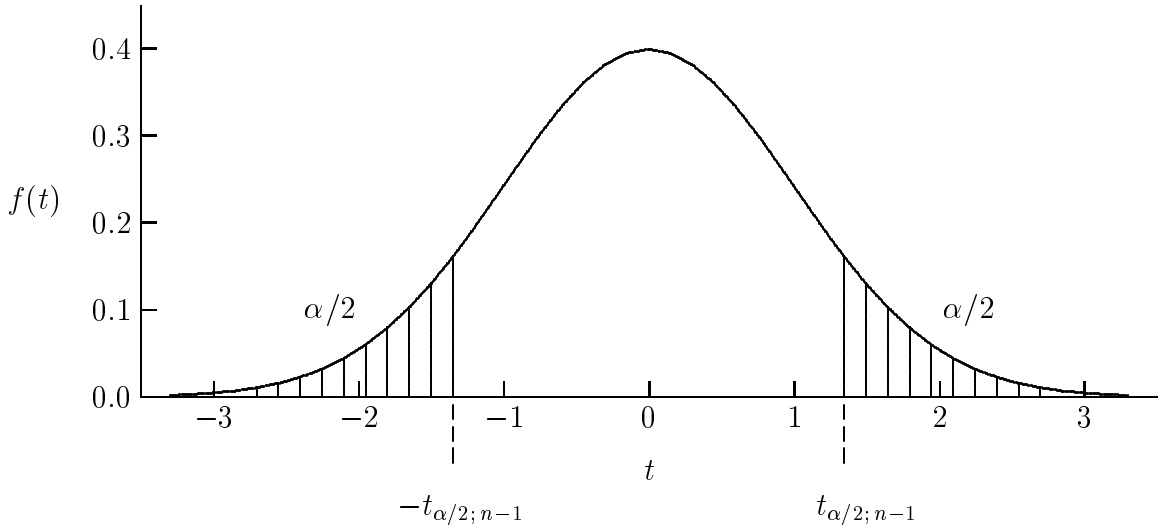


Рисунок 3.4 — До побудови довірчого інтервалу для оцінки математичного сподівання випадкової величини, розподіленої згідно із законом Стьюдента

Задамо так, щоб ця імовірність дорівнювала $1-\alpha$. Тоді з розв'язку рівняння

$$2\int_{0}^{t_{\alpha/2,\,n-1}} f(t)\,dt \;=\; 1-\alpha \tag{3.40}$$

можна знайти квантиль розподілу Стьюдента $t_{\alpha/2,\,n-1}$. Є спеціальні таблиці (див. додаток), які містять квантилі t–розподілу, що відповідають довірчій ймовірності $1-\alpha$ і обсягу вибірки n.

Будемо вважати квантилі $t_{\alpha/2,\,n-1}$, відповідні заданій довірчій ймовірності $1-\alpha$ і обсягу вибірки n, відомими. Перетворюємо нерівність, що міститься в лівій частині рівняння (3.39):

$$\Pr\left(-t_{\alpha/2,\,n-1}\,s/\sqrt{n} \;<\; (\overline{x}-a) \;<\; t_{\alpha/2,\,n-1}\,s/\sqrt{n}\right) \;=\; 1-\alpha \tag{3.41a}$$

або

$$\Pr\left(\overline{x}-t_{\alpha/2,\,n-1}\,s/\sqrt{n} \;<\; a \;<\; \overline{x}+t_{\alpha/2,\,n-1}\,s/\sqrt{n}\right) \;=\; 1-\alpha. \tag{3.41b}$$

Отже, довірчий інтервал $(\overline{x}-t_{\alpha/2,\,n-1}\,s/\sqrt{n},\ \overline{x}+t_{\alpha/2,\,n-1}\,s/\sqrt{n})$ накриває невідоме математичне сподівання із заданою ймовірністю $1-\alpha$. Точність (межова похибка) оцінки математичного сподівання наступна:

$$\varepsilon \;=\; \frac{s}{\sqrt{n}}\,t_{\alpha/2,\,n-1}. \tag{3.42}$$

Отриманий довірчий інтервал має такі ж властивості, що і довірчий інтервал для математичного сподівання при відомій σ.

Зауваження. Вище було зазначено, що при необмеженому зростанні обсягу вибірки розподіл Стьюдента прямує до нормального. Тому вже при $n \geq 30$ з метою побудови довірчих інтервалів для математичного сподівання можна замість розподілу Стьюдента використати нормальний розподіл. У цьому випадку наближені довірчі інтервали для математичного сподівання знаходяться за формулою (3.32), в якій потрібно прийняти $\sigma = s$.

3.9. Довірчий інтервал для середнього квадратичного відхилення нормальної випадкової величини

Нехай випадкова величина X $\to \mathcal{N}(a, \sigma)$, причому параметри a і σ невідомі. Для знаходження точкових оцінок параметрів a і σ з генеральної нормальної сукупності витягнута вибірка обсягом n. Нехай також на основі цієї вибірки знайдені точкові незсунені оцінки невідомих параметрів

$$\hat{a} \;=\; \overline{x}; \qquad \hat{\sigma} \;=\; s \;=\; \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.43}$$

З метою побудови довірчого інтервалу для середнього квадратичного відхилення σ складемо допоміжну випадкову величину

$$\chi^2 \;=\; \frac{(n-1)s^2}{\sigma^2}. \tag{3.44}$$

Ця випадкова величина має розподіл χ^2 з $n-1$ ступенями вільності.

Ймовірність того, що випадкова величина χ^2 влучить в інтервал $(C_1; C_2)$, дорівнює (рисунок 3.5)

$$\Pr\left(C_1 < \chi^2 < C_2\right) \;=\; \int_{C_1}^{C_2} f(\chi^2)\, d\chi^2. \tag{3.45}$$

Задамо цю ймовірність, щоб дорівнювала $1-\alpha$. Виберемо значення C_1 і C_2 з умов

$$\Pr\left(\chi^2 > C_2\right) \;=\; \int_{C_2}^{\infty} f(\chi^2)\, d\chi^2 \;=\; \frac{\alpha}{2}, \tag{3.46a}$$

$$\Pr\left(\chi^2 < C_1\right) \;=\; \int_{0}^{C_1} f(\chi^2)\, d\chi^2 \;=\; \frac{\alpha}{2}. \tag{3.46b}$$

На рис. 3.5 відповідні дві криволінійні трапеції, кожна з площею $\alpha/2$, заштриховані.

Розв'язуючи (наприклад, чисельно) рівняння (3.46), знаходять квантилі χ^2-розподілу $C_1 = \chi^2_{1-\alpha/2;\, n-1}$ і $C_2 = \chi^2_{\alpha/2;\, n-1}$. Будемо надалі вважати квантилі χ^2-розподілу відомими.

Перетворюючи формулу

$$\Pr\left(\chi^2_{1-\alpha/2;\, n-1} \;<\; \frac{(n-1)s^2}{\sigma^2} \;<\; \chi^2_{\alpha/2;\, n-1}\right) \;=\; 1-\alpha,$$

маємо

$$\Pr\left(s\sqrt{\frac{n-1}{\chi^2_{\alpha/2;\, n-1}}} \;<\; \sigma \;<\; s\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;\, n-1}}}\right) \;=\; 1-\alpha, \tag{3.47}$$

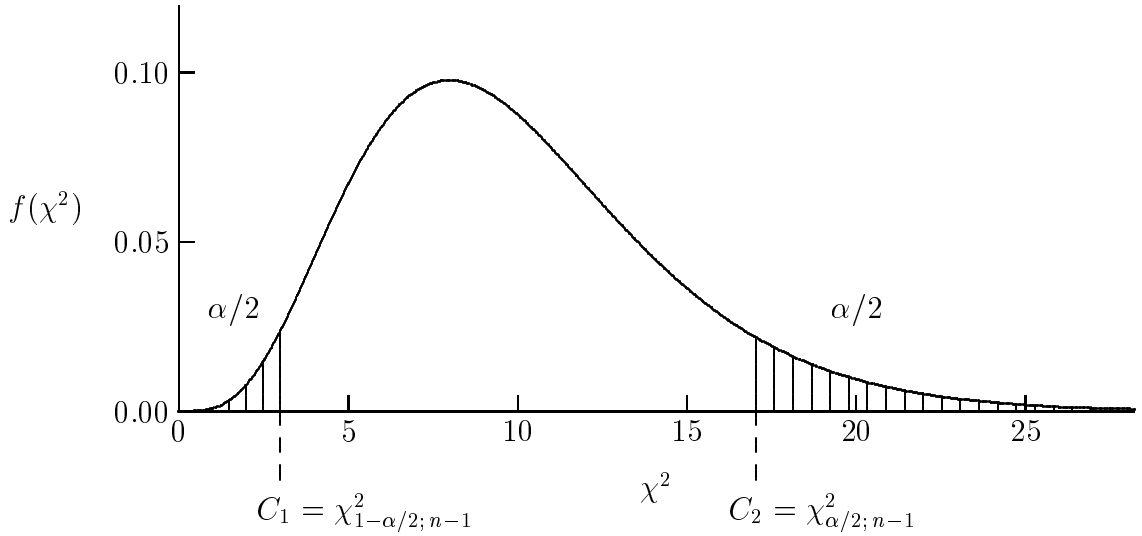


Рисунок 3.5 — До довірчого інтервалу для середнього квадратичного відхилення σ нормальної випадкової величини

або, більш стисло:

$$\Pr\left(s\,\gamma_1 < \sigma < s\,\gamma_2\right) \;=\; 1-\alpha, \tag{3.48}$$

де

$$\gamma_1 \;=\; \sqrt{\frac{n-1}{\chi^2_{\alpha/2;\,n-1}}}, \qquad \gamma_2 \;=\; \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;\,n-1}}}. \tag{3.49}$$

Коефіцієнти γ_1 та γ_2, відповідні довірчій імовірності $1-\alpha$ і кількості ступенів вільності n, вміщені в додатку.

Зауваження. Оскільки при $n \to \infty$ розподіл χ^2 наближається до нормального, то при досить великому обсязі вибірки ($n \geq 50$) довірчий інтервал можна знайти за формулою

$$\Pr\left(\frac{s}{1+u_{\alpha/2}/\sqrt{2n}} < \sigma < \frac{s}{1-u_{\alpha/2}/\sqrt{2n}}\right) \;=\; 1-\alpha, \tag{3.50}$$

де $u_{\alpha/2}$ – квантиль стандартизованого нормального розподілу, відповідний довірчій ймовірності $1-\alpha$.

3.10. Приклади

Приклад 3.1

Проводиться вивчення випадкової величини X з математичним сподіванням M[X]= a та дисперсією D[X]= σ^2. З генеральної сукупності вийнята вибірка $x_1, x_2, \ldots, x_N$ обсягом N, на підставі якої побудована оцінка середнього

$$X^*_N \;=\; \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{$*$}$$

Потрібно вивчити залежність дисперсії $\mathrm{D}[X_N^*]$ оцінки середнього від обсягу вибірки N.

Розв'язання

Оцінка $(*)$ є незсуненою

$$\mathrm{M}\left[X_N^*\right] \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathrm{M}\left[x_n\right] \;=\; \frac{1}{N}\sum_{n=1}^{N} a \;=\; a.$$

Введемо нову випадкову величину $\mathrm{Y} = X_N^* - a$. Оскільки при сталом значенні a маємо $\mathrm{D}[X_N^*] = \mathrm{D}[X_N^* - a] = \mathrm{D}[\mathrm{Y}]$, то далі будемо разглядати дисперсію $\mathrm{D}[\mathrm{Y}]$. Оскільки $\mathrm{M}[\mathrm{Y}] = 0$, для дисперсії оцінки $(*)$ маємо

$$\mathrm{D}\left[X_N^*\right] \;=\; \mathrm{D}[\mathrm{Y}] \;=\; \mathrm{M}\left[\mathrm{Y}^2\right] \;=\; \mathrm{M}\left[\left(\frac{1}{N}\sum_{n=1}^{N} y_n\right)^2\right],$$

де $y_n = x_n - a,\;\; n = 1, 2, \ldots, N$. Це дає

$$\mathrm{D}\left[X_N^*\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\left(\sum_{n=1}^{N} y_n\right)^2\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m\right].$$

В подвійній сумі, що виникла, виділимо N доданків, що співпадають, та $N(N-1)$ доданків, що не співпадають, тобто

$$\sum_{n=1,\;m=1}^{N} y_n y_m \;=\; \sum_{n=1}^{N}\left(y_n\right)^2 \;+\; \sum_{n=1,\;m=1,\;n\neq m}^{N} y_n y_m.$$

Математичне сподівання від кожного з $N(N-1)$ неспівпадаючих доданків дорівнює нулю, тому

$$\mathrm{D}\left[X_N^*\right] \;=\; \frac{1}{N^2}\mathrm{M}\left[\sum_{n=1}^{N} y_n^2\right] \;=\; \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{M}\left[y_n^2\right] \;=\; \frac{N}{N^2}\sigma^2 \;=\; \frac{1}{N}\mathrm{D}[\mathrm{X}].$$

Отже, з ростом обсягу вибірки дисперсія оцінки середнього $(*)$ зменшується зворотньо пропорційно N.

Приклад 3.2

Нехай випадкова величина X розподілена згідно з нормальним законом з параметрами $\mathrm{M}[\mathrm{X}] = a$ та $\sigma = \sqrt{\mathrm{D}[\mathrm{X}]}$, або, більш стисло, нехай $\mathrm{X} \to N(a;\sigma)$.

Потрібно за результатами спостережень $x_1, x_2, \ldots, x_n$ оцінити параметри a і σ та знайти оцінки асиметрії й ексцесу.

Розв'язання

Застосовуючи метод моментів, маємо

$$\left\{\begin{array}{l} \nu_1 = \nu_1^* \Rightarrow \mathrm{M}[\bar{\mathrm{X}}] \;=\; \hat{a} \;=\; \bar{x} \;=\; \frac{1}{n}\sum\limits_{i=1}^{n} x_i, \\ \mu_2 = \mu_2^* \Rightarrow \mathrm{D}[\bar{\mathrm{X}}] \;=\; \sigma_x^2 = s^2 \;=\; \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2, \end{array}\right. \qquad (*)$$

Покажемо, що середнє арифметичне значення, що спостерігається, є спроможною, незсуненою й ефективною оцінкою математичного сподівання.

Спроможність оцінки випливає з теореми Бернуллі

$$\lim_{n\to\infty} \Pr\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) \;=\; 1, \qquad (\varepsilon > 0),$$

де X_1, X_2, ..., X_n являють собою як би n примірників випадкової величини X з одним і тим же математичним сподіванням a. З теореми Бернуллі випливає, що середнє арифметичне є спроможною оцінкою математичного сподівання при будь-якому законі розподілу.

Знайдемо математичне сподівання середньої арифметичної

$$\mathrm{M}[\overline{x}] \;=\; \mathrm{M}\left[\frac{1}{n}\sum_i \mathrm{X}_i\right] \;=\; \frac{1}{n}\mathrm{M}\left[\sum_i \mathrm{X}_i\right] \;=\; \frac{1}{n}\sum_i \mathrm{M}[\mathrm{X}_i] \;=\; \frac{1}{n}\,na \;=\; a.$$

Отже, середнє арифметичне є незсуненою оцінкою математичного сподівання.

Досліджуємо ефективність оцінки $\hat{a} = \bar{x}$, вважаючи параметр σ відомим. Для цього обчислимо дисперсію середньої арифметичної

$$\mathrm{D}[\overline{x}] \;=\; \mathrm{D}\left[\frac{1}{n}\sum_i \mathrm{X}_i\right] \;=\; \frac{1}{n^2}\mathrm{D}\left[\sum_i \mathrm{X}_i\right] \;=\; \frac{1}{n^2}\sum_i \mathrm{D}[\mathrm{X}_i] \;=\; \frac{1}{n^2}\cdot n\sigma^2 \;=\; \frac{\sigma^2}{n}.$$

Знайдемо нижню межу нерівності Рао-Крамера. Для цього обчислимо

$$\mathrm{M}\left[\frac{\partial^2}{\partial a^2}\,\ln\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\right\}\right] \;=\; -\mathrm{M}\left[\frac{\partial^2}{\partial a^2}\frac{(x-a)^2}{2\sigma^2}\right] \;=\; -\frac{1}{\sigma^2}.$$

Оскільки нижня межа нерівності Рао-Крамера при $N = n$ збігається з дисперсією середньої арифметичної

$$-\,\frac{1}{n\,\mathrm{M}\left[\,\partial^2 f(x,\theta)/\partial\theta^2\right]} \;=\; \frac{\sigma^2}{n},$$

то середнє арифметичне при відомому значенні σ є ефективною оцінкою математичного сподівання.

Покажемо, що оцінка дисперсії, яка обчислюється за формулою $(*)$, є зсуненою. Знайдемо математичне сподівання емпіричної дисперсії, заздалегідь перетворивши формулу $(*)$:

$$s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - a) - (\overline{x} - a)\right]^2 \;=$$

$$=\; \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - a)^2 \;-\; 2\,(\overline{x} - a)\sum_{i=1}^{n}(x_i - a) \;+\; n\,(\overline{x} - a)^2\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\overline{x} - a)^2\,;$$

$$\mathrm{M}[s^2] \;=\; \frac{1}{n}\sum_i \mathrm{M}\left[(\mathrm{X}_i - a)^2\right] - \mathrm{M}\left[(\overline{x} - a)^2\right] \;=$$

$$= \sigma^2 - \mathrm{M}\left[\left(\frac{1}{n}\sum_i(\mathrm{X}_i - x)\right)^2\right] = \sigma^2 - \frac{1}{n^2}\mathrm{M}\left[\sum_i(\mathrm{X}_i - a)^2\right] =$$

$$= \sigma^2 - \frac{n\,\sigma^2}{n^2} = \frac{n-1}{n}\,\sigma^2.$$

Оскільки $\mathrm{M}[s^2] \neq \sigma^2$, то оцінка дисперсії, що визначається за формулою $(*)$, є зсуненою. Зсув оцінки компенсується помноженням її на величину $n/(n-1)$. При великих n $(n > 30)$ зсув оцінки незначний. При малому обсязі вибірки $(n \leq 30)$ незсунена оцінка обчислюється за формулою

$$\mathrm{D}[\mathrm{X}] = \hat{\sigma}^2_x = s^2_{\text{незм}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Наведемо без доведення деякі результати відносно властивостей оцінок дисперсії.

Якщо параметри нормального закону невідомі, то оцінка дисперсії, що визначається за формулою $(*)$, не володіє властивістю ефективності. Однак якщо заздалегідь математичне сподівання a є відомим, то оцінка $\hat{\sigma}^2 = \frac{1}{n}\sum_i(x_i - a)^2$ володіє властивістю незсуненості, спроможності і ефективності.

Оцінки асиметрії $\hat{A}$ й ексцесу $\hat{E}$, які є функціями центральних статистичних моментів, знаходяться за формулами:

$$\frac{\mu_3}{\sigma^3} = \left(\frac{\mu_3}{\sigma^3}\right)^* \Rightarrow \hat{A} = \left(\frac{\mu_3}{\sigma^3}\right)^* = \frac{\sum_i(x_i - \overline{x})^3}{ns^3};$$

$$\frac{\mu_4}{\sigma^4} - 3 = \left(\frac{\mu_4}{\sigma^4} - 3\right)^* \Rightarrow \hat{E} = \left(\frac{\mu_4}{\sigma^4}\right)^* - 3 = \frac{\sum_i(x_i - \overline{x})^4}{ns^4}.$$

Приклад 3.3

Нехай випадкова величина X розподілена згідно із законом Пуассона

$$\Pr(\mathrm{X} = k) = \frac{\lambda^k}{k!}\exp(-\lambda), \qquad k = 1, 2, \ldots.$$

Потрібно на основі спостережених значень $x_1 = k_1$, $x_2 = k_2$, ..., $x_n = k_n$ оцінити невідомий параметр λ цього розподілу.

Розв'язання

Запишемо функцію правдоподібності

$$L = \frac{\lambda^{x_1}}{x_1!}\exp(-\lambda)\frac{\lambda^{x_2}}{x_2!}\exp(-\lambda)\ldots\frac{\lambda^{x_n}}{x_n!}\exp(-\lambda) = \frac{\lambda^S}{x_1!x_2!\ldots x_n!}\exp(-n\lambda), \qquad S = \sum_{i=1}^{n} x_i.$$

Для знаходження оцінки λ, тобто такого значення параметра розподілу $\lambda = \lambda(x_1, x_2, \ldots, x_n)$, при якому функція правдоподібності досягає максимуму, зручно перейти до *логарифмічної функції правдоподібності*

$$\ln L = -n\lambda + \sum_{i=1}^{n} x_i \ln\lambda - \sum_{i=1}^{n}\ln(x_i!).$$

Отже,

$$\frac{\partial}{\partial\lambda}\ln L \;=\; -n+\frac{1}{\lambda}\sum_{i=1}^{n}x_i.$$

Прирівнюючи похідну до нуля, маємо $-n+\frac{1}{\lambda}\sum_{i=1}^{n}x_i \;=\; 0,$ що дає

$$\hat{\lambda} \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i \;=\; \overline{x}.$$

Приклад 3.4

Показати, що відносна частота m/n появи події A при n випробуваннях за схемою Бернуллі є спроможною, незсуненою й ефективною оцінкою ймовірнісного параметра p.

Розв'язання

Спроможність оцінки випливає з теореми Бернуллі:

$$\lim_{n\to\infty}\Pr\left(\left|\frac{m}{n}-p\right|<\varepsilon\right) \;=\; 1, \qquad (\varepsilon>0).$$

Визначимо математичне сподівання відносної частоти. Будемо вважати $m=\mathrm{X}$, тобто частоту появи події A в n випробуваннях схеми Бернуллі розглядаємо як випадкову величину X, розподілену згідно з біноміальним законом

$$\mathrm{M}\left[\frac{\mathrm{X}}{n}\right] \;=\; \frac{1}{n}\mathrm{M}[\mathrm{X}] \;=\; \frac{1}{n}np \;=\; p.$$

Оскільки $\mathrm{M}[m/n] \;=\; p,$ то відносна частота m/n є незсуненою оцінкою ймовірності p.

Визначимо дисперсію відносної частоти

$$\mathrm{D}\left[\frac{m}{n}\right] \;=\; \mathrm{D}\left[\frac{\mathrm{X}}{n}\right] \;=\; \frac{1}{n^2}\mathrm{D}[\mathrm{X}] \;=\; \frac{1}{n^2}npq \;=\; \frac{pq}{n}.$$

Знайдемо нижню межу нерівності Рао-Крамера, враховуючи, що

$$f(x,\theta) \;=\; f(\mathrm{X}=x;\,p) \;=\; C_n^{\mathrm{X}}\,p^{\mathrm{X}}(1-p)^{n-\mathrm{X}}.$$

Тоді

$$\mathrm{M}\left[\frac{\partial^2\,\ln f(x;p)}{\partial p^2}\right] \;=\; \mathrm{M}\left[\frac{\partial^2}{\partial p^2}\Big(\ln C_n^{\mathrm{X}}+\mathrm{X}\ln p+(n-\mathrm{X})\ln(1-p)\Big)\right] \;=$$

$$=\; \mathrm{M}\left[-\frac{\mathrm{X}}{p^2}-\frac{n-\mathrm{X}}{(1-p)^2}\right] \;=-\frac{\mathrm{M}[\mathrm{X}]}{p^2}-\frac{n-\mathrm{M}[\mathrm{X}]}{(1-p)^2} \;=-\frac{np}{p^2}-\frac{n-np}{(1-p)^2} \;=\; -\frac{n}{pq}.$$

Отже, вважаючи, що для оцінки ймовірності зроблений один експеримент $(n=1)$, знайдемо нижню межу нерівності Рао-Крамера, яка дорівнює

$$-\frac{1}{n\,\mathrm{M}\left[\partial^2 f(x;p)/\partial p^2\right]} \;=\; \frac{pq}{n}.$$

Оскільки нижня межа нерівності Рао-Крамера збігається з дисперсією відносної частоти, то відносна частота є ефективною оцінкою ідеальної ймовірності p.

Приклад 3.5

Випадкова величина X має нормальний розподіл.

Потрібно за результатами спостережених значень $x_1, x_2, \ldots, x_n$ цієї випадкової величини оцінити параметри a і σ нормального закону.

Розв'язання

Густина розподілу ймовірностей цієї нормально розподіленої випадкової величини X

$$f(x;a,\sigma) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Отже, функція правдоподібності має вигляд

$$L \;=\; \frac{1}{(\sqrt{2\pi}\,\sigma)^n}\,\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right).$$

Запишемо логарифмічну функцію правдоподібності

$$\ln L \;=\; -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2.$$

Диференціюючи логарифмічну функцію правдоподібності за a і σ, маємо

$$\frac{\partial}{\partial a}\ln L \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a) \;=\; 0;$$

$$\frac{\partial}{\partial \sigma}\ln L \;=\; -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-a)^2 \;=\; 0,$$

звідки знаходиться

$$\hat{a} \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i \;=\; \overline{x}; \qquad\qquad \hat{\sigma}^2 \;=\; s^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Ці оцінки збіглися з оцінками, отриманими методом моментів.

Приклад 3.6

Двовимірна випадкова величина (X,Y) розподілена згідно з нормальним законом з густиною ймовірностей

$$f(x,y) = \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-\rho^2}}\;\times$$

$$\times\;\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} - \frac{2\rho\,(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}} + \frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right\}.$$

Потрібно за результатами спостережень двовимірної випадкової величини $(\mathrm{X}_i,\mathrm{Y}_i)$, де $i=1,2,\ldots,n$, оцінити параметри розподілу m_{X}, m_{Y}, σ_{X}, σ_{Y} і ρ.

Розв'язання

Застосовуючи метод моментів, маємо

$$\nu_{10} = \nu_{10}^* \Rightarrow \hat{m}_{\mathrm{X}} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

$$\nu_{01} = \nu_{01}^* \Rightarrow \hat{m}_{\mathrm{Y}} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$\mu_{20} = \nu_{20}^* \Rightarrow \hat{\sigma}_{\mathrm{X}}^2 = s_{\mathrm{X}}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2;$$

$$\mu_{02} = \nu_{02}^* \Rightarrow \hat{\sigma}_{\mathrm{Y}}^2 = s_{\mathrm{Y}}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2;$$

$$\frac{\mu_{11}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \Rightarrow \hat{\rho} = \frac{1}{n}\frac{1}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Приклад 3.7

Випадкова величина X розподілена згідно з показниковим законом з густиною ймовірностей $f(x;\theta) = \theta\exp(-\theta x)$ з параметром θ.

Потрібно за результатами спостережених значень $x_1, x_2, \ldots, x_n$ цієї випадкової величини знайти оцінку $\hat{\theta}$ параметра θ.

Розв'язання

Запишемо функцію правдоподібності

$$L = \theta\exp(-\theta x_1)\,\theta\exp(-\theta x_2)\ldots\,\theta\exp(-\theta x_n) = \theta^n \exp\left(-\theta\sum_{i=1}^{n} x_i\right).$$

Логарифмуючи, маємо

$$\ln L = n\ln\theta - \theta\sum_{i=1}^{n} x_i.$$

Диференціюючи за параметром θ, знаходимо

$$\frac{\partial \ln L}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

Зрівнюючи похідну до нуля, отримаємо

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \theta = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\overline{x}.$$

Приклад 3.8

Знайти мінімальний обсяг вибірки n, на основі якого можна було б оцінити математичне сподівання параметра деякої технічної операції з похибкою, яка не

перевищує 10, та надійністю $(1-\alpha)=0,95$, якщо передбачити, що цей параметр X є випадковою величиною, яка має нормальний розподіл $\mathrm{X}\to N(a;\,50)$.

Розв'язання

З умови прикладу випливає, що дисперсія випадкової величини X відома: $\sigma^2 = 50^2$. Скористаємося формулою, що зв'язує межову похибку ε оцінки математичного сподівання з середньою арифметичною:

$$\varepsilon = u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},$$

що дає

$$n = \frac{1}{\varepsilon^2}\,u^2_{\alpha/2}\sigma^2.$$

Користуючись таблицею функції Лапласа, за довірчою ймовірністю $1-\alpha=0,95$ знаходимо квантиль $u_{0,025}=1,96$. Отже,

$$n = \frac{1}{100}\,1,96^2\cdot 2500 = 96 \text{ вимірювань.}$$

Приклад 3.9 (Збільшення вибірки)

Отримана вибірка обсягом N. Для неї знайдено вибіркове середнє x^*_N і вибіркова дисперсія D^*_N. Здобуто ще одне значення x_{N+1}. Розглядається об'єднана вибірка обсягом $N+1$.

Потрібно виразити вибіркові оцінки x^*_{N+1} та D^*_{N+1} через оцінки x^*_N, D^*_N, які вже знайдено, та нове значення x_{N+1}, що здобуто.

Розв'язання

Для вибіркового середнього маємо

$$x^*_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1} x_i = \frac{1}{N+1}\left(\sum_{i=1}^{N} x_i + x_{N+1}\right) = \frac{1}{N+1}\left(Nx^*_N + x_{N+1}\right). \qquad (1)$$

Знайдене вибіркове середнє x^*_{N+1} використаємо для знаходження вибіркової дисперсії

$$D^*_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1}\left(x_i - x^*_{N+1}\right)^2. \qquad (2)$$

В круглих дужках під знаком підсумовування віднімемо та додамо вибіркове середнє x^*_N. Це дає

$$D^*_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1}\left[\left(x_i - x^*_N\right)^2 + 2\left(x_i - x^*_N\right)\left(x^*_N - x^*_{N+1}\right) + \left(x^*_N - x^*_{N+1}\right)^2\right]. \qquad (3)$$

Перший додаток в (3) дорівнює

$$\frac{1}{N+1}\left[\sum_{i=1}^{N}\left(x_i - x^*_N\right)^2 + \left(x_{N+1} - x^*_N\right)^2\right] = \frac{1}{N+1}\left[ND^*_N + \left(x_{N+1} - x^*_N\right)^2\right]. \qquad (4)$$

Аналогично перетворюючи інші додатки в (3), отримаємо

$$D^*_{N+1} = \frac{1}{N+1}\left[ND^*_N + N\left(x^*_{N+1} - x^*_N\right)^2 + \left(x_{N+1} - x^*_{N+1}\right)^2\right]. \quad (5)$$

Приклад 3.10 (Об'єднання вибірок)

Є дві вибірки обсягом N та M відповідно. Для них знайдено вибіркові середні x^*_N, x^*_M та вибіркові дисперсії D^*_N, D^*_M. Розглядається об'єднана вибірка обсягом $N+M$.

Потрібно виразити вибіркові оцінки x^*_{N+M} та D^*_{N+M} через знайдені оцінки x^*_N, D^*_N та x^*_M, D^*_M.

Розв'язання

Маємо

$$x^*_{N+M} = \frac{1}{N+M}\left(Nx^*_N + Mx^*_M\right).$$

Вибіркове середнє x^*_{N+M}, що знайдено, використаємо для знаходження вибіркової дисперсії

$$D^*_{N+M} = \frac{1}{N+M}\left[ND^*_N + N\left(x^*_{N+M} - x^*_N\right)^2 + MD^*_M + M\left(x^*_{N+M} - x^*_M\right)^2\right].$$

Приклад 3.11

Нехай здобуті наступні результати незалежних рівноточних вимірювань товщини металевої пластини:

$$x_1 = 2{,}015;\ \ x_2 = 2{,}020;\ \ x_3 = 2{,}025;\ \ x_4 = 2{,}020;\ \ x_5 = 2{,}015.$$

Потрібно:

1. Оцінити за допомогою довірчого інтервалу істинну товщину (математичне сподівання) пластини. Довірчу ймовірність $1-\alpha$ прийняти, щоб дорівнювала 0,95.

2. Знайти мінімальне число вимірювань, які треба виконати, щоб з надійністю $1-\alpha = 0{,}95$ можна було стверджувати, що межова похибка точкової оцінки істинної товщини металевої пластини не перевищує 0,95.

Розв'язання

Будемо вважати результати вимірювання спостереженими значеннями випадкової величини X, яка розподілена згідно з нормальним законом з невідомими параметрами a і σ. Знайдемо точкові оцінки цих параметрів.

Номер спостереження	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	2,015	–0,004	0,000016
2	2,020	+0,001	0,000001
3	2,025	+0,006	0,000036
4	2,020	+0,001	0,000001
5	2,015	–0,004	0,000016

Отже,

$$\hat{a} = \overline{x} = \frac{10{,}095}{5} = 2{,}019;$$

$$\hat{\sigma} \;=\; s_{\text{незм}} \;=\; \sqrt{\frac{0,00007}{4}} \;=\; 0,004183.$$

Користуючись таблицею розподілу Стьюдента (див. додаток) за довірчою ймовірністю $1-\alpha = 0,95$ і кількістю ступенів вільності $\nu = n-1 = 4$ знаходимо квантиль розподілу $t_{0,025;\,4} = 2,776$. Отже, межова похибка точкової оцінки

$$\varepsilon \;=\; \frac{s}{\sqrt{n}}\, t_{\alpha/2;\,n-1} \;=\; \frac{0,004183}{\sqrt{5}} \cdot 2,776 \;=\; 0,0052.$$

Довірчий інтервал дорівнює $[\overline{x}-\varepsilon;\ \overline{x}+\varepsilon] \;=\; [2,0138;\ 2,0242]$.

Для розрахунку мінімальної кількості вимірювань, необхідних для визначення істинної товщини пластини з похибкою, що не перевищує 0,003, застосуємо формулу $\varepsilon \;=\; st_{\alpha/2,\,n-1}/\sqrt{n}$.

Отже,

$$n \;\geq\; \varepsilon^{-2} t^2_{0,025;\,4} s^2 \;=\; \frac{2,776^2 \cdot 0,004183^2}{0,000009} \;\approx\; 15 \text{ вимірювань.}$$

Таким чином, для того щоб межова похибка ε істинного значення товщини пластини не перевищувала $\varepsilon = 0,003$, потрібно, крім 5 пророблених вимірювань, виконати ще $(n-5) = 15-5 = 10$ вимірювань.

Приклад 3.12

Нехай випадкова величина X підпорядковується нормальному розподілу з відомим середнім квадратичним відхиленням $\sigma = 2$.

Потрібно знайти довірчий інтервал для математичного сподівання a, якщо середнє арифметичне значення результатів $n = 16$ прямих рівноточних вимірювань $\overline{x} = 20,09$. Довірчу ймовірність $(1-\alpha)$ прийняти, щоб дорівнювала 0,90.

Розв'язання

Користуючись таблицями функції Лапласа, за заданою ймовірністю $(1-\alpha)$ знаходимо $u_{\alpha/2} = u_{0,05} = 1,64$. Знайдемо точність (межову похибку) оцінки

$$\varepsilon \;=\; 1,64\,\frac{2}{\sqrt{16}} \;=\; 0,82,$$

отже, шуканий довірчий інтервал складає $[20,09-0,82;\ 20,09+0,82]$, або $[19,27;\ 20,91]$.

Значення отриманого результату: якщо буде зроблено досить велике число вибірок даного обсягу, то в 90 % з них довірчі інтервали накриють математичне сподівання і тільки у 10 % випадків математичне сподівання, що оцінюється, може вийти за межі довірчих інтервалів.

Приклад 3.13 (Нерівність Крамера–Рао)

Нехай отримана незалежна вибірка $\{x_1, x_2, \ldots, x_n\}$. В цій вибірці кожна x_k $(k = 1, 2, \ldots, n)$ має густину розподілу $f(x;\theta)$, де θ – невідомий параметр. Нехай

також $\theta^* = \theta^*(x_1, x_2, \ldots, x_n) \equiv \theta^*(x)$ – незсунена оцінка параметра θ. Тоді $\mathrm{M}[\theta^*] = \theta$. Цю рівність можна записати у вигляді

$$\theta \;=\; \int_{\theta^*} \theta^* f_\theta(x)\, dx, \tag{1}$$

де

$$f_\theta(x) \;=\; \prod_{i=1}^{n} f(x_i;\theta)$$

– спільна густина розподілу ймовірностей вибірки $\{x_1, x_2, \ldots, x_n\}$.

Потрібно отримати вираз, що обмежує дисперсію $\mathrm{D}[\theta^*]$ знизу.

Розв'язання

Вираз у правій частині (1) диференцюється за θ так само, як і вираз

$$\int_{\theta^*} f_\theta(x)\, dx \;=\; 1. \tag{2}$$

Легко бачити, що наступний вираз є позитивним:

$$I(\theta) \;=\; \int \left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2 f_\theta(x)\, dx \;=\; \mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right]. \tag{3}$$

Оскільки $f_\theta(x)$ – густина розподілу ймовірностей, то $\int f_\theta(x)\, dx = 1$, де $x = (x_1, x_2, \ldots, x_n)$. Після диференціювання (1) і (3) за θ, знайдемо

$$1 \;=\; \int_{\theta^*} \theta^*(x)\, \frac{\partial f_\theta(x)}{\partial\theta}\, dx, \tag{4a}$$

$$0 \;=\; \int_{\theta^*} \frac{\partial f_\theta(x)}{\partial\theta}\, dx. \tag{4b}$$

Помножимо тепер другу рівність в (4) на θ і віднімемо з першої, тоді отримаємо

$$1 \;=\; \int_{\theta^*} [\theta^*(x) - \theta]\, \frac{\partial f_\theta(x)}{\partial\theta}\, dx.$$

Оскільки $f_\theta(x) > 0$, то

$$\partial f_\theta(x)/\partial\theta \;=\; f_\theta(x)\, \partial \ln f_\theta(x)/\partial\theta.$$

Підставляючи цей вираз в (4), який зведемо в квадрат, і використовуючи нерівність Коші-Буняковського, знаходимо

$$1 \;=\; \left(\int [\theta^*(x) - \theta]\, \frac{\partial f_\theta(x)}{\partial\theta}\, f_\theta(x)\, dx\right)^2 \;\le$$

$$\le\; \int [\theta^*(x) - \theta]^2\; f_\theta(x)\, dx \cdot \int \left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\; f_\theta(x)\, dx.$$

Інакше,

$$1 \;\le\; \mathrm{D}[\theta^*] \cdot \mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right]. \tag{5}$$

Виразимо тепер другий співмножник в (5) через функцію $I(\theta)$, визначену в (3). З (4) при $n = 1$ випливає, що

$$\mathrm{M}\left[\frac{\partial \ln f(x_1;\theta)}{\partial\theta}\right] \;=\; 0. \tag{6}$$

Оскільки

$$\ln f_\theta(x) \;=\; \sum_{i=1}^{n} \ln f(x_i;\theta),$$

то

$$\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2 \;=\; \sum_{i=1}^{n}\sum_{k=1}^{n} \frac{\partial \ln f_\theta(x_i;\theta)}{\partial\theta}\,\frac{\partial \ln f_\theta(x_k;\theta)}{\partial\theta}.$$

Звідси, враховуючи незалежність співмножників при $i \neq k$, знаходимо

$$\mathrm{M}\left[\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2\right] \;=\; n\,\mathrm{M}\left[\left(\frac{\partial \ln f(x_1;\theta)}{\partial\theta}\right)^2\right] \;=\; nI_1(\theta),$$

де $I_1(\theta)$ — кількість інформації про параметр θ, що міститься в одному спостереженні.

Отже, приходимо до виразу

$$\mathrm{D}[\theta^*] \;\ge\; \frac{1}{nI_1(\theta)} \tag{7}$$

– *нерівність Крамера-Рао.*

3.11. Задачі для розв'язання

Задача 3.1

Нехай $(x_1, x_2, \ldots, x_n)$ – вибірка, здобута з генеральної сукупності, що має рівномірний розподіл на інтервалі з фіксованими кордонами.

Знайти оцінки максимальної правдоподібності для кордонів a і b інтервалу.

Задача 3.2

З рівномірного розподілу витягнута вибірка $x_1, x_2, \ldots, x_n$ обсягом n. Для статистичної оцінки довжини інтервалу $(a;\, b)$ пропонується широта $\eta - \xi$ вибірки, де $\xi = \min\limits_i\{x_i\}$, $\eta = \max\limits_i\{x_i\}$.

Чи буде ця оцінка володіти властивостями ефективності та спроможності? Навести числовий приклад.

Задача 3.3

Побудувати функцію правдоподібності вибірки з n незалежних величин, розподілених згідно з геометричним законом з однаковим параметром.

Задача 3.4

Здобута вибірка обсягом n, яка витягнута з розподілу Лапласа з густиною $f(x) = \alpha^{-1}\exp(-|x-\mu|/\alpha); \qquad -\infty < x < \infty.$

Знайти оцінки максимальної правдоподібності для параметрів α і μ.

Задача 3.5

Здійснені дві серії з n_1 і n_2 незалежних випробувань, причому в першій серії подія A сталася m_1 разів, а у другій серії – m_2 разів.

Знайти оцінку максимальної правдоподібності для невідомої ймовірності p події A в кожному випробуванні (вважаючи цю ймовірність сталою).

Задача 3.6

З розподілу з густиною $f(x) = \exp(\alpha - x)$, де $x \geq \alpha$, витягнута вибірка $x_1, x_2, \ldots, x_n$ обсягом n. Для статистичної оцінки невідомого параметра α пропонується $\hat{\alpha} = \min\limits_i\{x_i\}$.

Чи буде ця оцінка володіти властивостями ефективності і спроможності? Навести числовий приклад.

Задача 3.7

Вибірка $x_1, x_2, \ldots, x_n$ обсягом n здобута з рівномірного розподілу на відрізку $[a, b]$. Відома довжина цього відрізка $l = b - a$, але невідома середина інтервалу $c = (a+b)/2$. Для статистичної оцінки середини інтервалу пропонується середнє арифметичне екстремальних значень $\overline{c} = (\xi + \eta)/2$ вибірки, де $\xi = \min\limits_i\{x_i\}$, $\eta = \max\limits_i\{x_i\}$.

Чи буде ця оцінка володіти властивостями ефективності і спроможності? Навести числовий приклад.

Задача 3.8

Знайти оцінку максимальної правдоподібності параметра σ за вибіркою обсягом n з нормально розподіленої генеральної сукупності з відомим математичним сподіванням m. Показати, що отримана оцінка максимальної правдоподібності є незсуненою.

Задача 3.9

За допомогою n різних експериментальних приладів отримані n вимірювань випадкової величини X.

У припущенні, що X має нормальний розподіл, а дисперсія i-го вимірювання відома і дорівнює σ_i^2 $(i = 1, 2, \ldots, n)$, знайти оцінку максимальної правдоподібності математичного сподівання випадкової величини X. Показати, що отримана оцінка є незсуненою, й обчислити її дисперсію.

Задача 3.10

Оцінка величини опору для великої партії резисторів, визначена за результатами вимірювань 100 випадково відібраних примірників, дорівнює $\overline{x} = 10$ *кОм*.

а) Вважаючи, що дисперсія вимірювань відома: $\sigma^2 = 1$ *кОм*^{2}, знайти ймовірність того, що для резисторів усієї партії величина опору лежить в межах $10 \pm 0,1$ *кОм*.

б) Скільки вимірювань треба зробити, щоб з імовірністю 0,95 стверджувати, що для всієї партії резисторів величина опору лежить в межах $10 \pm 0,1$ *кОм*?

Задача 3.11

Було проведено n випробувань $(n \geq m)$ до того моменту, коли подія A відбулася рівно m разів.

Знайти оцінку максимальної правдоподібності ймовірності p появи події A в одному випробуванні.

Задача 3.12

З розподілу з густиною $f(x) = \frac{1}{2}\exp(-|x|)/[1-\exp(-\beta)]$, де $-\beta < x < \beta$, витягнута вибірка $x_1, x_2, \ldots, x_n$ обсягом n.

Побудувати оцінку $\hat{\beta}$ для параметра β. Вивчити властивості запропонованої оцінки.

Задача 3.13

З генеральної сукупності, розподіленої за нормальним законом з параметрами $(m;1)$, витягнута вибірка обсягом $n = 3$ і побудований варіаційний ряд: $x_1 \leq x_2 \leq x_3$. Для невідомого математичного сподівання m розглядається наступна оцінка: $\theta = \gamma x_1 + (1-2\gamma)x_2 + \gamma x_3$, де γ – параметр.

Вивчити властивості оцінки, що пропонується. При якому значенні параметра γ ця оцінка буде володіти властивостями незсуненості та ефективності?

Задача 3.14

Відмова приладу сталася при k-му випробуванні.

Знайти оцінку максимальної правдоподібності ймовірності відмови p при одному випробуванні й обчислити її математичне сподівання.

Задача 3.15

Внаслідок проведення n незалежних експериментів в одних і тих же умовах випадкова подія A сталася x разів.

а) Показати, що відносна частота $f = x/n$ появи події A буде незсуненою і спроможною оцінкою події A: $\Pr(\mathrm{A}) = p$ в одному експерименті.

б) Визначити таке значення p, при якому дисперсія відносної частоти f буде максимальною.

3.12. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

Побудуйте програму, за допомогою якої дослідить залежність вибіркового середнього x_N^* від обсягу вибірки N. Результати оформіть графично.

Варіант 1

Рівномірний закон.

Вхідні дані для програми:

a – ліва межа можливих значень;

b – права межа можливих значень;

N – обсяг вибірки.

Результат роботи програми – масив, який містить значення вибіркового середнього x_N^* від обсягу вибірки N.

Варіант 2

Нормальний закон Гаусса.

Вхідні дані для програми:

m_{x} – математичне сподівання;

σ_{x}^2 – дисперсія;

N – обсяг вибірки.

Результат роботи програми – масив, який містить значення вибіркового середнього x_N^* від обсягу вибірки N.

Завдання 2

Побудуйте програму, за допомогою якої дослідить залежність вибіркового середнього D_N^* від обсягу вибірки N. Результати оформіть графично.

Варіант 1

Рівномірний закон.

Вхідні дані для програми:

a – ліва межа можливих значень;

b – права межа можливих значень;

N – обсяг вибірки.

Результат роботи програми – масив, який містить значення вибіркового середнього D_N^* від обсягу вибірки N.

Варіант 2

Нормальний закон Гаусса.

Вхідні дані для програми:

m_{x} – математичне сподівання;

σ_{x}^2 – дисперсія;

N – обсяг вибірки.

Результат роботи програми – масив, який містить значення вибіркового середнього D_N^* від обсягу вибірки N.

3.13. Завдання для перевірки

1. У чому полягає сутність задачі знаходження точкових оцінок невідомих параметрів розподілу?

2. Яка оцінка параметра називається спроможною? Чому бажано, щоб оцінка була спроможною?

3. Яка оцінка параметра називається незсуненою? Чому бажано, щоб оцінка була незсуненою?

4. Які оцінки параметрів називаються ефективними? Достатніми?

5. У чому полягає сутність методу моментів точкової оцінки параметрів?

6. У чому полягає сутність методу найбільшої правдоподібності точкової оцінки параметрів?

7. Якими властивостями володіють точкові оцінки параметрів, знайдені за методом моментів? За методом найбільшої правдоподібності?

8. Яка оцінка ймовірності набуття події p в схемі Бернуллі володіє властивостями спроможності, незсуненості і ефективності?

9. Яка оцінка математичного сподівання нормально розподіленої випадкової величини має властивості спроможності, незсуненості і ефективності?

10. Наведіть приклад оцінок математичного сподівання нормальної випадкової величини, що не мають вказаних властивостей.

11. Яка оцінка середнього квадратичного відхилення нормально розподіленої випадкової величини не володіє властивістю незсуненості? Не володіє властивістю ефективності? Володіє властивістю спроможності?

12. Що називається довірчим інтервалом? Довірчою ймовірністю?

13. Що називається граничною похибкою точкової оцінки параметра?

14. Що відбувається з довжиною довірчого інтервалу при збільшенні обсягу вибірки? Збільшенні довірчої ймовірності?

15. Чи є кінці довірчих інтервалів постійними величинами? Випадковими величинами?

16. Сформулюйте загальну схему побудови довірчих інтервалів.

17. Як будується довірчий інтервал для математичного сподівання випадкової величини X, розподіленої за нормальним законом?

18. Як будується довірчий інтервал для середнього квадратичного відхилення σ випадкової величини X, розподіленої згідно з нормальним законом?

4. Статистична перевірка параметричних гіпотез

4.1. Постановка задачі. Основні визначення

Раніше були розглянуті методи отримання оцінок невідомих параметрів розподілу. Знаходження точкових або інтервальних оцінок є, як правило, деякою попередньою стадією статистичних досліджень. Кінцева мета дослідження може, наприклад, полягати в порівняльній оцінці різних технологічних процесів за їх продуктивністю, точністю або економічністю, в порівнянні характеристик приладів, виробів і т. п. Задачі такого роду носять назву – *задачі порівняння*.

На математичній мові задачі порівняння формулюються як задачі статистичної перевірки гіпотез відносно параметрів законів розподілу, оскільки зміна параметрів характеризує відмінності технологічних процесів, конструкцій, приладів і т. п.

Припустимо, що для розв'язання задачі порівняння з генеральної сукупності здобута вибірка $(x_1, x_2, \ldots, x_n)$ обсягом n. Нехай, далі, експериментатор візуально – на вигляд гістограми або полігона відносних частот, або з будь-яких інших міркувань – висунув гіпотезу про закон розподілу випадкової величини (ВВ) X, що досліджується. Оцінивши параметри цього закону, він побудував теоретико-ймовірнісну модель розподілу ймовірностей цієї випадкової величини, яка, на його думку, відображає основні особливості статистичного ряду.

Визначення 1. Статистична гіпотеза називається *непараметричною*, якщо в ній сформульовані припущення відносно вигляду функції розподілу.

Подальша задача експериментатора полягає у перевірці висуненої гіпотези, тобто у з'ясуванні, наскільки добре підібрана ймовірнісна модель. Для перевірки цієї гіпотези використовуються різні статистичні критерії.

Припустимо, що за допомогою цих критеріїв експериментатор переконався, що модель "добра", дослідні дані не суперечать цій моделі, тобто він "завантажив" дослідні дані $(x_1, x_2, \ldots, x_n)$ в модель вдало.

Якщо тепер, наприклад, оцінювати параметр θ імовірнісної моделі $F(x; \theta)$ статистичного ряду спостережень за двома незалежними вибірками, взятими, скажемо, до введення деякого удосконалення в технологічний процес і після введення, то отримаємо дві оцінки $\hat{\theta}_1$ і $\hat{\theta}_2$, які внаслідок їх випадкового характеру не будуть дорівнювати між собою. Запитується (*висувається гіпотеза*), чи є ці оцінки оцінками одного і того ж параметра θ ймовірнісної моделі або у зв'язку з введенням удосконалення в технологічний процес цей параметр змінився? Задачі такого роду є типовими задачами порівняння.

Визначення 2. Статистична гіпотеза називається *параметричною*, якщо в ній сформульовані припущення відносно значень параметрів функції розподілу відомого вигляду. Найбільш точні і безпомилкові висновки відносно істинності такого роду

гіпотез можна було б зробити при дослідженні всієї генеральної сукупності. Однак на практиці в більшості випадків суцільне дослідження з ряду причин провести не можна. Наприклад, обсяг генеральної сукупності часто буває нескінченним, суцільне обстеження всієї генеральної сукупності вимагає великих матеріальних витрат. Крім того, таке обстеження може привести до псування предмету, а результати в момент закінчення дослідження можуть виявитися неактуальними.

Таким чином, висновки про істинність (помилковість) статистичних гіпотез відносно вигляду функції розподілу генеральної сукупності $F(x;\theta_i)$ або про значення параметрів функції розподілу відомого вигляду приймаються на основі експериментальної вибірки обсягом n.

Процес використання вибірки для перевірки істинності (помилковості) статистичних гіпотез називається *статистичним доказом істинності (помилковості) гіпотези, що висунена.*

Нарівні з гіпотезою, що висунена, розглядають одну або декілька альтернативних (конкуруючих) гіпотез. Якщо гіпотеза, що висунена, буде відкинута, то її місце займає альтернативна гіпотеза. З цієї точки зору статистичні гіпотези поділяються на *нульові* та *альтернативні.*

Визначення 3. *Нульовою гіпотезою* називають основну гіпотезу, що висунена, Нульову гіпотезу часто позначають символом H_0. Звичайно нульові гіпотези стверджують, що відмінність між величинами (це можуть бути параметри або функції розподілу), які порівнюються, відсутня, а відхилення, що спостерігаються, пояснюються лише випадковими коливаннями вибірки.

Визначення 4. *Альтернативною гіпотезою* називається гіпотеза, конкуруюча з нульовою гіпотезою в тому значенні, що якщо нульова гіпотеза H_0 відкидається, то їй на зміну приймається альтернативна. Альтернативну гіпотезу позначають символом H_a або H_1.

Наведемо приклади нульових і альтернативних статистичних гіпотез параметричного вигляду. Нехай за вибіркою $(x_1, x_2, \ldots, x_n)$ побудована нормальна модель з параметрами a і σ. Тоді відносно параметрів генеральної сукупності a і σ можна висунути наступні гіпотези:

Н у л ь о в і	А л ь т е р н а т и в н і
1. $\{H_0:\ a = a_0\}$.	1. $\{H_a:\ a \neq a_0\}$; $a < a_0$; $a > a_0$.
2. $\{H_0:\ \sigma = \sigma_0\}$.	2. $\{H_a:\ \sigma \neq \sigma_0\}$; $\sigma < \sigma_0$; $\sigma > \sigma_0$.
3. $H_0:\ \begin{cases} a = a_0; \\ \sigma = \sigma_0. \end{cases}$	3. $H_a:\ \begin{cases} a \neq a_0; \\ \sigma \neq \sigma_0. \end{cases}$

Змістовна сутність цих гіпотез може набувати різного вигляду залежно від конкретної задачі дослідження.

Розглянемо, наприклад, наступну задачу: чи буде новий (запропонований) спосіб виготовлення електроламп збільшувати термін їх служби в порівнянні з існуючим способом, при якому середній термін служби $a_0 = 4500$ *годин*? Випробування невеликої партії електроламп дали $x = 4800$ *годин*. Чи можна на основі цих даних вважати, що новий спосіб виробництва електроламп краще старого?

Можна висунути нульову і альтернативну гіпотези:

$\{H_0:\ a = a_0\}$ – при новому способі виробництва спостережений термін служби електроламп залишився колишнім;

$\{H_a:\ a > a_0\}$ – при новому способі виробництва спостережений термін служби електроламп збільшився.

Якщо експериментатор має підстави прийти до висновку, що нормальна модель вдало відображає в собі закономірності двох статистичних рядів $(x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)})$ і $(x_1^{(2)}, x_2^{(2)}, \ldots, x_{n_2}^{(2)})$, то відносно параметрів цієї моделі можуть бути розглянуті наступні нульові і відповідні до них альтернативні гіпотези:

Нульові	Альтернативні
1. $\{H_0:\ a_1 = a_2\}$.	1. $\{H_a:\ a_1 \neq a_2\}$; $a_1 < a_2$; $a_1 > a_2$.
2. $\{H_0:\ \sigma_1 = \sigma_2\}$.	2. $\{H_a:\ \sigma_1 \neq \sigma_2\}$; $\sigma_1 < \sigma_2$; $\sigma_1 > \sigma_2$.
3. $H_0:\ \begin{cases} a_1 = a_2, \\ \sigma_1 = \sigma_2. \end{cases}$	3. $H_a:\ \begin{cases} a_1 \neq a_2; \\ \sigma_1 \neq \sigma_2. \end{cases}$

Нульову гіпотезу

$$H_0:\ \begin{cases} a_1 = a_2; \\ \sigma_1 = \sigma_2 \end{cases}$$

можна сформулювати таким чином: *дві вибірки витягнуті з однієї і тієї ж генеральної сукупності*; альтернативну гіпотезу

$$H_a:\ \begin{cases} a_1 \neq a_2; \\ \sigma_1 \neq \sigma_2 \end{cases}$$

— *дві вибірки витягнуті з різних генеральних сукупностей.*

Статистичні параметричні гіпотези можуть містити одне або декілька припущень відносно параметрів розподілу функції.

Визначення 5. Параметрична гіпотеза називається *простою*, якщо містить тільки одне припущення відносно параметра.

Наприклад, якщо a – математичне сподівання нормально розподіленої випадкової величини, то гіпотеза $\{H_0:\ a = 0\}$ – проста.

Визначення 6. Параметрична гіпотеза називається *складною*, якщо вона складається з кінцевої або нескінченної кількості простих гіпотез. Наприклад, якщо a — математичне сподівання нормально розподіленої випадкової величини, то гіпотеза $\{H_a:\ a > 0,5\}$ є складною, оскільки вона складається з безлічі простих гіпотез вигляду $\{H_a:\ a = a_i\}$, де a_i – будь-яке задане число, більше 0,5.

4.2. Статистичний критерій значущості перевірки нульової гіпотези

Перевірка статистичних гіпотез здійснюється на основі даних вибірки. Для цього використовують спеціальним чином підібрану випадкову величину (вибіркову

статистику), яка є функцією спостережених значень, у якої точний або наближений розподіл вже відомий.

Цю вибіркову статистику позначають різними буквами залежно від закону її розподілу, наприклад u, якщо вона розподілена згідно з нормальним законом, F – якщо вона має розподіл Фішера, t – якщо вона має розподіл Стьюдента. У даному параграфі з метою спільності будемо позначати її через K.

Визначення 1. *Статистичним критерієм (тестом)* називають випадкову величину K, за допомогою якої приймаються рішення про прийняття або відкидання нульової гіпотези, що висунена.

Для перевірки нульових гіпотез за вибірковими даними обчислюють приватні значення вхідних в критерій K величин і, таким чином, одержують значення критерію, що спостерігається.

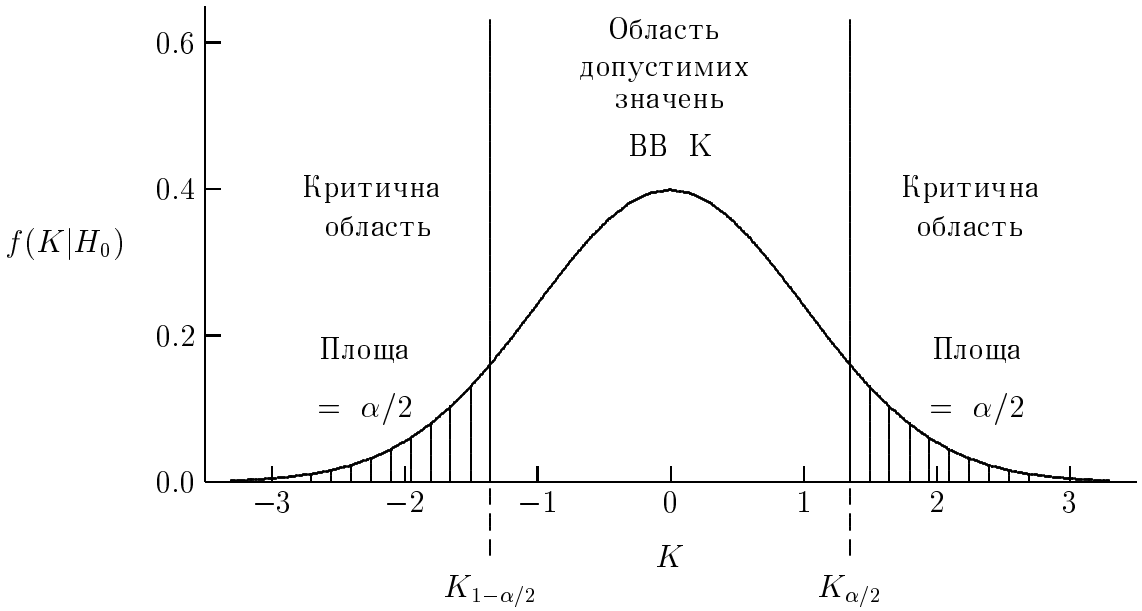


Рисунок 4.1 — Формування критичних областей і області допустимих значень випадкового критерію K при заданому рівні значущості α (двосторонній випадок)

Нехай для перевірки деякої нульової гіпотези H_0 відносно параметрів розподілу випадкової величини служить вибіркова статистика (критерій) K. Припустимо, що густина розподілу ймовірностей вибіркової статистики K за умовою справедливості нульової гіпотези H_0, що перевіряється, дорівнює $f(K|H_0)$, а математичне сподівання статистики K дорівнює K_0. Тоді ймовірність того, що випадкова величина K потрапить в довільний інтервал $[K_{1-\alpha/2}; K_{\alpha/2}]$, можна, користуючись формулою (рис. 4.1), знайти

$$\Pr\left(K_{1-\alpha/2} < \mathrm{K} < K_{\alpha/2}\right) = \int\limits_{K_{1-\alpha/2}}^{K_{\alpha/2}} f(K|H_0)\,dK. \qquad (4.1)$$

Задамо цю ймовірність, щоб дорівнювала $1-\alpha$, і обчислимо квантилі $K_{1-\alpha/2}$ і

$K_{\alpha/2}$ густини розподілу $f(K|H_0)$ за умов:

$$\Pr\left(\mathrm{K} \leq K_{1-\alpha/2}\right) = \int\limits_{-\infty}^{K_{1-\alpha/2}} f(K|H_0)\,dK = \frac{\alpha}{2}, \tag{4.2a}$$

$$\Pr\left(\mathrm{K} \geq K_{\alpha/2}\right) = \int\limits_{K_{\alpha/2}}^{\infty} f(K|H_0)\,dK = \frac{\alpha}{2}. \tag{4.2b}$$

Отже, ймовірність того, що ВВ K буде знаходитися всередині інтервалу $[K_{1-\alpha/2};\, K_{\alpha/2}]$, дорівнює $1-\alpha$. Ймовірність того, що ВВ K буде знаходитися поза цим інтервалом, дорівнює α.

Задамо ймовірність α настільки малою, щоб влучення ВВ K за межі інтервалу $[K_{1-\alpha/2};\, K_{\alpha/2}]$ можна було вважати малоймовірною подією. Тоді, виходячи з принципу практичної неможливості малоймовірних подій, можна вважати, що якщо нульова гіпотеза справедлива, то при її перевірці за допомогою критерію K за даними однієї вибірки значення критерію K, що спостерігається, повинне обов'язково влучити в інтервал $[K_{1-\alpha/2};\, K_{\alpha/2}]$.

Якщо ж значення критерію K, що спостерігається, потрапляє за межі інтервалу $[K_{1-\alpha/2};\, K_{\alpha/2}]$, то відбудеться малоймовірна, практично неможлива подія, тобто вважається, що з імовірністю $(1-\alpha)$ нульова гіпотеза, що перевіряється, не справедлива.

Тому область $[K_{1-\alpha/2};\, K_{\alpha/2}]$ називають *областю припустимих значень* ВВ K, при яких нульова гіпотеза не відхиляється, області $[-\infty;\, K_{1-\alpha/2}]$ і $[K_{\alpha/2};\, \infty]$ – *областями відхилення нульової гіпотези*, що перевіряється, або критичною областю критерію K. Якщо критичні області розташовуються ліворуч та праворуч від математичного сподівання випадкової величини K, так, як це зображено на рисунку 4.1, то критична область називається *двосторонньою*, а критерій K – *двостороннім критерієм значущості*.

У деяких випадках експериментатор має підстави бути твердо переконаним, що $\mathrm{K} > K_0$ або $\mathrm{K} < K_0$. У цьому випадку критичні області є односторонніми.

На рисунках 4.2 і 4.3 зображені правостороння і лівостороння критичні області відповідно.

Зауваження. Критична точка K_α (*квантиль K-розподілу*), що відокремлює області прийняття або відхилення нульової гіпотези, що перевіряється, для критерію з правосторонньою критичною областю знаходиться з умови

$$\Pr\left(\mathrm{K} \geq K_\alpha\right) = \int\limits_{K_\alpha}^{\infty} f(K|H_0)\,dK = \alpha. \tag{4.3}$$

Визначення 2. Перевірка гіпотези за допомогою статистичного критерію значущості є правилом відхилення нульової гіпотези, що полягає в розбитті області можливих значень ВВ K на дві підобласті, які не перетинаються, причому нульова гіпотеза відкидається, якщо значення критерію K, що спостерігається, належить критичній підобласті, і вважається узгодженою з дослідом, якщо значення критерію K, що спостерігається, не належить критичній підобласті.

Ця ж критична точка, яка відокремлює області прийняття або відхилення нульової гіпотези, що перевіряється, для критерію з лівосторонньою критичною областю знаходиться з умови

$$\Pr(\mathrm{K} \le K_{1-\alpha}) \;=\; \int\limits_{-\infty}^{K_{1-\alpha}} f(K|H_0)\,dK \;=\; \alpha. \tag{4.4}$$

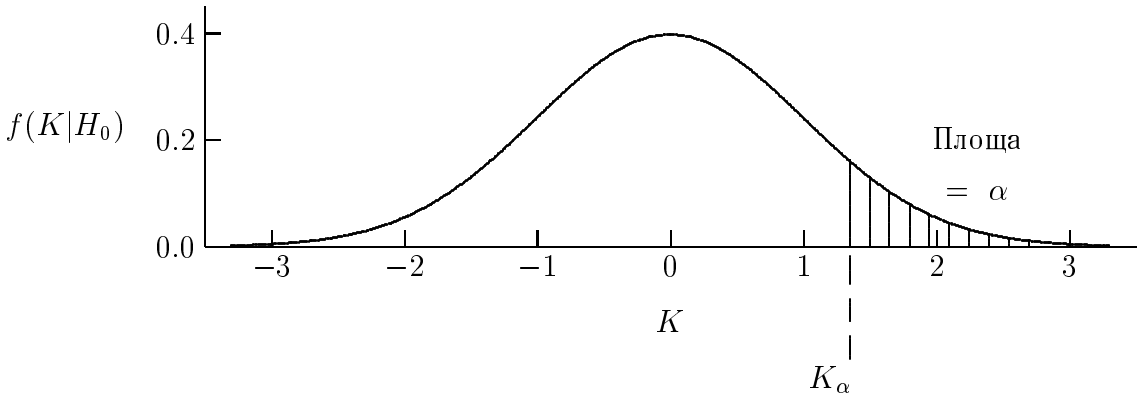


Рисунок 4.2 — Формування правосторонньої критичної області критерію K при заданому рівні значущості α

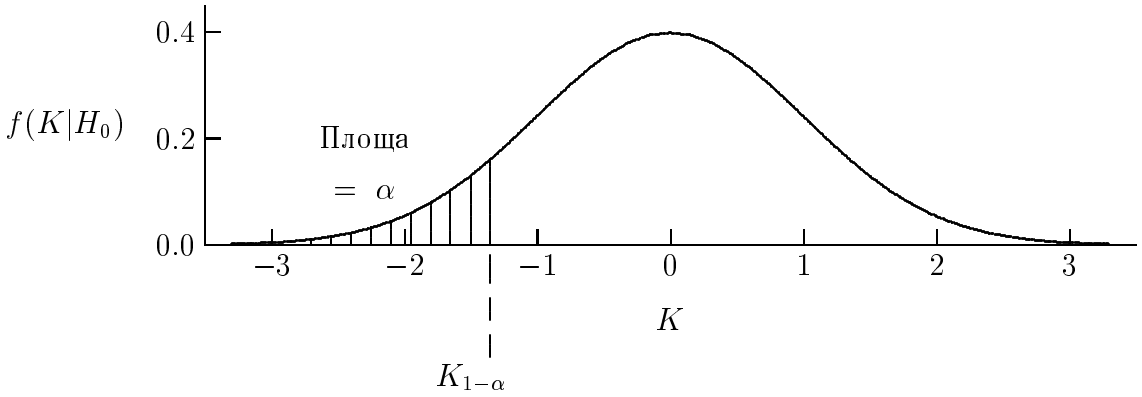


Рисунок 4.3 — Формування лівосторонньої критичної області критерію K при заданому рівні значущості α

Отже, можна сформулювати процедуру перевірки параметричних гіпотез за допомогою критеріїв значущості вказаним чином.

4.3. Помилки, що допускаються при перевірці статистичних гіпотез. Рівень значущості статистичного критерію

При перевірці статистичних гіпотез за вибірковими даними завжди існує ризик прийняття помилкового рішення. Це пояснюється тим, що обсяг вибірки кінцевий, і тому не можна точно визначити ні вигляд функції $F(x;\theta)$, ні значення її параметрів. Однак при багаторазовому застосуванні критеріїв теорія статистичної перевірки гіпотез дозволяє оцінити ймовірності прийняття помилкових рішень і, якщо ці ймовірності малі, то можна вважати, що даний статистичний критерій забезпечує малий ризик помилки.

Можливі помилки можуть бути подвійного роду.

Визначення 1. *Помилкою першого роду* називається помилка відхилення правильної нульової гіпотези H_0.

У попередньому параграфі було показано, що нульова гіпотеза відхиляється, якщо значення критерію К, що спостерігається, попаде в критичну область.

Визначення 2. *Рівнем значущості статистичного критерію* називається ймовірність α здійснення помилки першого роду.

Відхилення нульової гіпотези H_0 на рівні значущості $\alpha = 0,05$ означає, що, відхиляючи цю гіпотезу, ми або не помиляємося (тобто гіпотеза H_0 дійсно помилкова), або все-таки здійснюємо помилку першого роду, вважаючи правильну гіпотезу H_0 помилковою. В цьому останньому випадку ($\alpha = 0,05$) частота прийняття помилкового рішення дорівнює в середньому 5 на 100 випадків застосування даного статистичного критерію значущості.

Визначення 3. *Помилкою другого роду* називається помилка прийняття помилкової гіпотези H_0.

Ймовірність здійснення помилки другого роду прийнято позначати β. Якщо $f(K|H_a)$ – густина розподілу вибіркової статистики К за умови, що альтернативна гіпотеза H_a є вірною, то ймовірність здійснення помилки другого роду у випадку критерію з лівосторонньою критичною областю можна обчислити за формулою

$$\beta = \Pr(\mathrm{K} \geq K_{1-\alpha}) = \int\limits_{K_{1-\alpha}}^{\infty} f(K|H_a)\,dK. \tag{4.5}$$

Визначення 4. *Потужністю M критерію* К називається ймовірність $(1-\beta)$ нездійснення помилки другого роду (потужність критерію К – це ймовірність відхилення невірної гіпотези H_a), тобто $M = 1-\beta$.

На рис. 4.4 дана геометрична інтерпретація ймовірностей помилок першого роду, другого роду і потужності критерію К, що має лівосторонню критичну область.

З рис. 4.4 видно, що, пересуваючи квантиль $K_{1-\alpha}$ ліворуч (зменшуючи помилку першого роду), ми тим самим збільшуємо помилку другого роду. Можна показати, що ймовірність здійснення помилки другого роду β є функцією декількох змінних: кількості вимірювань n, рівня значущості α, характеру альтернативної гіпотези H_a, що застосовує критерій К, тобто

$$\beta = f(n, \alpha, H_a, K).$$

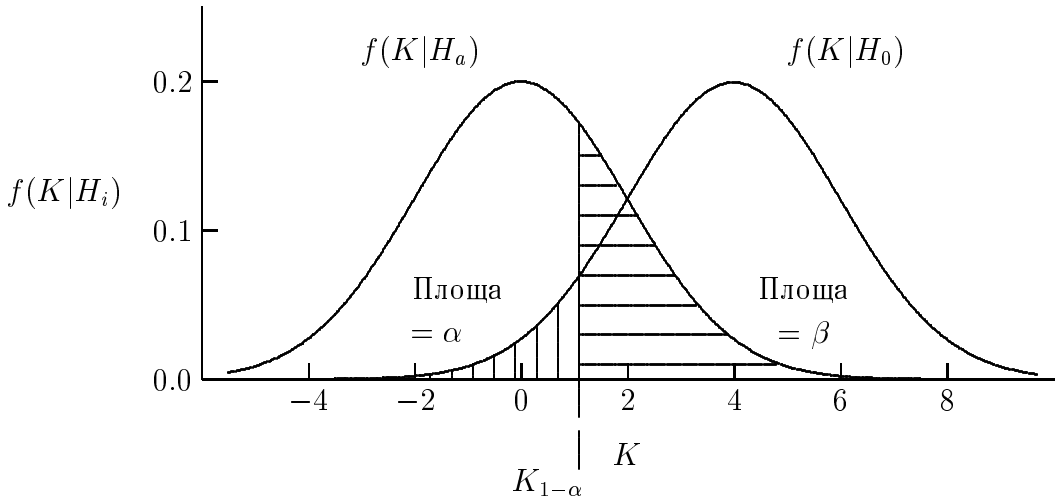


Рисунок 4.4 — Помилки першого і другого роду критерію K, що має лівосторонню критичну область

При цьому виконуються наступні граничні співвідношення:

$$\lim_{n\to\infty} \alpha = 0; \tag{4.6}$$

$$\lim_{n\to\infty} \beta = 0; \tag{4.7}$$

$$\lim_{\alpha\to 0} \beta = 1. \tag{4.8}$$

Рівності (4.6) і (4.7) вказують на те, що статистичні докази істинності гіпотез H_0 і H_a стають достовірними тільки при нескінченно великому обсязі вибірки. Крім того, з цих граничних співвідношень випливає, що єдиним способом одночасного зменшення ймовірностей помилок першого і другого роду є збільшення обсягу вибірки.

Рівність (4.8) говорить про те, що, зменшуючи ймовірність помилки першого роду до нуля, ми при фіксованому обсязі вибірки припускаємо необмежений ризик зробити помилку другого роду.

Як же слід вибирати рівень значущості α статистичних критеріїв? Відповідь на це питання залежить від втрат, що спричиняються помилками першого і другого роду. Наприклад, якщо здійснення помилки першого роду призведе до великих втрат в порівнянні з втратами, які мають місце при помилці другого роду, то потрібно прийняти по можливості менше значення α. Звичайно, не можна взяти $\alpha = 0$, оскільки з $\alpha = 0$ випливає $\beta = 1$, отже, в цьому випадку будуть прийматися всі нульові гіпотези, в тому числі і невірні.

У даному посібнику ми будемо розглядати тільки один вигляд статистичних критеріїв – статистичні критерії значущості, при застосуванні яких імовірність здійснення помилки першого роду (рівень значущості α) фіксується заздалегідь.

Статистичні критерії значущості – це односторонньо діючі критерії, тобто на основі їх застосування приймається із заздалегідь фіксованим ризиком тільки одне рішення: "*Відхилити нульову гіпотезу, що перевіряється*". Якщо ж для нульової гіпотези, що перевіряється, немає підстав відхилити її даним критерієм,

то стверджується: "*Результати вибірки не дають підстави для відхилення нульової гіпотези, що висунена*".

Таким чином, статистичні критерії значущості не дозволяють ухвалювати рішення: "*Нульова гіпотеза H_0 є правильною*", оскільки при застосуванні вказаних критеріїв імовірність помилки другого роду (ймовірність прийняття помилкової нульової гіпотези) залишається невідомою.

У більшості випадків на практиці достатньо застосування статистичних критеріїв значущості. Дійсно, практик-експериментатор, як правило, хоче перевірити, чи дають результати експерименту право відхилити нульову гіпотезу з тим, щоб прийняти замість неї деяку альтернативну гіпотезу, яку він відстоює (нова технологія виготовлення виробів і т.д.). Доказом істинності нульової гіпотези, наприклад, підтвердженням ефективності старої технології виготовлення виробу, він не займається.

Проілюструємо достатність застосування тільки статистичних критеріїв значущості на прикладі, що наведений на початку розділу. Припустимо, що запропоновано нову вдосконалену технологію виготовлення електроламп, яка припустимо збільшує термін їх служби. Для доказу збільшення терміну служби електроламп проведене випробування по визначенню тривалості їх роботи.

Висунемо нульову і альтернативну гіпотези:

$\{H_0 : \ a_1 = a_2\}$ – *при новій технології термін служби електроламп залишився колишнім*;

$\{H_a \ : \ a_1 > a_2\}$ – *внаслідок нової технології фактичний термін служби електроламп збільшився.*

Якщо статистичний критерій значущості відхилить нульову гіпотезу, що висунена, то підвищення терміну служби електроламп, виготовлених за новою технологією, визнається доведеним з відповідним малим ризиком здійснення помилки першого роду (рівнем значущості α).

Якщо ж на основі критерію значущості прийдемо до висновку: "*Немає підстав для відхилення нульової гіпотези*", то це означає, що статистичні дані, які повинні свідчити про ефективність нової технології, не можуть служити доказом підвищення терміну служби електроламп.

Цей приклад показує, що в більшості випадків для практичних додатків досить статистичних критеріїв значущості, які дозволяють тільки відхиляти нульову гіпотезу H_0, що висунена, з фіксованою малою ймовірністю помилки першого роду (рівнем значущості α).

При застосуванні статистичних критеріїв значущості вибір рівня значущості α в деякому сенсі довільний, оскільки в більшості випадків немає точної межі "дозволеної" ймовірності помилки першого роду α. Стало звичайним вибирати для α одне з стандартних значень: $\alpha = 0{,}005;\ 0{,}01;\ 0{,}05;\ 0{,}10$, хоча це не означає, що не можна вибирати $\alpha = 0,03$. Прийнята стандартизація має деяку перевагу, оскільки вона дозволяє скоротити обсяг таблиць критичних значень статистичних критеріїв.

Потрібно враховувати, що чим менше рівень значущості α, тим важче відхилити нульову гіпотезу. Тому не треба прагнути вибирати рівень значущості α дуже малим. При проведенні технічних досліджень найчастіше приймають $\alpha = 0,05$ і тільки у виключно важливих дослідженнях (наприклад, медичних) вважають $\alpha = 0,01$.

Зауваження:

1. Якщо суворі критерії для перевірки гіпотез відносно параметрів розподілу ВВ X, що має довільний закон розподілу, відмінний від нормального, відсутні, то далі статистичні критерії значущості, що описуються, можна вважати наближеними (відхилення від нормального закону призводить до збільшення рівня значущості).

2. Існують методи і критерії перевірки статистичних гіпотез, що враховують імовірності здійснення помилок як першого, так і другого роду, наприклад, критерії, побудовані за типом послідовного аналізу, критерій Неймана-Пірсона, критерії, що використовують теорію статистичних рішень.

3. Якщо в звичайному критерії значущості зафіксувати заздалегідь помилки першого і другого роду, то, користуючись кривими потужностей, можна визначити мінімальний обсяг вибірки, необхідний для розрізнення гіпотез H_0 і H_a з фіксованими ймовірностями здійснення помилок першого і другого роду.

Визначення найкращої критичної області для перевірки простих гіпотез.

На множині значень статистичного критерію можна вибрати скільки завгодно критичних областей V_k для заданого рівня значущості α, однак відповідні до них критерії будуть мати, взагалі кажучи, різні ймовірності помилок другого роду. *Найкращою критичною областю (НКО)* називають критичну область, яка при заданому рівні значущості α забезпечує мінімальну ймовірність помилки другого роду. Критерій, який використовує НКО, має максимальну потужність.

При перевірці простої гіпотези H_0 проти альтернативи H_a НКО визначається *лемою Неймана–Пірсона*: НКО критерію заданого рівня значущості α складається з точок вибіркового простору (вибірок обсягом n), для яких задовольняється нерівність

$$\frac{L(x_1, x_2, \ldots, x_n \,|\, H_0)}{L(x_1, x_2, \ldots, x_n \,|\, H_a)} \;<\; c_\alpha, \tag{4.9}$$

де c_α – стала, що залежить від заданого рівня значущості, $x_1, x_2, \ldots, x_n$ – елементи вибірки, а $L(x_1, x_2, \ldots, x_n|H_i)$ – функція правдоподібності, обчислена за умови, що вірна гіпотеза H_i.

Перейдемо до викладу конкретних статистичних критеріїв значущості для перевірки гіпотез про параметри нормального закону розподілу. Необхідно також зазначити, що, на жаль, розробка теорії перевірки статистичних гіпотез відносно параметрів законів розподілів, які відрізняються від нормального, пов'язана з великими труднощами.

4.4. Перевірка гіпотез про математичне сподівання випадкової величини, яка розподілена згідно з нормальним законом

Нехай для випадкової величини X відомо, що $X \to \mathcal{N}(a;\, \sigma)$. Потрібно перевірити нульову гіпотезу, згідно з якою математичне сподівання випадкової величини X дорівнює деякому гіпотетичному значенню a_0, тобто гіпотезу $\{H_0 : a = a_0\}$.

Іншими словами, треба встановити, значуще або незначуще розрізняються середнє арифметичне $\overline{x}$ і гіпотетичне математичне сподівання a генеральної сукупності.

Залежно від інформації, яку ми маємо про параметри генеральної сукупності, можна сформулювати дві основні моделі і побудувати для них відповідні критерії значущості.

Модель 1. Нехай генеральна сукупність має нормальний розподіл: $X \to \mathcal{N}(a;\, \sigma)$. Припустимо, що σ відоме. На основі випадкової вибірки $(x_1, x_2, \ldots, x_n)$ з цієї генеральної сукупності потрібно перевірити гіпотезу $\{H_0 : \; a = a_0\}$ проти альтернативної гіпотези $\{H_a : \; a \neq a_0\}$.

Критерій значущості для перевірки вказаної гіпотези грунтується на обчисленні вибіркової статистики

$$u \;=\; \frac{\overline{x} - a}{\sigma}\sqrt{n}. \tag{4.10}$$

Вище було показано, що якщо $X \to \mathcal{N}(a;\, \sigma)$, то $\overline{x} \to \mathcal{N}(a;\, \sigma/\sqrt{n})$. Отже, якщо нульова гіпотеза $\{H_0 : \; a = a_0\}$ справедлива, то нормована випадкова величина $u \to \mathcal{N}(0;\, 1)$.

Задамо рівень значущості даного критерію, щоб дорівнював α. За таблицею стандартизованого нормального розподілу (див. додаток) за заданим рівнем значущості знаходять критичні точки (квантилі) $u_{1-\alpha/2} = -u_{\alpha/2}$ та $u_{\alpha/2}$.

Множина значень u, що визначається нерівністю $|u| \;\geq\; u_{\alpha/2}$, є критичною областю критерію $u = (\overline{x} - a)\sqrt{n}/\sigma$ (рис. 4.5).

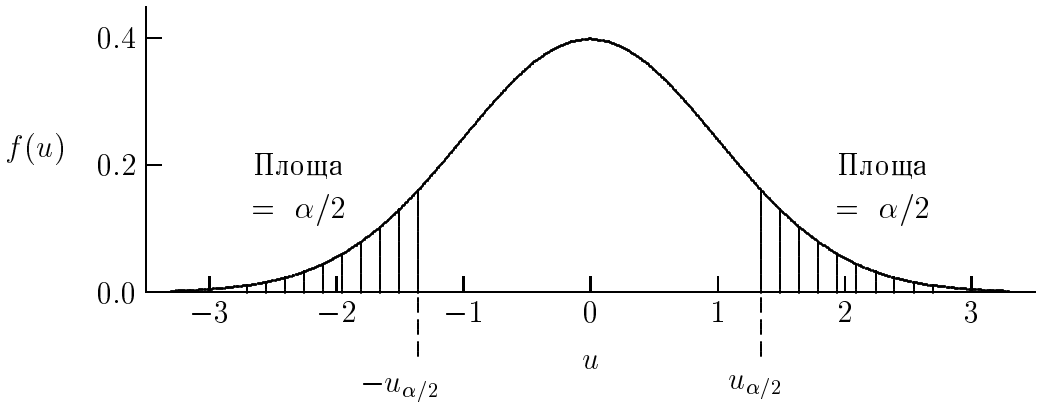


Рисунок 4.5 — Двостороння критична область критерію u

Отже, якщо обчислене за результатами вибірки з обсягом n значення критерію, що спостерігається, таке, що

$$|u_{\text{спос}}| \geq u_{\alpha/2}, \tag{4.11a}$$

то гіпотеза H_0 відхиляється.

Якщо

$$|u_{\text{спос}}| < u_{\alpha/2}, \tag{4.11b}$$

то немає підстав для відхилення нульової гіпотези.

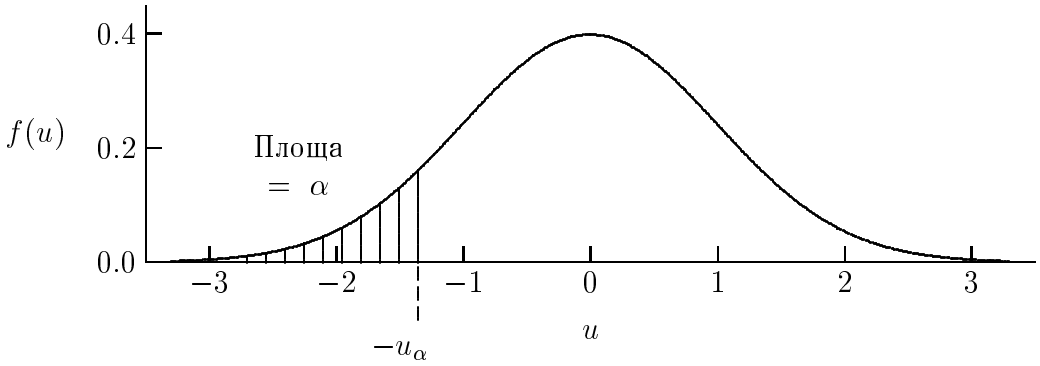


Рисунок 4.6 — Лівостороння критична область критерію u

Зауваження. Якщо гіпотеза H_a, що розглядається як альтернативна, має вигляд $\{H_a : a < a_0\}$, то використовують критерій u з лівосторонньою критичною областю (рис. 4.6).

У цьому (односторонньому) випадку критичні точки (квантилі) u_α стандартизованого нормального розподілу знаходяться з умови $\Pr(u \le -u_\alpha) = \alpha$. При такому виборі альтернативної гіпотези нульова гіпотеза відхиляється тільки в тому випадку, коли для значень вибіркової статистики, що спостерігається, виконується умова $u_{\text{спос}} \le -u_\alpha$.

Якщо альтернативна гіпотеза H_a, що розглядається, має вигляд $\{H_a : a > a_0\}$, то використовують критерій u із правосторонньою критичною областю.

У цьому випадку критичні точки (квантилі) u_α стандартизованого нормального розподілу знаходяться з умови $\Pr(u \ge u_\alpha) = \alpha$ (рис. 4.7). При такому вигляді альтернативної гіпотези нульова гіпотеза відхиляється тільки тоді, коли для значень вибіркової статистики, що спостерігається, виконується умова $u_{\text{спос}} \ge u_\alpha$.

Модель 2. Нехай генеральна сукупність має нормальний розподіл: $X \to \mathcal{N}(a, \sigma)$. При цьому параметри a і σ невідомі. За результатами випадкової вибірки обсягу n знайдені точкові оцінки параметрів

$$\hat{a} = \overline{x}; \qquad \hat{\sigma} = s_{\text{незс}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4.12}$$

Потрібно перевірити нульову гіпотезу $\{H_0 : a = a_0\}$ проти альтернативи $\{H_a : a \ne a_0\}$. Тоді критерій значущості для перевірки нульової гіпотези $\{H_0 : a = a_0\}$ грунтується на обчисленні тестової статистики

$$t = \frac{\overline{x} - a_0}{s}\sqrt{n} = \frac{\overline{x} - a_0}{s_{\text{незс}}}\sqrt{n-1}. \tag{4.13}$$

В курсі теорії ймовірностей показано, що якщо гіпотеза H_0 вірна, то випадкова величина t має розподіл Стьюдента з $\nu = n - 1$ ступенями вільності. За таблицею розподілу Стьюдента за заданим рівнем значущості α і кількістю ступенів вільності $\nu = n - 1$ знаходять критичні точки (квантилі) $t_{\alpha/2;\,n-1}$ розподілу Стьюдента.

Далі обчислюється значення критерію $t_{\text{спос}} = (\overline{x} - a_0)\sqrt{n}/s$, що спостерігається. Якщо

$$t_{\text{спос}} > t_{\alpha/2;\,n-1}, \tag{4.14a}$$

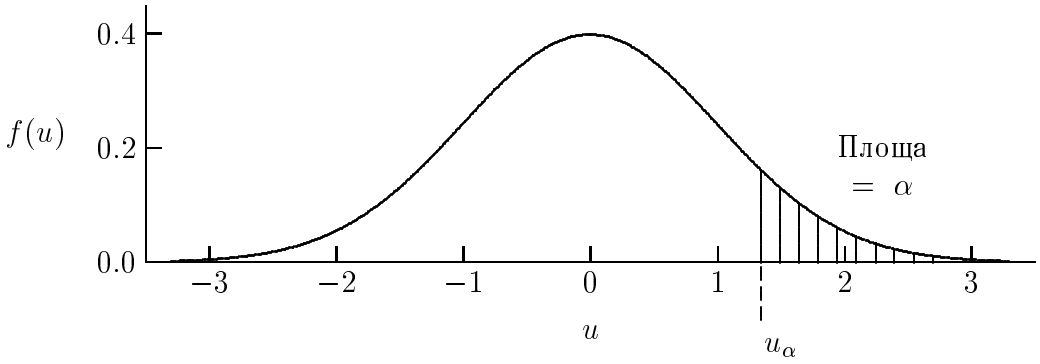


Рисунок 4.7 — Правостороння критична область критерію u

то нульова гіпотеза відхиляється на користь альтернативної гіпотези.

Якщо

$$t_{\text{спос}} \ < \ t_{\alpha/2;\,n-1}, \tag{4.14b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження 1. Вище наведене в моделі 1 зауваження залишається справедливим і для вибіркової статистики t. Це означає, що якщо альтернативна гіпотеза, що розглядається, має вигляд $\{H_a : a > a_0\}$, то застосовується правосторонній t-критерій, тобто критичні точки (квантилі) $t_{\alpha/2;\,n-1}$ знаходяться з умови $\Pr(t > t_{\alpha/2;\,n-1}) = \alpha$. Якщо ж альтернативна гіпотеза має вигляд $\{H_a : a < a_0\}$, то застосовується лівосторонній t–критерій, критичні точки якого (квантилі) $-t_{\alpha;\,n-1}$ знаходяться з умови $\Pr(t \le -t_{\alpha;\,n-1}) = \alpha$.

Зауваження 2. Якщо обсяг вибірки n досить великий ($n > 50$), то для перевірки гіпотези $\{H_0 : a = a_0\}$ можна застосовувати критерій u (модель 1), в якому потрібно використати

$$\sigma \ = \ s \ = \ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4.15}$$

У курсах математичної статистики показано, що критерії, засновані на обчисленні тестових статистик за формулами (4.10) і (4.13), є найкращими, оскільки вони забезпечують максимальну потужність. Нагадаємо, що потужність критерію залежить від вигляду альтернативної гіпотези, прийнятого рівня значущості α і обсягу вибірки n.

Наприклад, якщо бажано перевірити нульову гіпотезу $\{H_0 : a = a_0\}$ при простій альтернативній гіпотезі $\{H_a : a = a_1\}$, причому $a_1 \ne a_0$, то потужність критерію M виражається формулою

$$M \ = \ 1 - \beta \ = \ 1 + \frac{1}{2}\,\Phi\left(-u_\alpha + \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right) - \frac{1}{2}\,\Phi\left(-u_\alpha - \frac{a_1 - a_0}{\sigma/\sqrt{n}}\right). \tag{4.16}$$

Залежність потужності критерію від вигляду альтернативної гіпотези і рівня значущості використовується при плануванні експерименту для розрахунку обсягу вибірки, необхідного для отримання заданої потужності критерію.

4.5. Перевірка гіпотез рівності математичних сподівань двох нормальних випадкових величин

На практиці при обробці статистичних даних нерідко виникає потреба в розв'язанні задач "порівняння".

Наприклад, часто доводиться порівнювати новий і старий технологічні методи виготовлення деяких виробів, успішність в двох групах, що застосовують різні методи навчання, продуктивність праці на двох заводах і таке інше. Задачі такого типу можна розв'язувати, побудувавши теоретико-ймовірнісну модель (див. постановку задачі на початку даного розділу) генеральної сукупності $F(x;\theta)$. У більшості випадків закони розподілу цих сукупностей передбачають нормальними, а зміни в "технологіях" позначаються на зміні математичних сподівань нормальної сукупності, що моделюється.

Таким чином, більшість задач порівняння зводиться до перевірки гіпотез відносно математичних сподівань двох випадкових величин, розподілених згідно з нормальним законом.

Залежно від того, яка є в розпорядженні експериментатора інформація відносно параметрів нормальних сукупностей, що досліджуються, можна сформулювати дві основні моделі, в кожній з яких застосовується певний критерій значущості.

Модель 1. Нехай досліджуються дві випадкові величини X і Y, кожна з яких підкоряється нормальному закону: $\mathrm{X} \to \mathcal{N}(a_1;\, \sigma_1)$ і $\mathrm{Y} \to \mathcal{N}(a_2;\, \sigma_2)$. Припустимо, що середньоквадратичні відхилення σ_1 і σ_2 відомі, а значення a_1 і a_2 невідомі.

Потрібно на основі двох незалежних вибірок обсягом n_1 і n_2, відповідно вилучених з генеральних сукупностей, що досліджуються, виконати перевірку нульової гіпотези $\{H_0 :\ a_1 = a_2\}$ проти альтернативної гіпотези $\{H_a :\ a_1 \neq a_2\}$.

Обчислимо за вибірками середньоарифметичні $\overline{x}$ і $\overline{y}$. Відомо, що якщо $\mathrm{X} \to \mathcal{N}(a_1;\, \sigma_1)$ та $\mathrm{Y} \to \mathcal{N}(a_2;\, \sigma_2)$, то $\overline{x} \to \mathcal{N}(a_1;\, \sigma_1/\sqrt{n_1})$ і $\overline{y} \to \mathcal{N}(a_2;\, \sigma_2/\sqrt{n_2})$. Оскільки вибірки незалежні, то незалежні і середні арифметичні $\overline{x}$ і $\overline{y}$, а отже,

$$\mathrm{D}[\overline{x} - \overline{y}] \;=\; \mathrm{D}[\overline{x}] + \mathrm{D}[\overline{y}] \;=\; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{4.17}$$

Якщо гіпотеза $\{H_0 :\ a_1 = a_2\}$, що перевіряється, справедлива, то

$$\mathrm{M}[\overline{x} - \overline{y}] \;=\; \mathrm{M}[\overline{x}] - \mathrm{M}[\overline{y}] \;=\; 0. \tag{4.18}$$

Отже, нормована різниця (вибіркова статистика)

$$u \;=\; \frac{\overline{x} - \overline{y}}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \tag{4.19}$$

має стандартизований нормальний розподіл $u \to \mathcal{N}(0;\, 1)$.

Ця вибіркова статистика часто застосовується як статистичний критерій значущості для перевірки нульової гіпотези $\{H_0 :\ a_1 = a_2\}$.

Для перевірки цієї нульової гіпотези необхідно задати рівень значущості α. Потім за таблицею стандартизованого нормального розподілу за заданим рівнем

значущості α необхідно знайти критичне значення (квантиль) $u_{\alpha/2}$, що задовольняє умові $\Pr(|u| \geq u_{\alpha/2})$, яка визначає двосторонню критичну область u-критерію (див. додаток).

Таким чином, якщо обчислити згідно з (4.19) значення критерію $u_{\text{спос}}$, що спостерігається, і при цьому виявиться, що

$$|u_{\text{спос}}| \;\geq\; u_{\alpha/2}, \tag{4.20a}$$

то нульова гіпотеза відхиляється на користь альтернативної.

Якщо ж

$$|u_{\text{спос}}| \;<\; u_{\alpha/2}, \tag{4.20b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження. У тому випадку, коли альтернативна гіпотеза, що розглядається, має вигляд $\{H_a : \; a_1 < a_2\}$, застосовується лівосторонній u–критерій. При його застосуванні за таблицею стандартизованого нормального розподілу знаходиться таке критичне значення $-u_\alpha$, щоб виконувалась умова $\Pr(u < -u_\alpha) = \alpha$. У випадку ж, коли альтернативна гіпотеза має вигляд $\{H_a : \; a_1 > a_2\}$, застосовується правосторонній u–критерій. У цьому випадку за таблицею стандартизованого нормального розподілу знаходиться таке критичне значення u_α, щоб виконувалась умова $\Pr(u \geq u_\alpha) = \alpha$.

Модель 2. Нехай досліджуються дві випадкові величини X та Y, кожна з яких підкоряється нормальному закону $\text{X} \to \mathcal{N}(a_1;\, \sigma_1)$ та $\text{Y} \to \mathcal{N}(a_2;\, \sigma_2)$, причому середні квадратичні відхилення σ_1 та σ_2 хоч і невідомі, але передбачається, що $\sigma_1 = \sigma_2$, а параметри a_1 і a_2 невідомі.

Потрібно на основі двох незалежних вибірок обсягом n_1 і n_2 кожна ($n_1 \geq 30$ та $n_2 \geq 30$), витягнутих з генеральних нормальних сукупностей, перевірити нульову гіпотезу $\{H_0 : \; a_1 = a_2\}$ проти альтернативної гіпотези $\{H_a : \; a_1 \neq a_2\}$.

Критерій значущості для перевірки даної нульової гіпотези грунтується на обчисленні вибіркової статистики

$$t \;=\; (\overline{x} - \overline{y}) \left[\frac{n_1 \hat{\sigma}_1^2 + n_2 \hat{\sigma}_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1/2}, \tag{4.21}$$

де

$$\overline{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \qquad \overline{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i \tag{4.22a}$$

— середні арифметичні;

$$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2, \qquad \hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_i - \overline{y})^2 \tag{4.22b}$$

— оцінки середніх квадратичних відхилень першої і другої сукупностей відповідно.

Можна довести, що якщо нульова гіпотеза H_0 справедлива, то вибіркова статистика t має розподіл Стьюдента з $\nu = n_1 + n_2 - 2$ ступенями вільності. Для перевірки нульової гіпотези необхідно обчислити згідно з (4.21) значення t-критерію $t_{\text{спос}}$, що спостерігається, і задати рівень значущості α. За таблицею квантилів

розподілу Стьюдента за заданою ймовірністю α і кількістю ступенів вільності ν знайти критичні точки (квантилі) $t_{\alpha/2;\,n_1+n_2-2}$.

Якщо при цьому виявиться

$$|t_{\text{спос}}| \;\geq\; t_{\alpha/2;\,n_1+n_2-2}, \tag{4.23a}$$

то нульова гіпотеза відхиляється на користь альтернативної.

Якщо ж

$$|t_{\text{спос}}| \;<\; t_{\alpha/2;\,n_1+n_2-2}, \tag{4.23b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження 1. Якщо альтернативна гіпотеза, що розглядається, має вигляд $\{H_a : \; a_1 < a_2\}$, то, як і в моделі 1, застосовується лівосторонній t–критерій. Якщо ж альтернативна гіпотеза має вигляд $\{H_a : a_1 > a_2\}$, то застосовується правосторонній t-критерій. Умови знаходження критичних точок $t_{\alpha;\,n_1+n_2-2}$ цих критеріїв залишаються тими ж, що і в моделі 1.

Зауваження 2. Іноді в практиці трапляється, що результати двох вибірок розглядають як вимірювання однієї і тієї ж випадкової величини до і після проведення деякої технологічної операції. Нехай є впорядковані пари таких чисел $(x_i,\, y_i)$, i – номер вимірювання.

Будемо розглядати різниці $z_i = x_i - y_i$ як компоненти однієї вибірки і обчислимо їх середнє арифметичне

$$\overline{z} = \frac{1}{n}\sum\nolimits_{i=1}^{n} z_i$$

і середнє квадратичне

$$s_z^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n}(z_i - \overline{z})^2$$

відхилення.

Будемо замість гіпотези $\{H_0 : \; a_1 = a_2\}$ перевіряти еквівалентну до неї гіпотезу $\{H_0 : \; a_1 - a_2 = 0\}$. Перевірка цієї гіпотези проводиться за допомогою критерію $t = (\overline{z}/s_z)\sqrt{n-1}$ методами, що описані в попередньому параграфі.

Зауваження 3. Якщо обсяги вибірок n_1 і n_2 досить великі $(n_1 \geq 30$ та $n_2 \geq 30)$, то замість t–критерію (модель 2) можна застосовувати u-критерій (модель 1).

В цьому випадку при обчисленні вибіркової статистики u за формулою (4.13) в ній потрібно замінити середні квадратичні відхилення σ_1 і σ_2 їх точковими оцінками s_1 і s_2.

Критерії, що викладені в даному параграфі, є найбільш потужними.

4.6. Перевірка гіпотез про дисперсію нормальної випадкової величини

У практичних задачах перевірка гіпотез про дисперсію грає велику роль, оскільки саме дисперсія характеризує такі важливі технологічні і конструкторські

показники, як точність роботи машин, похибки свідчення вимірювальних приладів, ритмічність виробництва, стійкість роботи автоматичних ліній і таке інше.

Критерій значущості, що застосовується для перевірки рівності невідомої дисперсії генеральної нормальної сукупності деякому гіпотетичному (передбачуваному) значенню σ_0^2, засновується на ряді початкових імовірнісних припущень відносно генеральної сукупності, що досліджується. Будемо розглядати ці припущення як деяку ймовірнісну модель.

Модель. Нехай випадкова величина X $\to\mathcal{N}(a;\,\sigma)$, причому параметри a і σ невідомі. З генеральної сукупності $\mathcal{N}(a;\,\sigma)$ витягнута випадкова вибірка обсягом n і знайдені точкові оцінки параметрів нормального закону:

$$\hat{a} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{4.24}$$

Потрібно на основі цієї отриманої інформації перевірити нульову гіпотезу $\{H_0:\ \sigma^2 = \sigma_0^2\}$ проти альтернативної гіпотези $\{H_a:\ \sigma^2 \neq \sigma_0^2\}$.

Критерій значущості для перевірки даної нульової гіпотези засновується на обчисленні вибіркової статистики

$$\chi^2 \;=\; \frac{n\,s^2}{\sigma_0^2} \;=\; \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{4.25}$$

Якщо передбачити, що нульова гіпотеза справедлива, то можна довести, що вибіркова статистика має χ^2 (хі–квадрат) розподіл з $\nu = n-1$ ступенями вільності. Задамо рівень значущості даного критерію, щоб дорівнював α. Тоді за таблицею χ^2-розподілу за рівнем значущості α і кількістю ступенів вільності $\nu = n-1$ можна знайти критичні точки (квантилі) χ^2-розподілу $\chi^2_{1-\alpha/2;\,\nu}$ і $\chi^2_{\alpha/2;\,\nu}$ (рис. 4.8). Тепер необхідно обчислити спостережене значення критерію

$$\chi^2_{\text{спос}} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{4.26}$$

Якщо при цьому виявиться, що

$$\chi^2_{\text{спос}} \geq \chi^2_{\alpha/2;\,\nu} \qquad \text{або} \qquad \chi^2_{\text{спос}} \leq \chi^2_{1-\alpha/2;\,\nu}, \tag{4.27a}$$

то нульова гіпотеза відхиляється на користь альтернативної.

Якщо ж виявиться, що

$$\chi^2_{1-\alpha/2;\,\nu} < \chi^2_{\text{спос}} < \chi^2_{\alpha/2;\,\nu}, \tag{4.27b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження 1. Якщо використана при аналізі альтернативна гіпотеза має вигляд $\{H_a:\ \sigma^2 < \sigma_0^2\}$, то застосовується критерій χ^2 з лівосторонньою критичною областю. У цьому випадку критичні точки (квантилі) $\chi^2_{1-\alpha/2;\,\nu}$ знаходяться з умови $\Pr\left(\chi^2 \leq \chi^2_{1-\alpha;\,\nu}\right)$, а нульова гіпотеза відхиляється, якщо $\chi^2_{\text{спос}} \leq \chi^2_{1-\alpha;\,\nu}$.

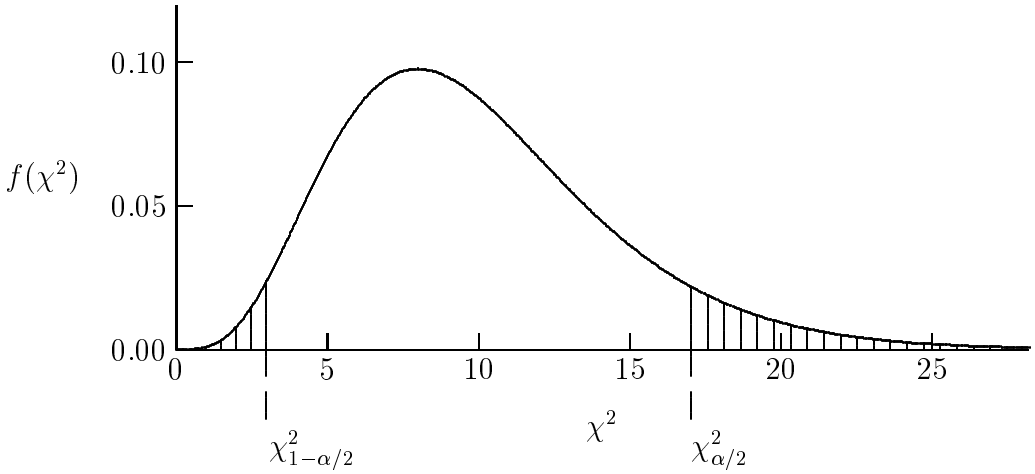


Рисунок 4.8 — Двостороння критична область критерію χ^2

Якщо ж альтернативна гіпотеза, що розглядається, має вигляд $\{H_a : \sigma^2 > \sigma_0^2\}$, то застосовується критерій χ^2 Пірсона, який відповідає правосторонній критичній області. У цьому випадку критичні точки $\chi^2_{\alpha;\nu}$ знаходяться з умови $\Pr\left(\chi^2 \geq \chi^2_{\alpha;\nu}\right)$, а нульова гіпотеза відхиляється, якщо $\chi^2_{\text{спос}} \geq \chi^2_{\alpha;\nu}$.

Зауваження 2. Якщо для кількості ступенів вільності ν справедлива нерівність $\nu = n - 1 > 30$, перевірку нульової гіпотези, що використана, можна проводити, обчислюючи вибіркову статистику $u = \sqrt{2\chi^2_{\text{спос}}} - \sqrt{2\nu}$. Можна показати, що ця статистика має асимптотично нормальний стандартизований розподіл $u \to \mathcal{N}(0;\, 1)$.

Критерій, заснований на обчисленні тестової статистики, що визначається формулою (4.25), є найбільш потужним. Дослідження потужності даного критерію можна знайти у підручниках з математичної статистики.

4.7. Перевірка гіпотез про дисперсії двох нормальних випадкових величин

У випадку, коли статистичні дослідження деякої кількісної ознаки проводяться в двох генеральних сукупностях, часто з'являється потреба в перевірці гіпотези про рівність міри розсіювання ознаки, що досліджується в цих сукупностях. Нехай є дві вибірки, дисперсії яких відповідно дорівнюють s_1^2 і s_2^2. Чи можна вважати при наявності деяких відмінностей між величинами s_1^2 і s_2^2, що дані вибірки належать одній і тій же самій генеральній сукупності?

Можна сформулювати досить поширену типову задачу: проведено дві серії дослідів, з яких один дослід проводиться з урахуванням чинника А, а інший – без урахування. Чи надає чинник А вплив на розсіювання ознаки, що досліджується?

Для відповіді на поставлені питання необхідно зробити перевірку нульової гіпотези $\{H_0 : \sigma_1^2 = \sigma_2^2\}$. Критерій значущості, що грунтується для перевірки рівності невідомих дисперсій нормальних сукупностей, тобто для перевірки вказаної

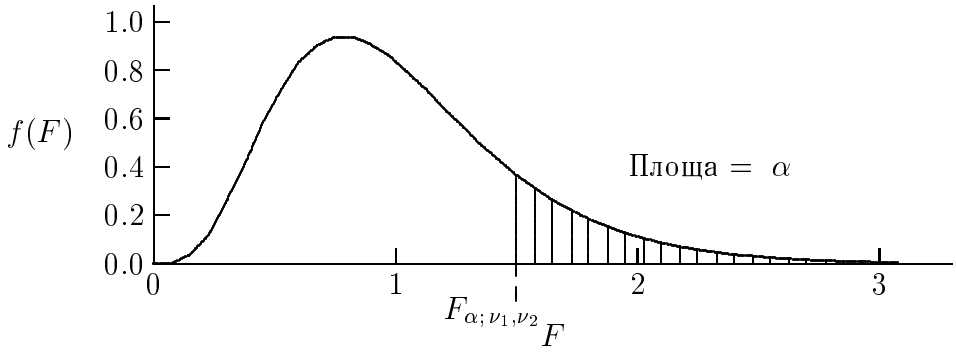


Рисунок 4.9 — Формування критичної області критерію F при заданому рівні значущості α

нульової гіпотези, засновується на ряді початкових імовірнісних припущень відносно цих сукупностей. Будемо розглядати ці припущення як деяку ймовірнісну модель.

Модель. Нехай ВВ $\mathrm{X}_1 \to \mathcal{N}(a_1;\, \sigma_1)$ і $\mathrm{X}_2 \to \mathcal{N}(a_2;\, \sigma_2)$. Параметри нормальних законів розподілу a_1 і a_2 — невідомі. З цих двох генеральних нормальних сукупностей витягнуті вибірки обсягом n_1 і n_2. На основі цих вибірок обчислені точкові оцінки параметрів нормального закону:

$$\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}, \qquad s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(x_{i1}-\overline{x}_1\right)^2,$$

$$\overline{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}, \qquad s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(x_{i2}-\overline{x}_2\right)^2.$$

Потрібно на основі отриманої в дослідженні інформації перевірити нульову гіпотезу $\{H_0:\ \sigma_1^2 = \sigma_2^2\}$ проти альтернативної гіпотези $\{H_a:\ \sigma_1^2 > \sigma_2^2\}$.

Критерій значущості для перевірки даної нульової гіпотези засновується на обчисленні вибіркової статистики

$$F \;=\; s_1^2/s_2^2, \tag{4.28}$$

де s_1^2 та s_2^2 – відповідно найбільша і найменша дисперсії. Якщо припустити, що нульова гіпотеза вірна, то вибіркова статистика F має розподіл Фішера з $\nu_1 = n_1 - 1$ та $\nu_2 = n_2 - 1$ ступенями вільності. Задамо рівень значущості даного критерію, щоб дорівнював α. Тоді за таблицею квантилів F-розподілу за рівнем значущості α і кількістю ступенів вільності $\nu_1 = n_1 - 1$ та $\nu_2 = n_2 - 1$ можна знайти критичну точку (квантиль) $F_{\alpha;\nu_1;\nu_2}$, що задовольняє умові (рис. 4.9)

$$\Pr\left(F > F_{\alpha;\nu_1;\nu_2}\right) \;=\; \alpha. \tag{4.29}$$

Обчислимо тепер згідно з (4.28) значення критерію $F_{\text{спос}}$.

Якщо при цьому виявиться, що

$$F_{\text{спос}} \;\geq\; F_{\alpha;\nu_1;\nu_2}, \tag{4.30a}$$

то нульова гіпотеза відхиляється на користь альтернативної.

Якщо ж

$$F_{\text{спос}} \;<\; F_{\alpha;\,\nu_1;\,\nu_2}, \tag{4.30b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження 1. Якщо альтернативна гіпотеза, що розглядається, має вигляд $\{H_a : \sigma_1^2 \neq \sigma_2^2\}$, тоді потрібно застосовувати F-критерій з двосторонньою критичною областю. У цьому випадку критичні точки $F_{\alpha/2;\,\nu_1;\,\nu_2}$ знаходяться за рівнем значущості $\alpha/2$.

Зауваження 2. Дуже часто F-критерій застосовується для перевірки припущення про рівність дисперсій, закладеного в критерії Стьюдента.

Критерій, заснований на обчисленні тестової статистики (4.28), є найбільш потужним.

4.8. Перевірка гіпотез про дисперсії декількох нормальних величин

Перевірку гіпотез про дисперсії декількох нормальних сукупностей за вибірками однакового обсягу можна провести методом, викладеним у попередньому параграфі, тобто порівнюючи за критерієм Фішера найбільшу і найменшу з k емпіричних дисперсій, що розглядаються. Якщо при цьому виявиться, що відмінність між ними є незначущою, то тим більше є незначущою і відмінність між іншими дисперсіями.

У випадку, якщо обсяги вибірок, витягнутих з тих нормальних сукупностей, що досліджуються, різні, то можна застосовувати спеціальний критерій значущості – критерій Бартлетта. Застосування критерію Бартлетта засновується на ряді початкових припущень відносно генеральних сукупностей, що досліджуються. Будемо розглядати ці початкові припущення як деяку ймовірнісну модель.

Модель. Нехай є k нормальних сукупностей: $\mathrm{X}_i \to \mathcal{N}(a_i;\,\sigma_i)$. З цих сукупностей витягнуті незалежні вибірки обсягом n_i. Результати кожної вибірки визначимо x_{ij}, $(i = 1, 2, \ldots, k,\;\; j = 1, 2, \ldots, n_i)$. За результатами цих вибірок обчислені точкові оцінки параметрів нормальних законів розподілу:

а) середні арифметичні

$$\overline{x}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \qquad i = 1, 2, \ldots, k; \tag{4.31}$$

б) незсунені оцінки дисперсій

$$s_i^2 \;=\; \frac{1}{n_i - 1}\sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i\right)^2; \tag{4.32}$$

в) зважена за кількістю ступенів вільності середня арифметична незсунених емпіричних дисперсій

$$\overline{s^2} \;=\; \frac{1}{n - k}\sum_{i=1}^{k} \left(x_{ij} - \overline{x}_i\right)^2. \tag{4.33}$$

Потрібно на основі інформації, що здобута в дослідженні, перевірити нульову гіпотезу $\{H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2\}$ проти альтернативної гіпотези $\{H_a :$ *не всі ці дисперсії рівні між собою*$\}$.

Перевірка нульової гіпотези "однорідності" дисперсій засновується на обчисленні вибіркової статистики

$$\chi^2 \;=\; \frac{2,303}{C}\left[(n-k)\,\lg(\overline{s^2}) - \sum_{i=1}^{k}(n_i-1)\lg(s_i^2)\right], \tag{4.34}$$

де

$$C \;=\; 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right).$$

Вибіркова статистика χ^2 за умови справедливості нульової гіпотези H_0 має наближено χ^2-розподіл з $\nu = k-1$ ступенями вільності. Більш точно, розподіл вибіркової статистики (4.34) асимптотично збігається до χ^2-розподілу, причому ця збіжність є дуже швидкою, так що критерій Бартлетта можна застосовувати навіть при дуже малих обсягах вибірок.

Для перевірки нульової гіпотези за формулою (4.34) обчислюється значення $\chi^2_{\text{спос}}$ вибіркової статистики χ^2, що спостерігається.

Потім за таблицею квантилів χ^2-розподілу за заданим рівнем значущості α і кількістю ступенів вільності $\nu = k-1$ знаходять критичну точку $\chi^2_{\alpha;\,\nu}$, що задовольняє умові

$$\Pr\left(\chi^2 \geq \chi^2_{\alpha;\,\nu}\right) \;=\; \alpha. \tag{4.35}$$

Якщо виявиться, що

$$\chi^2_{\text{спос}} \;\geq\; \chi^2_{\alpha;\,\nu}, \tag{4.36a}$$

то нульова гіпотеза відхиляється на користь альтернативної гіпотези.

Якщо

$$\chi^2_{\text{спос}} \;<\; \chi^2_{\alpha;\,\nu}, \tag{4.36b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Якщо перевіряється гіпотеза про дисперсії двох випадкових величин, то критерій Бартлетта має меншу потужність, ніж критерій Фішера, заснований на обчисленні тестової статистики (4.28).

4.9. Перевірка гіпотез про параметр біноміального закону розподілу

У практичних ситуаціях дуже часто при обробці статистичної інформації можна зустрітися з ознаками, що не піддаються кількісній оцінці.

Наприклад, неможливо дати кількісну оцінку математичним здібностям студентів і т. д. У цих випадках прийнято підраховувати частку або відсоток елементів генеральної сукупності, що мають ту чи іншу якісну ознаку. Наприклад, можна підрахувати:

а) частку або відсоток студентів даного університету, що займаються в бібліотеці університету більше 6 годин на тиждень;

б) частку або відсоток цих же студентів, які знають дві іноземні мови;

в) частку або відсоток бракованої продукції в деякій партії;

г) частку або відсоток чоловіків зросту від 168 *см* до 172 *см* і т. д.

Долю елементів генеральної сукупності, що мають якісну ознаку, будемо позначати p $(0 \le p \le 1)$. Вибірковою оцінкою частки є частість (відносна частота) m/n.

Перевірка гіпотез про частість засновується на моделі біноміального розподілу, оскільки частість представляє параметр p в цьому розподілі. Існує багато методів перевірки гіпотез про частість.

Нижче ми розглянемо тільки один тип критеріїв, який можна застосовувати при досить великому обсязі вибірки $(n \ge 100)$. Ці критерії засновані на тому, що вибіркова оцінка частості генеральної сукупності має асимптотично нормальний закон розподілу з параметрами

$$\mathrm{M}[m/n] \;=\; p, \qquad \sigma[m/n] \;=\; \sqrt{p(1-p)/n}. \tag{4.37}$$

Нижче наводяться дві моделі. Перша з них відображає ймовірнісні передумови, що необхідні для перевірки нульової гіпотези рівності частості генеральної сукупності p деякому гіпотетичному числу p_0, тобто гіпотези $\{H_0 : \; p = p_0\}$.

Модель 2 відображає ймовірнісні передумови, необхідні для перевірки нульової гіпотези рівності часток двох генеральних сукупностей.

Модель 1. Нехай кількість елементів x даної генеральної сукупності, що мають деяку якісну ознаку, розподілена згідно з біноміальним законом з параметром p, тобто

$$\mathrm{P}_x \;=\; C_n^x p^x \,(1-p)^{n-x}, \qquad (x = 0, 1, 2, \ldots, n), \tag{4.38}$$

де p— частка елементів генеральної сукупності, що мають деяку якісну ознаку. З цієї генеральної сукупності витягнута незалежна вибірка обсягом n $(n \ge 100)$ і за нею обчислена точкова оцінка параметра p: $\hat{p} = m/n$. Потрібно на основі інформації, що здобута, перевірити нульову гіпотезу $\{H_0 : \; p = p_0\}$, де p_0 — деяке гіпотетичне число, проти альтернативної гіпотези $\{H_a : \; p \ne p_0\}$. У курсі теорії ймовірностей доводиться, що частість m/n має асимптотично нормальний розподіл з математичним сподіванням $\mathrm{M}[m/n] = p$ і середньоквадратичним відхиленням $\sigma[m/n] = \sqrt{p(1-p)/n}$.

Отже, якщо нульова гіпотеза вірна, то нормована вибіркова статистика

$$u \;=\; \frac{m/n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \tag{4.39}$$

має стандартизований нормальний розподіл: $u \to N(0;\, 1)$.

Ця вибіркова статистика застосовується для перевірки нульової гіпотези $\{H_0 : \; p = p_0\}$.

Для перевірки нульової гіпотези за таблицею квантилів стандартизованого нормального розподілу за заданим рівнем значущості α знаходять критичне значення

$u_{\alpha/2}$, що задовольняє умові

$$\Pr\left(|u| \geq u_{\alpha/2}\right) \;=\; \alpha. \tag{4.40}$$

Потім обчислюється значення критерію $u_{\text{спос}}$.

Якщо виявиться, що

$$|u_{\text{спос}}| \;\geq\; u_{\alpha/2}, \tag{4.41a}$$

то нульова гіпотеза відкидається на користь альтернативної.

Якщо ж

$$|u_{\text{спос}}| \;<\; u_{\alpha/2}, \tag{4.41b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження. Якщо альтернативна гіпотеза має вигляд $\{H_a : \; p < p_0\}$, то застосовується u-критерій з лівосторонньою критичною областю. Якщо ж альтернативна гіпотеза має вигляд $\{H_a : \; p > p_0\}$, то застосовується u-критерій з правосторонньою критичною областю. Критичні значення $-u_\alpha$ для критерію з лівосторонньою критичною областю знаходяться з умови $\Pr(u \leq -u_\alpha) = \alpha$, для критерію з правосторонньою критичною областю – з умови $\Pr(u \geq u_\alpha) = \alpha$.

Модель 2. Нехай дані дві генеральні сукупності, що мають біноміальний закон розподілу з параметрами p_1 і p_2 (тут p_1, p_2 – невідомі частості елементів генеральних сукупностей, що мають задану якісну ознаку). З цих генеральних сукупностей витягнуті вибірки обсягом n_1 і n_2 ($n_1 \geq 100$ і $n_2 \geq 100$), після чого обчислені точкові оцінки параметрів p_1 і p_2:

$$\hat{p}_1 \;=\; m_1/n_1, \qquad \hat{p}_2 \;=\; m_2/n_2.$$

Потрібно на основі цієї інформації перевірити нульову гіпотезу $\{H_0 : \; p_1 = p_2\}$ проти альтернативної гіпотези $\{H_a : \; p_1 \neq p_2\}$.

Перевірка нульової гіпотези засновується на обчисленні вибіркової статистики

$$u \;=\; \frac{m_1/n_1 - m_2/n_2}{\sqrt{\overline{p}\,\overline{q}}}\,\sqrt{n}, \tag{4.42}$$

де

$$n = \frac{n_1 \cdot n_2}{n_1 + n_2}; \qquad \overline{p} = \frac{m_1 + m_2}{n_1 + n_2}; \qquad \overline{q} = 1 - \overline{p}. \tag{4.43}$$

Якщо нульова гіпотеза вірна, то ця вибіркова статистика має асимптотично стандартизований нормальний розподіл $\mathcal{N}(0;\,1)$.

Правило перевірки гіпотези залишається тим же, що і в моделі 1.

Якщо

$$|u_{\text{спос}}| \;\geq\; u_{\alpha/2}, \tag{4.44a}$$

то нульова гіпотеза відхиляється на користь альтернативної.

Якщо

$$|u_{\text{спос}}| \;<\; u_{\alpha/2}, \tag{4.44b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

Зауваження. Якщо альтернативна гіпотеза має вигляд $\{H_a : \; p_1 < p_2\}$, то застосовується критерій з лівосторонньою критичною областю. Якщо ж альтернативна гіпотеза має вигляд $\{H_a : \; p_1 > p_2\}$, то застосовується критерій з правосторонньою критичною областю.

4.10. Перевірка гіпотез про математичні сподівання декількох нормальних величин методом однофакторного дисперсійного аналізу

Критерій дисперсійного аналізу є одним з основних понять розділу математичної статистики, що швидко розвивається, – теорії планування експерименту.

Метод дисперсійного аналізу дозволяє перевірити, чи впливають на математичні сподівання випадкових величин певні чинники, які можна довільно змінювати в ході експерименту, вибрати найбільш важливі чинники і оцінити ступінь їх впливу.

Якщо на вказані математичні сподівання впливає тільки один чинник, то відповідний критерій значущості називається *однофакторним дисперсійним аналізом*, якщо ж декілька – *багатофакторним дисперсійним аналізом*. Нижче обмежимося розглядом тільки однофакторного дисперсійного аналізу. Ідея однофакторного дисперсійного аналізу полягає в розбитті загальної дисперсії $s^2_{\text{заг}}$ ВВ X на два незалежних доданки – *факторну дисперсію*, що породжується впливом чинника, який досліджується, і *залишкову дисперсію*, зумовлену різними іншими неврахованими і випадковими чинниками, тобто $s^2_{\text{заг}} = s^2_{\text{факт}} + s^2_{\text{зал}}$.

Зауваження. При застосуванні однофакторного дисперсійного аналізу результати вимірювання випадкової величини X розбиваються залежно від ступіні дії чинника A на групи. З цієї точки зору факторну дисперсію називають іноді міжгруповою, а залишкову – внутрішньогруповою (всередині груп чинник A не діє). Внаслідок порівняння факторної і залишкової дисперсій за критерієм Фішера $F = s^2_{\text{факт}}/s^2_{\text{зал}}$ приходять до висновку про значущість розходження середніх значень в групах.

Сформулюємо основні припущення і обмеження, що містяться в обгрунтуванні дисперсійного аналізу у вигляді ймовірнісної моделі.

Модель. Передбачимо, що з метою вивчення впливу чинника A на деяку результативну ознаку ми розбили, залежно від варіації ознаки A, результати вимірювань на k груп по n_i вимірювань в кожній групі. Будемо розглядати результати вимірювань $\{x_{ij}\}$ (тут $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, де i – номер рівня чинника A, j — номер результату вимірювання на даному рівні) як вибірки з генеральних нормальних сукупностей: $\mathrm{X}_i \to N(a_i; \sigma_i)$ $(i = 1, 2, \ldots, k)$. Параметри (a_i, σ_i) хоч і невідомі, але передбачається, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k$. Виконання останньої рівності можна перевірити за допомогою критерію Бартлетта. Представимо результати вимірювань x_{ij} у вигляді суми двох доданків:

$$x_{ij} = a_i + \varepsilon_{ij}, \tag{4.45}$$

де a_i – математичне сподівання ВВ X_i; ε_{ij} – випадкова помилка (залишок), що характеризує вплив на результати X_i неврахованих і випадкових чинників. Передбачається, що $\varepsilon_{ij} \to N(0; \sigma)$.

Передбачимо, що за вибірковими даними обчислені:

а) групові середні арифметичні

$$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \qquad i = 1, 2, \ldots, k; \tag{4.46}$$

б) загальна середня арифметична

$$< x > = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}, \qquad n = \sum_{i=1}^{k} n_i. \tag{4.47}$$

На основі цієї інформації потрібно перевірити нульову гіпотезу $\{H_0 : a_1 = a_2 = \ldots = a_k\}$ проти альтернативної гіпотези $\{H_a$: *не всі математичні сподівання рівні між собою*$\}$.

Перевірка нульової гіпотези грунтується на обчисленні вибіркової статистики

$$F = \left[\frac{1}{k-1} \sum_{i=1}^{k} (\overline{x}_i - < x >)^2 n_i\right] \left[\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 n_i\right]^{-1}. \tag{4.48}$$

Якщо нульова гіпотеза вірна, то ця вибіркова статистика має F-розподіл з $\nu_1 = k - 1$ і $\nu_2 = n - k$ ступенями вільності. Далі обчислюється значення критерію, що спостерігається. Для зручності всі обчислення розташовуються в таблиці дисперсійного аналізу.

Обчислене значення критерію F, що спостерігається, порівнюється з критичним значенням $F_{\alpha;\nu_1;\nu_2}$, знайденим за таблицею квантилів F-розподілу за заданим рівнем значущості α і кількістю ступенів вільності ν_1 і ν_2.

Якщо

$$F_{\text{спос}} \geq F_{\alpha;\nu_1;\nu_2}, \tag{4.49a}$$

то нульова гіпотеза відхиляється на користь альтернативної гіпотези.

Якщо

$$F_{\text{спос}} < F_{\alpha;\nu_1;\nu_2}, \tag{4.49b}$$

то вважається, що немає підстав для відхилення нульової гіпотези.

4.11. Приклади

Приклад 4.1

На верстаті-автоматі виготовляють деталі з номінальним розміром $a = 12$ *мм*. Відомо, що розподіл розміру є нормальним $X \to \mathcal{N}(a; 0,5)$. Відділ технічного контролю протягом зміни зробив вимірювання 36 випадково відібраних деталей і підрахував середній розмір параметра a: $\overline{x} = 11,7$ *мм*.

Чи можна стверджувати, що верстат-автомат виготовляє деталі зменшеного розміру і тому потрібно налагодити верстат?

Розв'язання

З умови прикладу випливає, що необхідно перевірити нульову гіпотезу $\{H_0 : a = 12$ *мм*$\}$ (*автомат виготовляє деталі номінального розміру*) проти альтернативної гіпотези $\{H_a : a < 12$ *мм*$\}$ (*автомат виготовляє деталі, розмір яких менше номінального*).

Оскільки відділ технічного контролю має підозру, що автомат розрегулювався і виробляє деталі зменшеного розміру, то для перевірки нульової гіпотези застосуємо критерій, що відповідає моделі 1 з лівосторонньою критичною областю.

Обчислимо значення критерію, що спостерігається,

$$u_{\text{спос}} = \frac{\overline{x} - a}{\sigma}\sqrt{n} = \frac{11,7 - 12,0}{0,5}\sqrt{36} = -3,6.$$

За таблицею функції Лапласа за рівнем значущості $\alpha = 0,05$ знаходимо значення (квантиль) $-u_{0,05}$, що задовольняє умові $\Pr(u \le -u_{0,05})$. Це значення (див. додаток) дорівнює $-u_{0,05} = -1,64$.

Оскільки значення критерію u, що спостерігається, знаходиться в критичній області $u_{\text{спос}} = -3,6 < -1,64$, то гіпотезу $\{H_0 : a = 12\ мм\}$ потрібно відхилити на користь альтернативної гіпотези. Це означає, що з імовірністю помилки, меншої ніж 0,05, можна стверджувати, що розмір контрольованих деталей, що виготовляються верстатом-автоматом, є заниженим в порівнянні з номінальним розміром і тому необхідно налагодити верстат.

Приклад 4.2

Технічна норма передбачає в середньому 40 c на виконання певної технічної операції на конвеєрі. Від робітниць, зайнятих на цій операції, надійшли скарги, що вони насправді витрачають на цю операцію більше часу. Для перевірки скарги у 16 робітниць зроблені хронометричні вимірювання часу виконання цієї технічної операції і отримані наступні результати:

$$\overline{x} = 42\ c \qquad \text{(середній час виконання операці)},$$

$$s_{\text{незс}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{16}(x_i - \overline{x})^2} \approx 3,5\ c.$$

Чи можна за хронометричними даними на рівні значущості $\alpha = 0,01$ відхилити гіпотезу, що дійсний середній час виконання цієї технічної операції відповідає нормі?

Розв'язання

З умови прикладу випливає, що нам потрібно перевірити нульову гіпотезу $\{H_0 : a = 40\ c\}$ (*технічна норма встановлена вірно*) проти альтернативної гіпотези $\{H_a : a \ne 40\ c\}$ (*технічна норма встановлена невірно*).

Для перевірки даної нульової гіпотези застосуємо t–критерій значущості (модель 2) з правосторонньою критичною областю.

Обчислимо значення t-критерію

$$t_{\text{спос}} = \frac{\overline{x} - a_0}{s_{\text{незс}}}\sqrt{n-1} = \frac{42 - 40}{3,5}\sqrt{15} = 2,21.$$

За таблицею квантилів розподілу Стьюдента (див. додаток) за рівнем значущості $\alpha = 0,01$ і кількістю ступенів вільності $\nu = n - 1 = 15$ знаходимо значення $t_{0,01;\,15}$, що задовольняє умові $\Pr(t \ge t_{0,01;\,15}) = 0,01$. Це значення $t_{\alpha;\,n-1} = t_{0,01;\,15} = 2,602$.

Оскільки $t_{\text{спос}} = 2,21 < 2,602$, то немає підстав для відхилення нульової гіпотези (перегляду технічної норми часу виконання даної операції).

Таким чином, ми довели, що при $\alpha = 0,01$ різниця (за хронометражем) між середнім часом, що витрачається на дану технічну операцію, і нормою часу є істотно незначущою (випадковою).

Приклад 4.3

Розглядається гіпотеза, що застосування нового типу різця скорочує час виготовлення деякої деталі. Проведено 10 вимірювань часу, що витрачається на виготовлення цієї деталі старим і новим різцем. Отримані наступні результати (в хвилинах):

старий тип різця – 58, 58, 56, 38, 70, 38, 42, 75, 68, 67;
новий тип різця – 57, 55, 63, 24, 67, 43, 33, 68, 56, 54.

Перевірити гіпотезу рівності середнього часу, що витрачається на виготовлення цієї деталі за допомогою двох типів різців. Рівень значущості прийняти $\alpha = 0,05$.

Розв'язання

Передбачимо, що час виготовлення деталі старим і новим типом різця є випадковою величиною, розподіленою згідно з нормальним законом $X \to \mathcal{N}(a_1; \sigma_1)$ та $Y \to \mathcal{N}(a_2; \sigma_2)$, причому a_1 і a_2 – невідомі, а σ_1 і σ_2 хоча і невідомі, але передбачається, що $\sigma_1 = \sigma_2$. Згідно з умовою нам необхідно перевірити нульову гіпотезу $\{H_0 : a_1 = a_2\}$ (*середній час, що витрачається на виготовлення деталі старим і новим типом різця, однаковий*) проти альтернативної гіпотези $\{H_a : a_1 > a_2\}$ (*новий тип різця скорочує середній час опрацювання даної деталі*).

Оскільки обсяги вибірок $n_1 = n_2 = 10$ малі і умови, що закладені в моделі 2, виконуються, то для перевірки нульової гіпотези використаємо правосторонній t-критерій. Обчислимо значення статистики, що спостерігається. Для цього проведемо (див. нижче таблицю) допоміжні обчислення даних для старого типу різця (змінні $\{x_i\}$) і нового типу різця (змінні $\{y_i\}$).

i	x_i	$(x_i - \overline{x})^2$	y_i	$(y_i - \overline{y})^2$
1	58	1	57	25
2	58	1	55	9
3	56	1	63	121
4	38	361	24	784
5	70	169	67	225
6	38	361	43	81
7	42	225	33	361
8	75	324	68	256
9	68	121	56	16
10	67	100	54	4
	$\sum_i x_i = 570$ $\overline{x} = 57,0$	$ns_1^2 = \sum_i (x_i - \overline{x})^2$ $= 1664$	$\sum_i y_i = 520$ $\overline{y} = 52,0$	$ns_2^2 = \sum_i (y_i - \overline{y})^2$ $= 1882$

З наведених в таблиці даних випливає: $s_1^2 = 166,4$ і $s_2^2 = 188,2$.

Отже,

$$t_{\text{спос}} = (\overline{x} - \overline{y}) \left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1/2} =$$

$$= (57 - 52) \left[\frac{1664 + 1882}{10 + 10 - 2} \left(\frac{1}{10} + \frac{1}{10} \right) \right]^{-1/2} = 0,797.$$

За таблицею квантилів розподілу Стьюдента за заданим рівнем значущості $\alpha = 0,05$ і кількістю ступенів вільності $\nu = n_1 + n_2 - 2 = 18$ знаходимо квантиль $t_{0,05;\,18} = 1,734$.

Оскільки $t_{\text{спос}} = 0,797 < 1,734$, то немає підстав для відхилення нульової гіпотези рівності середнього часу, що витрачається на виготовлення деталі двома типами різців.

Це означає, що різниця в середніх арифметичних $(\overline{x} - \overline{y})$ на користь нового типу різця є статистично незначущою (випадковою).

Іншими словами, перевага нового типу різця залишилася недоведеною, хоч це не означає, що цієї переваги немає. При більшому обсязі вибірок ця перевага, якщо вона дійсно є, може бути доведена.

Приклад 4.4

Контролер автопарку визначив, що витрата пального на одній машині в середньому склала $m = 10,0$ *л* на 100 *км*. З метою зменшення витрати пального була проведена модернізація двигунів $n = 25$ автомашин. Після модернізації виявилося, що витрата пального у цих 25 автомашин склала $X^*_{25} = 9,3$ *л* на 100 *км*. Відомо, що розглянута вибірка є нормальною з $\overline{X} = m$ і $\sigma^2 = 4\,л^2$.

Потрібно перевірити гіпотезу: *модернізація не вплинула на витрату пального.*

Розв'язання

1) Розглянемо дві гіпотези:

а) $\{H_0 :\ m = 10,0\}$ (*модернізація не вплинула на витрату*);

б) $\{H_1 :\ m < 10,0\}$ (*модернізація призвела до зменшення витрати*).

2) Приймемо рівень значущості $\alpha = 0,05$.

3) Використовуємо як статистику критерію оцінку математичного сподівання X^*_{25}.

4) Оскільки вибірка отримана з нормальної сукупності, то вибіркове значення також нормальне з дисперсією, що задовольняє співвідношенню $\sigma^{*2}/n = 4/25$, тобто $\sigma^* = 0,4\,(л)$.

Якщо справедлива гіпотеза H_0, то $m = 10,0$.

Перейдемо до стандартизованої змінної $U = (X^*_{25} - m)\,/0,4$, яка є нормальною випадковою величиною $N(0,\,1)$.

5) Альтернативна гіпотеза $\{H_1 :\ m < 10,0\}$.

Тому потрібно використати односторонній критерій. У цьому випадку цей критерій є лівостороннім. Критична область визначається з нерівності $U < u_\alpha$, тобто

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_\alpha} \exp\left(-t^2/2\right)\, dt \;=\; \alpha \;=\; 0,05.$$

За таблицею функції Лапласа за рівнем значущості $\alpha = 0,05$ знаходимо квантиль розподілу $u_{0,05}$. Це значення (див. додаток) дорівнює –1,64.

6) Вибіркове значення стандартизованої статистики критерію складає

$$u_{\text{спос}} \;=\; \frac{X^*_{25} - 10,0}{0,4} \;=\; \frac{9,3 - 10,0}{0,4} \;=\; -1,75.$$

7) Ухвалимо статистичний розв'язок.

Оскільки $u_{\text{спос}} < u_{0,05}$, тобто вибіркове значення статистики належить критичній області, то гіпотеза H_0 відхиляється.

Відповідь: приймається розв'язок H_1: *модернізація двигунів призвела до зменшення витрати пального.*

Примітка. Для прийнятого рівня значущості α межа критичної області складає $X_{\text{кр}} = 10,0 - 0,4 \cdot 1,645 = 9,342$.

Приклад 4.5

Отримана вибірка в 50 електроламп заводу А показала середню тривалість роботи $\overline{x} = 1282$ *годин* із середнім квадратичним відхиленням $s_1 = 80$ *годин*, а така ж за обсягом вибірка того ж типу ламп із заводу Б показала $\overline{y} = 1208$ *годин* із середнім квадратичним відхиленням $s_2 = 94$ *години*.

Перевірити гіпотезу про те, що ці заводи випускають електролампи однакової якості (середній термін служби електроламп обох заводів однаковий). Рівень значущості прийняти $\alpha = 0,05$.

Розв'язання

Оскільки обсяги вибірок досить великі, то застосуємо модель 1. Додатково передбачимо, що тривалість роботи електроламп, що випускаються заводами А і Б, є випадковими величинами, розподіленими згідно з нормальним законом $X \to \mathcal{N}(a_1; s_1)$ та $Y \to \mathcal{N}(a_2; s_2)$, причому $s_1 = 80$, $s_2 = 94$, а величини a_1 і a_2 — невідомі. Згідно з умовою нам необхідно перевірити нульову гіпотезу $\{H_0 : a_1 = a_2\}$ (*середній термін служби ламп, що випускаються заводами* А *і* Б, *однаковий*) проти альтернативної гіпотези $\{H_a : a_1 > a_2\}$ (*лампи, що випускаються заводом* А, *мають більший термін служби*).

Для перевірки нульової гіпотези H_0 застосуємо правосторонній u-критерій. Обчислимо значення статистики, що спостерігається,

$$u_{\text{спос}} = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{1282 - 1208}{\sqrt{80^2/50 + 94^2/50}} = 4,24.$$

За таблицею функції Лапласа знайдемо критичну точку (квантиль) $u_{0,05}$, що задовольняє умові $\Pr(u \geq u_{0,05})$. Це значення дорівнює 1,64.

Оскільки виконується $u_{\text{спос}} = 4,24 > 1,64$, то нульова гіпотеза відхиляється на користь альтернативної.

Іншими словами, вважається "статистично доведеним", що термін служби ламп, що випускаються заводом А, більше терміну служби ламп, що випускаються заводом Б.

Приклад 4.6

Точність роботи верстата-автомата перевіряється за дисперсією розміру деталей, що контролюються, яка не повинна перевищувати $\sigma_0^2 = 0,04$. Взята проба з 11 випадково відібраних деталей. Отримані наступні результати (в міліметрах):

100,6 99,6 100,0 100,1 100,3 100,0 99,9 100,2 100,4 100,6 100,5

На основі даних, що є, перевірити, чи забезпечує верстат задану точність. Рівень значущості прийняти, щоб дорівнював 0,05.

Розв'язання

З умови випливає, що необхідно перевірити нульову гіпотезу $\{H_0 : \sigma = 0,04\}$ (*верстат забезпечує задану точність*) проти альтернативної гіпотези $\{H_a : \sigma > 0,04\}$ (*верстат не забезпечує задану точність*).

Альтернативна гіпотеза сформульована у вигляді $\{H_a : \sigma > 0,04\}$, тому випадок, коли $\sigma < 0,04$, не є істотним. Якщо насправді і виявиться, що $\sigma < 0,04$, то це означає, що станок добре налагоджений і випускає деталі більш високої якості, ніж передбачалося.

Знайдемо точкові оцінки параметрів нормального закону:

$$\hat{a} = \overline{x} = 100,2; \qquad \hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{11}(x_i - \overline{x})^2 = \frac{1,00}{11} = 0,091.$$

Для перевірки нульової гіпотези застосуємо критерій χ^2 з правосторонньою критичною областю. Обчислимо значення тестової статистики, що спостерігається,

$$\chi^2_{\text{спос}} = \frac{1}{\sigma_0^2}\sum_{i=1}^{11}(x_i - \overline{x})^2 = \frac{1}{0,04} = 25.$$

За таблицею квантилів χ^2-розподілу (див. додаток) за заданим рівнем значущості $\alpha = 0,05$ та кількістю ступенів вільності $\nu = 10$ знаходимо критичну точку $\chi^2_{\alpha;\,n-1} = \chi^2_{0,05;\,10}$, що задовільняє умові $\Pr\left(\chi^2 \geq \chi^2_{0,05;\,10}\right)$. Це значення дорівнює 18,307.

Оскільки $\chi^2_{\text{спос}} = 25 > 18,307$, нульова гіпотеза відхиляється на користь альтернативної.

Це означає, що верстат не забезпечує задану точність і вимагає підналадки.

Приклад 4.7

Двома методами проведено вимірювання однієї і тієї ж фізичної величини. Першим методом ця величина вимірювалася 10 разів. Отримані наступні результати:

$$\overline{x}_1 = 10,28; \qquad s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{10}(x_{i1} - \overline{x}_1)^2 = 0,00084.$$

Другим методом ця ж величина вимірювалася 8 разів, що дало

$$\overline{x}_2 = 10,30; \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{8}(x_{i2} - \overline{x}_2)^2 = 0,00041.$$

Чи можна вважати, що обидва методи забезпечують однакову точність? Рівень значущості прийняти $\alpha = 0,05$. Передбачається, що результати вимірювань розподілені нормально і вибірки незалежні.

Розв'язання

З умови прикладу випливає, що необхідно перевірити нульову гіпотезу $\{H_0 : \sigma_1^2 = \sigma_2^2\}$ (*обидва методи забезпечують однакову точність*) проти альтернативної гіпотези $\{H_a : \sigma_1^2 > \sigma_2^2\}$ (*другий метод вимірювань забезпечує більш високу точність*).

Обчислимо значення F–критерію: $F_{\text{спос}} = 0,00084/0,00041 = 2,05$.

За таблицею квантилів F-розподілу Фішера (див. додаток) за рівнем значущості $\alpha = 0,05$ і кількістю ступенів вільності $\nu_1 = 10 - 1 = 9$ і $\nu_2 = 8 - 1 = 7$ знаходимо критичну точку $F_{0,05;\,9;\,7} = 3,68$.

Оскільки $F_{\text{спос}} = 2,05 < 3,68$, то немає підстав для відхилення нульової гіпотези.

Іншими словами, інформація про точність цих методів не дає підстав вважати, що другий метод вимірювання краще першого.

Приклад 4.8

На підприємстві група соціологів досліджувала вплив стажу роботи за професією на продуктивність праці робітників механічного цеху заводу. Отримані наступні результати:

	Стаж роботи		
Результативна ознака	до 10 років	від 10 років до 15 років	від 15 років до 25 років
Кількість деталей, що виробляються за зміну одним робітником, штук	135 156 165 —	176 196 204 180	155 160 149 171 140

Передбачаючи, що продуктивність праці робітників, що мають різний стаж роботи, підкоряється нормальному закону: $\mathrm{X}_i \to \mathcal{N}(a_i;\, \sigma_i)$ (тут $i = 1,\, 2,\, 3$), причому $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, потрібно перевірити методом дисперсійного аналізу нульову гіпотезу $\{H_0 : a_1 = a_2 = a_3\}$ (*середня продуктивність праці не залежить від стажу роботи*). Рівень значущості $\alpha = 0,05$.

Розв'язання

Згідно з умовою прикладу, нульова та альтернативна гіпотези мають вид $\{H_0 : a_1 = a_2 = a_3\}$ – (*середня продуктивність праці не залежить від стажу роботи*) $\{H_a : a_1 \neq a_2 \neq a_3\}$ – (*продуктивність праці залежить від стажу роботи*).

Обчислимо допоміжні величини, необхідні для складання таблиці дисперсійного аналізу :

а) групові середні арифметичні

$$n \;=\; n_1 + n_2 + n_3 \;=\; 3 + 4 + 5 \;=\; 12;$$

$$\overline{x}_1 = \frac{456}{3} = 152; \qquad \overline{x}_2 = \frac{756}{4} = 189; \qquad \overline{x}_3 = \frac{775}{5} = 155;$$

б) загальну середню арифметичну

$$< x > = \frac{1}{n} \sum_{i=1}^{3} \sum_{j=1}^{n_i} x_{ij} \;=\; \frac{1987}{12} \;=\; 165,58.$$

Згідно з (4.48) перевірка нульової гіпотези засновується на обчисленні

вибіркової статистики ($k = 3$)

$$F_{\text{спос}} = s^2_{\text{факт}}/s^2_{\text{зал}} = \left[\frac{1}{k-1}\sum_{i=1}^{k}(\overline{x}_i - \langle x\rangle)^2 n_i\right]\left[\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 n_i\right]^{-1}.$$

Якщо нульова гіпотеза справедлива, то ця статистика має F–розподіл з $\nu_1 = k - 1 = 2$ та $\nu_2 = n - k = 9$ вільностями свободи.

Складемо таблицю дисперсійного аналізу.

Джерело мінливості	Суми квадратів	Кількість ступенів вільності	Дисперсія
Стаж роботи (між групами)	3306,92	2	1653,5
Залишкова (всередині груп)	6228	9	692,0
Повна мінливість	9534,92	11	

Маємо $s^2_{\text{факт}} = 1653,56$ та $s^2_{\text{зал}} = 692$, звідки спостережене значення F-критерію дорівнює $F_{\text{спос}} = 2,389$.

Знайдемо за теблицею квантилів F-розподілу $F_{\alpha;\nu_1;\nu_2}$, відповідне рівню значущості $\alpha = 0,05$ і кількості ступенів вільності $\nu_1 = k - 1 = 2$ та $\nu_2 = n - k = 9$ (див. додаток). Це значення дорівнює $F_{0,05;2;9} = 4,26$.

Оскільки $F_{\text{спос}} = 2,389 < 4,26 = F_{0,05;2;9}$, то приймаємо рішення, що не має підстав для відхилення нульової гіпотези, тобто вважається статистично доведеним, что средняя продуктивність праці залежить від стажу роботи.

Приклад 4.9

Анкетним обстеженням встановлено, що 32% студентів університету є слухачами телевізійних лекцій з математичної статистики. Кафедра прикладної математики надрукувала конспекти телевізійних лекцій і розіслала їх студентам. Після цього знову було зроблено анкетування і встановлено, що 80 студентів з 200 опитаних виявилися слухачами телевізійних лекцій, ($m/n = 80/200 = 0,4$).

Чи можна вважати, що видання телевізійного конспекту лекцій сприяло збільшенню контингенту студентів, що слухають телевізійні лекції з математичної статистики (рівень значущості $\alpha = 0,05$)?

Розв'язання

Згідно з умовою необхідно перевірити нульову гіпотезу $\{H_0 : p = 0,32\}$ проти альтернативної гіпотези $\{H_a : p > 0,32\}$. Для перевірки даної нульової гіпотези застосуємо u–критерій з правосторонньою критичною областю. Обчислимо значення критерію, що спостерігається,

$$u_{\text{спос}} = \frac{m/n - p_0}{\sqrt{\overline{p}_0\,\overline{q}_0}}\sqrt{n} = \frac{0,40 - 0,32}{\sqrt{0,32\cdot 0,68}}\sqrt{200} = 2,43.$$

За таблицею функції Лапласа (див. додаток) за заданим рівнем значущості $\alpha = 0,05$ знайдемо критичне значення u_α, що задовольняє умові $\Pr(u \geq u_\alpha)$. Це значення дорівнює 1,64.

Оскільки $u_{\text{спос}} = 2,43 > 1,64$, то нульова гіпотеза відкидається на користь альтернативної, тобто вважається, що відсоток студентів-заочників, що слухають телевізійні лекції з математичної статистики, значно збільшився.

Приклад 4.10

На іспиті з певної дисципліни викладач ставить студенту тільки одне питання за однією з чотирьох частин курсу. Зі 100 студентів 26 отримали питання з першої частини, 32 – з другої, 17 – з третьої і інші з четвертої.

Чи можна за цими результатами прийняти гіпотезу, що для студента, який прийшов на екзамен, є однакова ймовірність отримати питання за будь-якою з чотирьох частин? Прийняти $\alpha = 0{,}05$.

Розв'язання

У цьому випадку маємо: $m_1 = 26,\ m_2 = 32,\ m_3 = 17,\ m_4 = 25,$ тому $p_i = 0{,}25,$ $n = 100,\ np_i = 25\ \ (i = 1,\ 2,\ 3,\ 4)$. Знаходимо

$$\chi_0^2 \ = \ \frac{(26-25)^2}{25} + \frac{(32-25)^2}{25} + \frac{(17-25)^2}{25} + \frac{(25-25)^2}{25} \ = \ 4,56.$$

Оскільки жоден з параметрів передбачуваного розподілу нами не знаходився за вибіркою, то $s = 0$ і кількість ступенів вільності дорівнює $k - s - 1 = 4 - 0 - 1 = 3$. За таблицею знаходимо межу критичної області. Для $\alpha = 0{,}05$ вона дорівнює 7,815.

Оскільки $4,56 < 7,815,$ то гіпотеза підтвердилася.

Приклад 4.11

Два заводи виготовляють однотипні деталі. Для оцінки їх якості взяті вибірки з продукції цих заводів і отримані наступні результати:

завод № 1

– обсяг вибірки $n_1 = 200,$ кількість бракованих деталей $m_1 = 20;$

завод № 2

– обсяг вибірки $n_2 = 300,$ кількість бракованих деталей $m_2 = 15.$

Визначити, чи є істотна відмінність якості деталей, що виготовляються цими заводами. Рівень значущості $\alpha = 0,05.$

Розв'язання

Згідно з умовою потрібно перевірити нульовою гіпотезу $\{H_0 : \ p_1 = p_2\}$ (*частості бракованих деталей, що виготовляються заводами* № 1 *та* № 2, *рівні*) проти альтернативної гіпотези $\{H_a : \ p_1 \neq p_2\}$. Обчислимо значення u-критерію, що спостерігається. Оскільки

$$m_1/n_1 \ = \ 20/200 \ = \ 0,10; \qquad m_2/n_2 \ = \ 15/300 \ = \ 0,05; \qquad n = \frac{n_1 \cdot n_2}{n_1 + n_2} = 120;$$

$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2} = 0,07; \qquad \overline{q} = 1 - \overline{p} = 0,93,$$

то

$$u_{\text{спос}} \ = \ \frac{m_1/n_1 - m_2/n_2}{\sqrt{\overline{p}\,\overline{q}}}\sqrt{n} \ = \ \frac{0,10 - 0,05}{\sqrt{\overline{p}\,\overline{q}}}\sqrt{120} \ = \ 2,15.$$

За таблицею функції Лапласа (див. додаток) за заданим рівнем значущості $\alpha = 0,05$ знайдемо критичне значення $u_{0,025} = 1,96$.

Оскільки $u_{\text{спос}} > 1,96$, то нульова гіпотеза відхиляється на користь альтернативної, тобто вважається, що якість деталей, що виготовляються цими заводами, різна.

Приклад 4.12

Потрібно перевірити, що три марки будівельного бетону мають однакове розсіювання міцності на стиснення. Для перевірки цієї гіпотези проведено вимірювання міцності на стиснення і отримані наступні результати ($кг/см^2$).

Бетон марки № 1	Бетон марки № 2	Бетон марки № 3
195	215	201
200	201	204
204	202	221
205	198	210
201	—	199

Рівень значущості прийняти $\alpha = 0,05$.

Розв'язання

Передбачимо, що результати вимірювань міцності на стиснення трьох марок бетону підкорюються нормальному або приблизно нормальному розподілам. Згідно з умовою, нам необхідно перевірити нульову гіпотезу $\{H_0 : \ \sigma_1 = \sigma_2 = \sigma_3\}$ проти альтернативної гіпотези $\{H_a : \ \textit{не всі ці дисперсії рівні між собою}\}$.

Об'єднаємо допоміжні обчислення, необхідні для розрахунку вибіркової статистики χ^2 за формулою (4.34), в таблицю.

№	x_{1i}	x_{2i}	x_{3i}	$(x_{1i} - \overline{x}_1)^2$	$(x_{2i} - \overline{x}_2)^2$	$(x_{3i} - \overline{x}_3)^2$
1	195	215	201	36	121	36
2	200	201	204	1	9	9
3	204	202	221	9	4	196
4	205	198	210	16	36	9
5	201	—	199	0	—	64
Суми	1005	816	1035	62	170	314

Звідси:

$$\overline{x}_1 = 201; \qquad \overline{x}_2 = 204; \qquad \overline{x}_3 = 207;$$

$$\overline{s_1^2} = 15,5; \qquad \overline{s_2^2} = 56,7; \qquad \overline{s_3^2} = 78,7.$$

Обчислимо зважену середню арифметичну емпіричних дисперсій:

$$\overline{s^2} = \frac{1}{14-3} \cdot 546 = 49,64;$$

$$\lg \overline{s^2} = 1,696; \qquad (n-k) \lg \overline{s^2} = (14-3) \cdot 1,696 = 18,656.$$

Далі обчислимо

i	s_i^2	$\lg s_i^2$	$n_i - 1$	$(n_i - 1)\lg s_i^2$
1	15,5	1,190	4	4,760
2	56,7	1,754	3	5,262
3	78,7	1,895	4	7,580

Звідси знайдемо суму

$$\sum_{i=1}^{3}(n_i - 1)\lg s_i^2 = 17,602.$$

Стала C дорівнює

$$C \;=\; 1 + \frac{1}{3\cdot 2}\left[\left(\frac{1}{4}+\frac{1}{3}+\frac{1}{4}\right) - \frac{1}{11}\right] \;=\; 1,124.$$

Розрахуємо значення вибіркової статистики χ^2, що спостерігається:

$$\chi^2_{\text{спос}} \;=\; \frac{2,303}{1,124}\cdot(18,656 - 17,602) \;=\; 2,049\cdot 1,053 \;=\; 2,158.$$

За таблицею квантилів χ^2-розподілу (див. додаток) за рівнем значущості $\alpha = 0{,}05$ і кількістю ступенів вільності $\nu = k - 1 = 2$ знайдемо критичну точку $\chi^2_{0,05;\,2} = 5{,}99$.

Оскільки $\chi^2_{\text{спос}} = 2,158 < 5,99$, то немає підстав для відхилення нульової гіпотези. Це означає, що інформація про розсіювання міцності на стиснення трьох марок бетону не дає підстав вважати, що їх рівень розсіювання є різним.

Приклад 4.13 (Перевірка гіпотези про незалежність випадкових величин)

Потрібно перевірити, що внаслідок універсальності χ^2-критерію і його застосовності до багатовимірних розподілів він може служити і для перевірки гіпотези про незалежність випадкових величин.

Передбачимо, що область значень величини X розбита на r_1 інтервалів, а область значень величини Y – на r_2 інтервалів.

Розв'язання

Нехай $\hat{P}_{ij}$ і p_{ij} – випадкова частота та імовірність влучення вектора $[X^{\mathrm{T}}\,Y^{\mathrm{T}}]^{\mathrm{T}}$ в перетин i-го інтервалу значень X та j-го інтервалу значень Y $(i = 1,2,\ldots,r_1;\ j = 1,2,\ldots,r_2)$. Якщо X і Y незалежні, то $p_{ij} = p_{i\cdot}\cdot p_{\cdot j}$, де $p_{i\cdot}$ та $p_{\cdot j}$ – ймовірності влучення X в i-й інтервал і Y в j-й $(i = 1,2,\ldots,r_1;\ j = 1,2,\ldots,r_2)$.

Ймовірності $p_{i\cdot}$ і $p_{\cdot j}$ можна розглядати як $r_1 + r_2 - 2$ невідомих параметра розподілу вектора $[X^{\mathrm{T}}\,Y^{\mathrm{T}}]^{\mathrm{T}}$, при цьому треба мати на увазі співвідношення $\sum_i p_{i\cdot} = 1$ та $\sum_j p_{\cdot j} = 1$.

Ефективними асимптотично нормальними оцінками ймовірностей $p_{i\cdot}$ і $p_{\cdot j}$ можуть виступати відповідні частоти

$$\hat{P}_{i\cdot} \;=\; \sum_{j=1}^{r_2}\hat{P}_{ij}, \qquad \hat{P}_{\cdot j} \;=\; \sum_{i=1}^{r_1}\hat{P}_{ij}.$$

Тому величина

$$Z \;=\; n\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\frac{\left(\hat{P}_{ij}-\hat{P}_{i.}\,\hat{P}_{.j}\right)^2}{\hat{P}_{i.}\,\hat{P}_{.j}}$$

має асимптотично χ^2-розподіл з

$$k \;=\; r_1r_2-(r_1+r_2-2)-1 \;=\; (r_1-1)(r_2-1)$$

кількостями ступенів вільності.

Це дає можливість перевіряти гіпотезу про незалежність двох величин (як скалярних, так і векторних).

Приклад 4.14 (Перевірка гіпотез про збіг розподілів)

Нехай внаслідок N незалежних послідовностей дослідів, що містять n_1, n_2, ..., n_N спостережень, $n_1+n_2+\ldots+n_N=n$, отримані частоти влучення величини X, що спостерігається, в інтервали Δ_ν, $\nu=1,2,\ldots,r$, на які розбита область її можливих значень.

Потрібно перевірити гіпотезу про збіг розподілів величини X (або N різних величин, що спостерігаються) в цих N послідовностях дослідів.

Розв'язання

Нехай $\hat{P}_{\mu\nu}$ – випадкова частота влучення величини, що спостерігається, в ν-й інтервал в μ-ї послідовності дослідів, p_ν – імовірність влучення в ν-й інтервал ($\nu=1,2,\ldots,r$ та $\mu=1,2,\ldots,N$).

Оскільки сума незалежних величин з χ^2-розподілом кожна має χ^2-розподіл з сумарною кількістю ступенів вільності, то при даних імовірностях p_1, p_2, ..., p_r випадкова величина

$$Z \;=\; \sum_{\mu=1}^{N} n_\mu \sum_{\nu=1}^{r}\frac{\left(\hat{P}_{\nu\mu}-p_\nu\right)^2}{p_\nu}$$

описується асимптотичним χ^2-розподілом з $N(r-1)$ кількостями ступенів вільності, якщо розподіл величини, що спостерігається, один і той же у всіх серіях дослідів.

Це дає можливість перевіряти гіпотезу про те, що у всіх N послідовностях дослідів величина, що спостерігається, має один і той же розподіл, для якого ймовірності влучення в інтервали мають дані значення p_1, p_2, $\cdots$, p_r.

4.12. Задачі для розв'язання

Задача 4.1

З автоматичної лінії, що виробляє підшипники, було відібрано 400 штук, причому 10 виявилися бракованими.

Знайти 90 %-й довірчий інтервал для ймовірності появи бракованого підшипника.

Скільки підшипників треба перевірити, щоб з імовірністю P = 0,9973 можна було б стверджувати, що ймовірність появи бракованого підшипника не відрізняється від частоти більш ніж на 5 % ?

Задача 4.2

З великої партії транзисторів одного типу були випадково відібрані і перевірені 100 штук. У 36 транзисторів один номінальний параметр виявився нижче допустимого.

Знайти 95 %-й довірчий інтервал для частки транзисторів з таким дефектом з усієї партії.

Задача 4.3

Для перевірки твердження про те, що ймовірність відмови приладу p дорівнює 0,01, було проведено випробування 100 приладів. При цьому один прилад відмовив.

Побудувати 95 %-ву довірчу межу одностороннього довірчого інтервалу для p за цими даними.

Задача 4.4

При перегляді 10000 волокон з партії льону виявлено 1200 недозрілих.

Скільки треба переглянути волокон льону з цієї партії, щоб з імовірністю P = 0,997 можна було ручатися за точність визначення частки недозрілих волокон з усієї партії в межах 1% ? Відбір безповторний.

Задача 4.5

Вибірково обстежували якість цегли. З 1600 проб у 32 випадках цегла виявилася бракованою.

Потрібно визначити, в яких межах укладається частка браку для всієї продукції, якщо результат необхідно гарантувати з імовірністю P = 0,945.

Задача 4.6

У 10000 сеансах гри гральним з автоматом виграш з'явився 4000 разів.

Знайти 95 %-й довірчий інтервал для ймовірності виграшу. Скільки сеансів гри потрібно провести, щоб з імовірністю 0,99 імовірність виграшу відрізнялася від частоти не більш ніж на 1 %?

Задача 4.7

За схемою повторної вибірки зроблене вибіркове вимірювання виробки на земляних роботах у 145 робітників. Внаслідок обстеження середнє виробки визначене в 4,95 $м^3$ на одного робітника, а її середнє квадратичне відхилення дорівнює 1,5 $м^3$.

Знайти довірчі межі для генерального середнього, що відповідають ймовірності P = 0,9973.

Задача 4.8

У контейнері містяться болти з номінальним значенням розміру $m_0 = 40$ *мм*. Була взята вибірка болтів обсягом $n = 36$. Вибіркове середнє розміру болтів, що контролюються, виявилося $\overline{x} = 40,2$ *мм*. Результати попередніх вимірювань дають підставу передбачати, що дійсні розміри болтів утворять нормальну сукупність з дисперсією $\sigma^2 = 1$ $мм^2$.

Чи можна за результатами проведеного вибіркового обстеження стверджувати, що розмір болтів, що контролюються, не має позитивного зсунення по відношенню до номінального розміру ? Прийняти $\alpha = 0,10$. Яка критична область у цьому випадку ?

Задача 4.9

З урни, що містить невідмінні на дотик чорні і білі кулі в невідомій пропорції, випадково витягується 100 куль (з поверненням). Серед них виявилося 39 чорних куль.

Знайти: а) 90 %–й і б) 95 %–ві довірчі інтервали для частки чорних куль.

Задача 4.10

Розглядається випадкова величина $Z = X - Y$, де X і Y – незалежні випадкові величини. Вибіркові оцінки для X й Y визначалися за результатами $n_1 = 16$ і $n_2 = 36$ спостережень відповідно.

Знайти 95 %-й довірчий інтервал для математичного сподівання Z, якщо $\overline{x} = 10$, $\overline{y} = 4$, σ_X та σ_Y відомі і такі: $\sigma_X = 1$ та $\sigma_Y = 4$.

4.13. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

Цукор, що привозять до складу, упакований в мішки, при цьому кожний мішок повинен містити у середньому 50 *кг* корисної ваги. З'явилися підстави припускати, що в дійсності в мішках міститься цукор меншої ваги. Для перевірки припущення проведено виміри ваги цукру в $n = 16$ мішках (дані про вибіркових середніх наводяться).

Потрібно з'ясувати, можна ли за вимірами, що проведено, відхилити нульову гіпотезу, про те, що дійсна корисна вага відповідає нормі.

Рівень значущості прийняти рівним $\alpha = 0{,}01$.

Результати оформіть графично.

Результат роботи – прийняття рішення відносно нульової гіпотези.

При якому значені рівня значущості α нульову гіпотезу буде відхилено?

Варіант 1

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,2\,\text{кг}, \qquad s_{\text{незс}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,3\,\text{кг}.$$

Варіант 2

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,5\,\text{кг}, \qquad s_{\text{незс}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,2\,\text{кг}.$$

Варіант 3

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 49,8\,\text{кг}, \qquad s_{\text{незс}} = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{1/2} = 0,1\,\text{кг}.$$

Завдання 2

На першому та другому верстаті виробляються деталі. Розглядається гіпотеза, що другий верстат (нового типу) скорочує час вироблення деякої деталі. Проведено 10 вимірів часу, що тратиться на вироблення цієї деталі першим та другим верстатом (дані наводяться в таблиці).

Потрібно перевірити нульову гипотезу про рівність середнього часу, що тратиться на вироблення цієї деталі на кожному з цих верстатів.

Рівень значущості прийняти рівним $\alpha = 0,05$.

Результати оформіть графічно.

Результат роботи – прийняття рішення відносно нульової гіпотези.

При якому значені рівня значущості α нульову гіпотезу буде відхилено?

Варіант 1

Перший верстат	38	38	36	18	50	18	22	55	48	47
Другий верстат	27	35	43	14	47	23	13	48	36	34

Варіант 2

Перший верстат	35	37	38	21	52	19	24	54	49	47
Другий верстат	26	34	43	15	46	22	14	43	33	34

Варіант 3

Перший верстат	48	48	46	28	60	28	32	75	58	57
Другий верстат	37	45	53	24	57	33	23	58	46	44

4.14. Завдання для перевірки

1. Сформулюйте визначення статистичної гіпотези.

2. Що називається параметричною статистичною гіпотезою? Непараметричною статистичною гіпотезою?

3. Наведіть приклади нульової, альтернативної, простої і складної гіпотез. Поясніть принцип перевірки нульових гіпотез за допомогою статистичних критеріїв значущості.

4. Що називається помилкою першого роду? Помилкою другого роду? Дайте геометричну інтерпретацію ймовірностям здійснення помилок першого і другого роду.

5. Як змінюються ймовірності здійснення помилок першого і другого роду при збільшенні обсягу вибірки?

6. Чи залежать імовірності здійснення помилок першого і другого роду від вигляду альтернативної гіпотези? Від критерію, що застосовується?

7. У чому полягає однобічність дії статистичних критеріїв значущості?

8. Чи береться до уваги ймовірність здійснення помилки другого роду при перевірці нульових гіпотез за допомогою статистичних критеріїв значущості?

9. Чи можна, застосовуючи статистичний критерій значущості, зробити висновок: "*Нульова гіпотеза, що перевіряється, вірна*"?

10. У чому полягає відмінність між побудовою двосторонньої критичної області і побудовою довірчого інтервалу для одного і того ж параметра?

11. Як знаходяться критичні точки (квантилі) статистичних критеріїв значущості (u, t, χ^2, F) у випадку двосторонньої критичної області? У випадку лівосторонньої критичної області? У випадку правосторонньої критичної області?

12. За допомогою яких вибіркових статистик проводиться перевірка гіпотез про математичне сподівання однієї випадкової величини? Двох випадкових величин? Декількох випадкових величин?

13. За допомогою яких вибіркових статистик проводиться перевірка гіпотез про дисперсії однієї випадкової величини? Двох випадкових величин? Декількох випадкових величин?

14. За допомогою яких вибіркових статистик проводиться перевірка гіпотез про параметр p біноміального закону розподілу?

15. Як змінюється ймовірність здійснення помилок другого роду при $\alpha \to 0$?

5. Статистична перевірка непараметричних гіпотез

5.1. Основні поняття

У попередніх розділах було приділено увагу перевірці гіпотез відносно параметрів законів розподілу, вигляд яких передбачався відомим (нормальний, біноміальний і т. д.).

Під час опрацювання статистичних даних для характеристики частотних властивостей ряду спостережень $x_1, x_2, \ldots, x_n$ експериментатор добирає теоретико-ймовірнісну модель (нормальну, показникову, біноміальну і т. д.). Експериментатор візуально на вигляд гістограми (полігона частостей) або з будь-яких інших міркувань може висунути гіпотезу про множину функцій Ω певного вигляду (нормальних, показникових, біноміальних тощо), до якій може належати функція розподілу випадкової величини X. Припущення такого роду називаються *непараметричними гіпотезами.*

Визначення. *Нульовою непараметричною гіпотезою* називається гіпотеза відносно загального вигляду функції розподілу ймовірностей випадкової величини X, тобто гіпотеза вигляду $\{H_0 : F(x) = F_0(x)\}$.

Гіпотетична (передбачувана) функція розподілу випадкової величини X може бути визначена повністю або з точністю до її параметрів, тобто нульова гіпотеза може мати вигляд $\{H_0 : F(x) \in \Omega\}$, де Ω означає сукупність функцій певного вигляду (нормальних, показникових, біноміальних і т. д.).

Припустимо, що клас таких функцій вибраний і зроблена точкова оцінка параметрів цих функцій всередині вибраного класу. Подальша задача експериментатора полягає в перевірці гіпотези, що висунена, про клас функцій Ω, тобто в з'ясуванні, наскільки добре підібрана ймовірнісна модель ряду спостережень.

Перевірка гіпотези про передбачуваний закон розподілу проводиться за допомогою непараметричних критеріїв значущості. Принципи побудови таких критеріїв і методика перевірки залишаються практично тими ж, що і при перевірці параметричних гіпотез, тобто перевірка непараметричних гіпотез проводиться на основі обчислення деякої вибіркової статистики (критерію), закон розподілу якої отриманий в припущенні істинності нульової гіпотези, і порівняння значення цієї вибіркової статистики, що спостерігається, з критичним значенням.

Непараметричні критерії значущості умовно можна поділити на дві групи.

До першої групи відносяться *критерії згоди*, за допомогою яких перевіряються нульові гіпотези відносно загального вигляду функції розподілу. Найбільш поширеними критеріями згоди є критерій згоди χ^2 Пірсона і λ-критерій Колмогорова.

До іншої численної *групи непараметричних критеріїв* відносяться критерії, за допомогою яких перевіряється нульова гіпотеза про належність двох вибірок однієї

і тієї ж генеральної сукупності (або про те, що дві генеральні сукупності мають одну і ту ж функцію розподілу).

У практиці за допомогою непараметричних критеріїв значущості особливо часто перевіряється нульова гіпотеза про те, що генеральна сукупність, яка досліджується, має нормальний закон розподілу. Наприклад, при застосуванні параметричних критеріїв для перевірки гіпотез відносно параметрів нормального закону передбачалося, що генеральна сукупність є нормальною. Це припущення заздалегідь потрібно перевірити за допомогою критеріїв згоди, а потім, якщо нульова гіпотеза не відхилена, застосовувати параметричні критерії значущості.

5.2. Критерій згоди χ^2 Пірсона

Статистичний критерій χ^2 Пірсона дозволяє проводити перевірку згоди емпіричної функції розподілу з гіпотетичною інтегральною функцією $F(x)$, що належить до деякої множини Ω функцій певного вигляду (нормальних, показникових, біноміальних та інших). Сформулюємо основні ймовірнісні передумови і обмеження, які повинні бути виконані при застосуванні критерію χ^2 у вигляді моделі.

Модель. Нехай генеральна сукупність має функцію розподілу $F(x)$, що належить деякому класу функцій Ω. З генеральної сукупності вилучена вибірка обсягом n $(n \geq 50)$.

Розіб'ємо весь діапазон отриманих результатів на k часткових інтервалів рівної довжини і нехай в кожному частковому інтервалі виявилося m_i вимірювань, причому

$$\sum_{i=1}^{k} m_i \;=\; n. \tag{5.1}$$

Складемо згрупований статистичний ряд:

Інтервали спостережених значень ВВ X	$[x_0; x_1]$	$[x_1; x_2]$	$\ldots$	$[x_{i-1}; x_i]$	$\ldots$	$[x_{k-1}; x_k]$
Частоти	m_1	m_2	$\ldots$	m_i	$\ldots$	m_k

Потрібно на основі інформації, що є, перевірити нульову гіпотезу про те, що гіпотетична функція розподілу $F(x)$ значуще подає дану вибірку, тобто гіпотезу $\{H_0 :\; F(x) \in \Omega\}$.

При перевірці нульової гіпотези за допомогою критерію згоди χ^2 дотримуються наступної послідовності дій:

1) На основі гіпотетичної функції розподілу $F(x)$ обчислюють імовірності влучення випадкової величини X в часткові інтервали групування $[x_{i-1};\, x_i]$:

$$p_i \;=\; \Pr\left(x_{i-1} \leq \mathrm{X} < x_i\right) \;=\; \int\limits_{x_{i-1}}^{x_i} f(x)\,dx \;=\; F(x_i) - F(x_{i-1}), \tag{5.2}$$

де $i = 1, 2, \ldots, k$.

2) Помножуючи отримані імовірності p_i на обсяг вибірки n, обчислюють знайдені теоретично частоти np_i часткових інтервалів $[x_{i-1},\, x_i]$, тобто частоти, які потрібно чекати, якщо нульова гіпотеза справедлива.

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

$$\chi^2_{\text{спос}} \;=\; \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}. \tag{5.3}$$

Можна показати, що якщо нульова гіпотеза вірна, то при $n \to \infty$ закон розподілу вибіркової статистики (5.3), незалежно від вигляду функції $F(x)$, прямує до закону розподілу χ^2 з $\nu = k - r - 1$ ступенями вільності (тут k – кількість часткових інтервалів; r – кількість параметрів гіпотетичної функції $F(x)$, що оцінюються за даними вибірки).

4) Для того щоб перевірити нульову гіпотезу, необхідно знайти за таблицею квантилів χ^2-розподілу (див. додаток) за заданим рівнем значущості α і кількістю ступенів вільності $\nu = k - r - 1$ критичне значення $\chi^2_{\alpha;\,\nu}$, що задовольняє умові

$$\Pr\left(\chi^2 \geq \chi^2_{\alpha;\,\nu}\right) \;=\; \alpha. \tag{5.4}$$

Критерій χ^2 сконструйований таким чином, що чим ближче до нуля значення критерію χ^2, тим імовірніше, що нульова гіпотеза справедлива. Тому для перевірки нульової гіпотези застосовується критерій χ^2 з правосторонньою критичною областю.

Якщо виявиться, що

$$\chi^2_{\text{спос}} \;\geq\; \chi^2_{\alpha;\,\nu}, \tag{5.5a}$$

то нульова гіпотеза $\{H_0 :\; F(x) \in \Omega\}$ відкидається на користь альтернативної $\{H_1 :\; F(x) \notin \Omega\}$, тобто вважається, що гіпотетична функція не узгоджується з дослідженими даними.

Якщо ж

$$\chi^2_{\text{спос}} \;<\; \chi^2_{\alpha;\,\nu}, \tag{5.5b}$$

то вважається, що немає підстав для відхилення нульової гіпотези, тобто гіпотетична функція $F(x)$ узгоджується з даними, що були отримані під час експерименту.

Зауваження. При застосуванні критерію χ^2 необхідно, щоб в кожному частковому інтервалі було не менше ніж 5 елементів. Якщо число елементів (частота) менш ніж 5, то рекомендується об'єднувати такі часткові інтервали з сусідніми.

5.3. Критерій згоди λ Колмогорова

Вище був викладений критерій згоди χ^2, що дозволяє перевірити гіпотезу про згоду даних вибірки з конкретним теоретичним законом розподілу для будь-якої випадкової величини, як неперервної, так і дискретної. *Критерій згоди λ Колмогорова* застосовується для перевірки гіпотез про закони розподілу тільки неперервних величин.

Його відмінність від критерію згоди χ^2 Пірсона полягає в тому, що при застосуванні критерію згоди χ^2 порівнювалися емпіричні і теоретичні частоти розподілу;

при застосуванні λ-критерію Колмогорова порівнюються емпірична $F^*(x)$ і гіпотетична $F(x)$ функції розподілу. Крім того, при застосуванні λ-критерію Колмогорова передбачається, що теоретичні значення параметрів гіпотетичної функції відомі (в критерії згоди χ^2 вони можуть визначатися за даними вибірки). Ці обмеження звужують область практичного застосування λ-критерію Колмогорова. Проте цей критерій широко застосовується на практиці.

При його використанні невідомі теоретичні параметри гіпотетичного розподілу оцінюються за даними вибірок великого обсягу, паралельно з тими, що досліджуються, або за даними вибірки, що досліджується.

В останньому випадку λ-критерій Колмогорова стає наближеним в тому значенні, що дійсний рівень значущості α приблизно дорівнює заданому рівню α ($\alpha_{\text{факт}} < \alpha_{\text{задан}}$). У випадку, коли параметри гіпотетичного закону розподілу оцінюються за даними вибірки, що досліджується, статистичний λ-критерій Колмогорова показує кращу згоду з емпіричними даними, ніж критерій згоди χ^2 Пірсона. Тому при його застосуванні рекомендується використати трохи більший рівень значущості $\alpha = 0{,}10 - 0{,}20$.

Нижче наводяться дві ймовірнісні моделі. У першій з них вказані ймовірнісні передумови і правило перевірки нульової гіпотези про вигляд інтегральної функції розподілу неперервної випадкової величини за допомогою λ-критерію Колмогорова. У другій моделі дані ймовірнісні передумови і правило перевірки нульової гіпотези про приналежність двох вибірок до однієї і тієї ж генеральної сукупності (або дві генеральні сукупності мають одну і ту ж функцію розподілу). Критерій, що використовується при цьому, носить назву *λ-критерій Смірнова-Колмогорова*.

Модель 1. Нехай відомо, що випадкова величина X має неперервну функцію розподілу $F(x)$. З генеральної сукупності з функцією розподілу $F(x)$ вилучена випадкова вибірка обсягом n ($n \geq 50$). На основі здобутої інформації потрібно перевірити нульову гіпотезу $\{H_0 :$ *емпіричні дані узгоджуються з гіпотетичною функцією розподілу*$\}$.

Перевірку нульової гіпотези за допомогою критерію згоди Колмогорова виконують за наступною схемою:

1) розташовують результати спостережень в зростаючому порядку або подають їх у вигляді інтервального статистичного ряду;

2) знаходять емпіричну функцію розподілу

$$F^*(x) = \frac{n_x}{n}; \tag{5.6}$$

3) обчислюють, користуючись гіпотетичною функцією розподілу, значення теоретичної функції розподілу $F(x)$, відповідні до спостережених значень випадкової величини X;

4) знаходять для кожного поточного значення x_i модуль різниці між емпіричною і теоретичною функціями розподілу:

$$|F^*(x) - F(x)|; \tag{5.7}$$

5) обчислюють значення вибіркової λ-статистики:

$$\lambda = D\sqrt{n}, \tag{5.8}$$

де

$$D \;=\; \max_x |F^*(x) - F(x)|. \tag{5.9}$$

Академік А. М. Колмогоров показав, що якщо нульова гіпотеза вірна, то вибіркова статистика $\lambda = D\sqrt{n}$ при $n \to \infty$ має функцію розподілу, яку прийнято позначати $K(\lambda)$, наступного вигляду:

$$K(\lambda) \;=\; \Pr(D\sqrt{n} < \lambda) \;=\; \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda^2\right). \tag{5.10}$$

Задамо рівень значущості α.

Тоді зі співвідношення

$$\Pr(\lambda \geq \lambda_\alpha) \;=\; \Pr(D\sqrt{n} \geq \lambda_\alpha) \;=\; 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda_\alpha^2\right) \;=\; \alpha \tag{5.11}$$

можна знайти квантилі λ-розподілу Колмогорова.

Таблиця квантилів (критичних значень) розподілу Колмогорова наведена у додатку.

Порівняємо значення вибіркової статистики $\lambda_{\text{спос}} = D\sqrt{n}$, що спостерігається, з критичним значенням λ_α, що визначається за таблицею квантилів розподілу Колмогорова за заданим рівнем значущості α.

Якщо при цьому виявиться, що

$$D\sqrt{n} \;\geq\; \lambda_\alpha, \tag{5.12a}$$

то гіпотеза, що перевіряється, відхиляється.

Якщо ж

$$D\sqrt{n} \;<\; \lambda_\alpha, \tag{5.12b}$$

то вважається, що в цьому випадку немає підстав для відхилення нульової гіпотези, тобто гіпотетична функція розподілу вважається узгодженою з емпіричними даними.

Модель 2. Нехай внаслідок спостережень отримані дві випадкові вибірки обсягом n_1 і n_2 ($n_1 \geq 50$ і $n_2 \geq 50$). Потрібно на основі здобутої інформації перевірити нульову гіпотезу $\{H_0$: *дві вибірки вилучені з однієї і тієї ж генеральної сукупності з гіпотетичною функцією розподілу* $F(x)\}$.

Перевірка нульової гіпотези базується на обчисленні вибіркової статистики λ-критерію Смірнова-Колмогорова:

$$\lambda \;=\; D^*\sqrt{n} \;=\; \max_x |F_1^*(x) - F_2^*(x)|\,\sqrt{n}, \tag{5.13}$$

де $n \;=\; n_1 n_2/(n_1 + n_2)$, а $F_1^*(x)$ і $F_2^*(x)$ – емпіричні функції розподілу, побудовані за даними першої і другої вибірок відповідно.

Якщо нульова гіпотеза вірна, то при $n \to \infty$ розподіл вибіркової статистики λ асимптотично збігається до розподілу Смірнова-Колмогорова *незалежно від вигляду функції* $F(x)$. За таблицею квантилів розподілу Смірнова-Колмогорова (див. додаток) для заданого рівня значущості α знаходять критичні значення λ_α, що задовольняють умові

$$\Pr(\lambda \geq \lambda_\alpha) \;=\; \alpha. \tag{5.14}$$

Якщо виявиться, що

$$D^*\sqrt{n} \;\geq\; \lambda_\alpha, \tag{5.15a}$$

то нульова гіпотеза відхиляється.

Якщо ж

$$D^*\sqrt{n} \;<\; \lambda_\alpha, \tag{5.15b}$$

то вважається, що немає підстав для відхилення гіпотези про те, що дві сукупності, що досліджуються, мають ту саму функцію розподілу.

5.4. Критерій знаків

На практиці часто доводиться мати справу з випадковими величинами, закон розподілу яких є невідомим. У цьому випадку замість параметричних критеріїв значущості можна використати *непараметричні критерії значущості*, при застосуванні яких не потрібно робити будь-які припущення про закон розподілу випадкової величини, що досліджується.

Непараметричні критерії в порівнянні з параметричними мають дещо меншу потужність, тобто вони менш ефективні. Однак цей недолік компенсується більш простою побудовою вибіркових статистик, на обчислення яких витрачається набагато менше часу.

Критерій знаків – один з найпростіших непараметричних критеріїв, за допомогою якого перевіряється нульова гіпотеза про те, що дві вибірки вилучені з однієї і тієї ж генеральної сукупності. Його головна перевага полягає в тому, що при застосуванні не треба ставити ніяких обмежень відносно вигляду функції розподілу, крім її неперервності.

Критерій знаків застосовується як критерій порівняння спарених спостережень. Звичайно порівнюються результати двох вибірок однакового обсягу. Критерій знаків, наприклад, часто застосовується для перевірки стійкості генеральної сукупності у часі, тобто для порівняння двох спарених спостережень, які відповідають двом моментам часу. Його можна також застосовувати при обробці експериментів для порівняння величин, що вимірюються в експериментальній і контрольній групах.

Модель. Нехай дані дві випадкові вибірки $(x_1, x_2, \ldots, x_n)$ та $(y_1, y_2, \ldots, y_n)$ однакового обсягу n. Потрібно перевірити нульову гіпотезу про те, що вони вилучені з однієї і тієї ж генеральної сукупності.

Для перевірки даної нульової гіпотези за допомогою критерію знаків досліджують знаки різниць спарених результатів обох вибірок і знаходять число тих знаків, яких менше.

Визначимо їх кількість r.

Якщо нульова гіпотеза вірна, то

$$\Pr(x - y > 0) \;=\; \Pr(x - y < 0) \;=\; \frac{1}{2}, \tag{5.16}$$

кількість r є дискретною випадковою величиною, розподіленою згідно з біноміальним

законом з параметром $p = \frac{1}{2}$,

$$\mathrm{P}_n(r) \;=\; C_n^r \left(\frac{1}{2}\right)^n. \tag{5.17}$$

Нехай тепер знайдено, що r_α – найменше значення кількості r, для якого виконується нерівність $\mathrm{P}_n(r) \le \alpha$. Таблиця критичних значень кількості знаків r_α, відповідних до заданого рівня значущості α і обсягу вибірки n, наведена в додатку.

Якщо

$$r_{\text{спос}} \;\le\; r_\alpha, \tag{5.18a}$$

то нульова гіпотеза відхиляється, тобто вважається, що вибірки вилучені з генеральних сукупностей з різними функціями розподілу.

Якщо ж

$$r_{\text{спос}} \;>\; r_\alpha, \tag{5.18b}$$

то вважається, що немає підстав для відхилення нульової гіпотези про те, що дві вибірки вилучені з однієї і тієї ж генеральної сукупності.

5.5. Методичні вказівки з застосування критеріїв згоди

Критерії згоди, як і всі непараметричні критерії, є статистичними критеріями значущості. Це означає, що за їх допомогою нульова гіпотеза про вигляд функції розподілу або відхиляється, або вважається, що здобута у досліді інформація не дає достатніх підстав для відхилення гіпотезі, що висунена, про вигляд функції розподілу.

Якщо обсяг вибірок n малий ($n \le 50$) або результати вимірювань розташовуються у досить вузькому інтервалі змін випадкової величини X, то експериментальні дані можуть досить добре узгоджуватися з рядом різних імовірнісних моделей, тобто з різними законами розподілу. Тому не треба надавати дуже великого значення позитивному результату ("*Нульова гіпотеза не відхиляється*") перевірки нульових гіпотез про вигляд функції розподілу за допомогою функції згоди.

У сучасній практиці все ширше застосовують критерії згоди не стільки для перевірки згоди експериментальних даних з деякою гіпотетичною функцією розподілу, скільки для добору найкращої функції розподілу (ймовірнісної моделі) з ряду функцій (моделей), що розглядаються. Вибір відповідного закону розподілу повинен базуватися передусім на розумінні механізму явища, що вивчається. Однак якщо механізм явища, що вивчається, є невідомим, то попередній вибір закону розподілу може бути зроблений, виходячи з наступних міркувань:

1) Із зовнішнього вигляду гістограми частостей статистичного ряду розподілу. Вигляд гістограми дає орієнтування на можливий закон розподілу. Наприклад, якщо гістограма має багатомодовий (багатогорбний) вигляд, таку її якість, можливо, слід з'ясувати змішуванням різнорідних за своїми якостями об'єктів спостередження.

Переваги – простота застосування, наочність. Недоліки – гістограма може одночасно нагадувати декілька законів розподілу. Наприклад, із вигляду гістограми практично не можна розрізнити логарифмічно нормальний закон і закон розподілу Вейбулла навіть при великому обсязі вибірки.

2) За допомогою графічного відображення емпіричної функції розподілу на ймовірнісних шкалах (це зручно виконувати за допомогою ЕОМ).

Якщо закон розподілу вибраний правильно, то при нанесенні емпіричної функції розподілу на ймовірнісний папір (зі шкалами, що відповідають гіпотетичному закону розподілу) ці значення будуть розташовуватися на прямій лінії або поблизу прямої лінії.

Переваги методу – простота, наочність. Недоліки – необхідно мати спеціальні папери; відсутність кількісного критерію можливого відхилення значень емпіричної функції розподілу від прямої лінії; неоднозначність вибору закону, викликана тим, що іноді на таких імовірнісних паперах зі шкалами, що відповідають різним законам розподілу, значення емпіричної функції розподілу розташовуються приблизно однаково.

3) Попередній вибір закону може здійснюватися за величиною емпіричного коефіцієнта варіації $Var = s/\overline{x}$. Відомо, що кожному закону розподілу відповідає певний наближений діапазон значень коефіцієнта варіації (див. зведення-таблицю). Переваги методу – простота. Недоліки – коефіцієнт варіації не відображає ступінь симетрії емпіричної кривої розподілу; неоднозначність вибору.

Закон розподілу випадкової величини X	Межі зміни	Середнє значення
Нормальний закон	[0,08; 0,40]	0,25
Закон Вейбулла	[0,40; 0,85]	0,71
Логарифмічний закон	[0,35; 0,80]	0,68
Експоненційний закон	[0,60; 1,30]	0,92

4) За дослідними даними раніше проведених досліджень.

Недолік методу – можуть бути значні розходження в механізмах явищ, що описуються випадковими величинами, відмінних від раніше описаних.

5) Як наближений критерій для попереднього вибору закону розподілу можуть бути використані вибіркові коефіцієнти асиметрії і ексцесу.

Якщо із вибірки обсягом n знайдені точкові оцінки асиметрії і ексцесу

$$\hat{A} = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3, \qquad \hat{E} = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3 \tag{5.19}$$

та їх середніх арифметичних відхилень

$$s_{_A} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}, \qquad s_{_E} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+4)}}, \tag{5.20}$$

то емпірична функція вважається узгодженою з гіпотетичною функцією за умови, що вибіркові коефіцієнти асиметрії і ексцесу відрізняються за абсолютною величиною від своїх математичних сподівань не більш ніж на 3 середніх квадратичних відхилення. Таким чином, якщо

$$|\hat{A} - m_{_A}| < 3\sigma_{_A} \qquad \text{та} \qquad |\hat{E} - m_{_E}| < 3, \tag{5.21}$$

то вважається, що нульова гіпотеза узгоджується з експериментальними даними.

Якщо хоч одна з цих нерівностей не виконується, то нульова гіпотеза, що висунена, відхиляється.

Зауваження. Для нормального закону розподілу математичні сподівання вибіркових коефіцієнтів асиметрії і ексцесу дорівнюють нулю. Тому гіпотеза нормальності приймається, якщо $|\hat{A}| < 3\sigma_{\hat{A}}$ та $|\hat{E}| < 3$.

Перевага методу – облік симетрії і крутості, тобто форми кривої. Недолік – немає суворої кількісної оцінки допустимого розходження між вибірковими коефіцієнтами асиметрії і ексцесу та їх математичними сподіваннями, оскільки правило "трьох сигм" є емпіричним.

Після попереднього вибору закону розподілу за одним або декількома перерахованими методами рекомендується застосовувати суворі критерії згоди.

При цьому потрібно мати на увазі, що крім викладених в даному посібнику основних критеріїв згоди χ^2 і критерію λ Колмогорова, є ряд інших специфічних критеріїв згоди.

Наприклад, критерій ω^2 Мізеса-Смірнова в протилежність критерію χ^2 не вимагає об'єднання числових даних в розряди, тобто більш повно використовує інформацію, що міститься у вибірці. Є, наприклад, спеціальний критерій для перевірки гіпотез нормальності за сукупністю досить великого числа вибірок малого обсягу.

5.6. Розгорнений приклад опрацювання даних для нормального закону розподілу

Наведемо розгорнене розв'язання прикладу, в якому використовується нормальний закон розподілу і застосовуються основні поняття математичної статистики, засновані на використанні нормального закону розподілу.

Приклад

Під час свердлування 80 отворів одним свердлом і подальшим вимірюванням діаметрів отворів отримані такі дані (в *мм*):

40,26	40,35	40,44	40,35	40,39	40,40	40,42	40,32	40,35	40,34
40,44	40,35	40,30	40,34	40,31	40,32	40,33	40,41	40,32	40,34
40,33	40,38	40,33	40,33	40,28	40,30	40,40	40,36	40,37	40,38
40,42	40,35	40,29	40,33	40,31	40,33	40,36	40,34	40,32	40,36
40,41	40,40	40,33	40,37	40,34	40,30	40,43	40,34	40,37	40,35
40,34	40,31	40,43	40,36	40,34	40,34	40,28	40,44	40,35	40,30
40,31	40,31	40,36	40,34	40,29	40,39	40,39	40,37	40,32	40,32
40,36	40,41	40,27	40,38	40,37	40,37	40,36	40,35	40,30	40,30

Потрібно:

1. Скласти інтервальні статистичні ряди розподілу частот і частостей спостережених значень неперервної випадкової величини X – діаметрів отворів x_i.

2. Побудувати гістограму і полігон частостей діаметрів отворів.

3. Знайти емпіричну функцію розподілу і побудувати її графік.

4. Обчислити числові характеристики вибірки:

– середнє арифметичне;

– вибіркову дисперсію;

– вибіркове середнє квадратичне відхилення;

– вибіркові коефіцієнти асиметрії і ексцесу;

– вибірковий коефіцієнт варіації.

5. З вигляду гістограми і полігона частостей, а також за значеннями вибіркових коефіцієнтів асиметрії і ексцесу, і, виходячи з механізму утворення випадкової величини X, зробити попередній вибір вигляду закону розподілу цієї випадкової величини.

6. Знайти точкові оцінки параметрів нормального закону розподілу (передбачається, що випадкова величина розподілена згідно з нормальним законом), записати густину імовірностей і функцію розподілу випадкової величини X.

7. Знайти теоретичні частоти нормального закону розподілу, перевірити згоду емпіричної функції розподілу з нормальним законом за допомогою основних критеріїв згоди – критерію χ^2 Пірсона і λ-критерію Колмогорова.

8. Знайти інтервальні оцінки обох параметрів нормального закону розподілу. Довірчу ймовірність прийняти $\mathrm{P}= 1 - \alpha = 0,95$.

Розв'язання

1. Вивчення неперервних випадкових величин починається з угрупування статистичного матеріалу, тобто з розбиття інтервалу спостережених значень випадкової величини X на k часткових інтервалів однакової довжини і підрахунку частот влучення спостережених значень ВВ X в часткові інтервали групування. Кількість інтервалів вибирається довільно. Звичайно число інтервалів буває не менше за 5 і не більше за 15.

Розіб'ємо весь діапазон спостережених значень на 5 інтервалів (розрядів). Довжину часткового інтервалу визначимо за формулою

$$h \;=\; \frac{1}{5}\,(x_{\max} - x_{\min}) \;=\; \frac{1}{5}\,(40,44 - 40,26) \;\approx\; 0,04.$$

За початок першого інтервалу приймаємо величину a_0, яка дорівнює $a_0 = 40,26 - 0,02 = 40,24$, тоді перший інтервал буде $[40,24;\ 40,28]$, другий – $[40,28;\ 40,32]$ і так далі. Шкала інтервалів і угрупування початкових статистичних даних зведені в таблицю. В результаті отримуємо статистичний ряд розподілу частот:

Інтервали спостережених значень ВВ X	[40,24; 40,28]	[40,28; 40,32]	[40,32; 40,36]	[40,36; 40,40]	[40,40; 40,44]
Частота m_i	4	19	32	15	10

Контроль: $n = \sum\limits_{i=1}^{5} m_i = 80.$

Для отримання статистичного ряду частостей розділимо частоти m_i на обсяг вибірки n. У результаті отримуємо інтервальний статистичний ряд розподілу частостей.

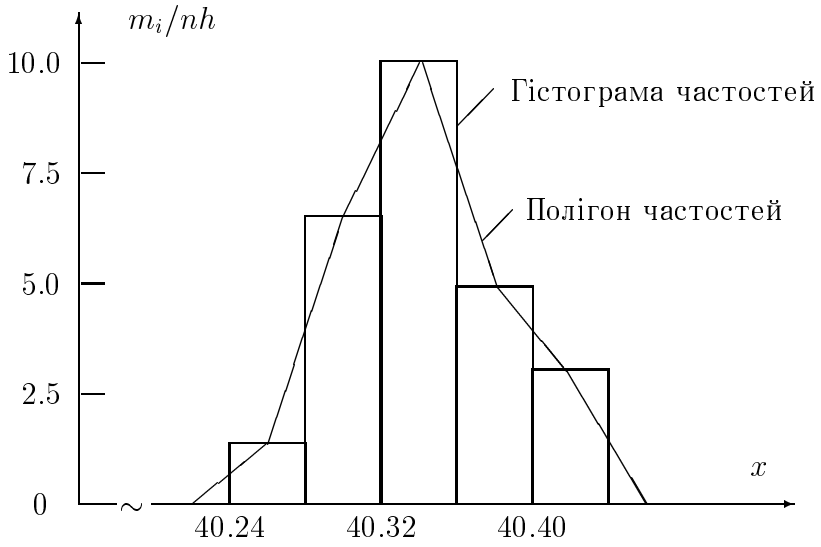


Рисунок 5.1 — Гістограма частостей і полігон частостей

Інтервали спостережених значень ВВ	[40,24; 40,28]	[40,28; 40,32]	[40,32; 40,36]	[40,36; 40,40]	[40,40; 40,44]
Частості m_i/n	0,0500	0,2379	0,4000	0,1875	0,1250
Накопичені частості $F^*(x)$	0,0500	0,2875	0,6875	0,8750	1,0000

Контроль: $\sum\limits_{i=1}^{5} m_i/n = 1.$

2. Для побудови гістограми частостей на осі $0x$ відкладаються часткові інтервали.

На кожному з цих інтервалів будується прямокутник, площа якого має дорівнювати частості даного часткового інтервалу. Якщо частості віднести до середин часткових інтервалів, то отримана замкнена лінія утворить *полігон* частостей. На рис. 5.1 зображена гістограма і полігон частостей.

3. Значення емпіричної функції розподілу виписані в останньому рядку статистичного ряду розподілу частостей.

Запишемо значення емпіричної функції розподілу в аналітичному вигляді:

$$F^*(x) = \begin{cases} 0,0000, & \text{якщо} \quad x \le 40,24; \\ 0,0500, & \text{якщо} \quad 40,24 < x \le 40,28; \\ 0,2875, & \text{якщо} \quad 40,28 < x \le 40,32; \\ 0,6875, & \text{якщо} \quad 40,32 < x \le 40,36; \\ 0,8750, & \text{якщо} \quad 40,36 < x \le 40,40; \\ 1,0000, & \text{якщо} \quad 40,40 < x \le 40,44; \\ 1,0000, & \text{якщо} \quad x > 40,44. \end{cases}$$

Зауваження. Значення емпіричної функції розподілу віднесені до верхньої межі часткового інтервалу.

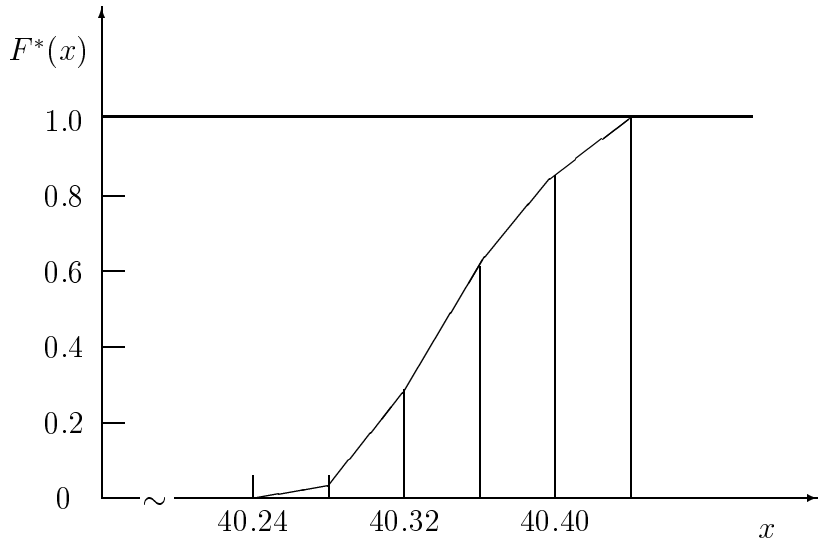


Рисунок 5.2 — Емпірична функція розподілу (кумулята)

Графік емпіричної функції зображений на рис. 5.2.

4. У тих випадках, коли значення випадкової величини, які спостерігаються, задаються багатозначними числами і обсяг вибірки досить великий ($n > 25$), обчислення досить об'ємні.

Тому спочатку визначимо x_i – середину i-го інтервалу ($i = 1, \ldots, 5$) і знайдемо середню арифметичну $\overline{x}$ за формулою

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{5} x_i m_i.$$

Визначимо:

$$q_{1i} = (x_i - \overline{x})\, m_i, \qquad q_{2i} = (x_i - \overline{x})^2\, m_i,$$
$$q_{3i} = (x_i - \overline{x})^3\, m_i, \qquad q_{4i} = (x_i - \overline{x})^4\, m_i$$

і перейдемо до обчислення центральних моментів порядку k ($k = 2, 3, 4$):

i	Інтервали значень ВВ X	x_i	m_i	q_{1i}	q_{2i}	q_{3i}	q_{4i}
1	40,24 – 40,28	40,26	4	–0,3360	0,0282	–0,0024	0,0002
2	40,28 – 40,32	40,30	19	–0,8360	0,0368	–0,0016	0,0001
3	40,32 – 40,36	40,34	32	–0,1280	0,0005	–0,0000	0,0000
4	40,36 – 40,40	40,38	15	+0,5400	0,0194	+0,0007	0,0000
5	40,40 – 40,44	40,42	10	+0,7600	0,0578	+0,0044	0,0003
	Сума		80	0,0000	0,1427	+0,0011	0,0006

Отже,

$$\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{5} m_i x_i \;=$$

$$=\; \frac{40,26\cdot 4 + 40,30\cdot 19 + 40,34\cdot 32 + 40,38\cdot 15 + 40,42\cdot 10}{80} \;= 40,344;$$

$$\hat{\mathrm{D}}[\mathrm{X}] \;=\; s_x^2 \;=\; \frac{1}{n}\sum_{i=1}^{5}(x_i-\overline{x})^2 m_i \;=\; \frac{0,1427}{80} \;=\; 0,001784;$$

$$s_x \;=\; \sqrt{0,001784} \;=\; 0,0422;$$

$$\hat{A} \;=\; \frac{1}{n s_x^3}\sum_{i=1}^{5}(x_i-\overline{x})^3 m_i \;=\; \frac{0,001}{80\cdot 0,042^3} \;=\; 0,1822;$$

$$\hat{E} \;=\; \frac{1}{n s_x^4}\sum_{i=1}^{5}(x_i-\overline{x})^4 m_i - 3 \;=\; \frac{0,0006}{80\cdot 0,042^4} - 3 \;=\; -0,5288;$$

$$\mathrm{Var} \;=\; \frac{s_x}{\overline{x}}\,100\,\% \;=\; \frac{0,0422}{40,344}\,100\,\% \;=\; 0,1047\,\%.$$

5. Для попереднього вибору закону розподілу ВВ обчислимо спочатку середні квадратичні помилки визначення асиметрії

$$s_{_A} \;=\; \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} \;=\; \sqrt{\frac{6\cdot 79}{81\cdot 83}} \;=\; 0,2655$$

і ексцесу

$$s_{_E} \;=\; \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+4)}} \;=\; \sqrt{\frac{24\cdot 80\cdot 78\cdot 77}{79^2\cdot 83\cdot 84}} \;=\; 0,5148.$$

Критерієм розподілу діаметрів отворів за нормальним законом є рівність нулю асиметрії і ексцесу. З наведених розрахунків видно, що значення вибіркових коефіцієнтів асиметрії $\hat{A}$ і ексцесу $\hat{E}$ відрізняються від нуля не більш ніж на подвоєні середні квадратичні помилки їх визначення, що відповідає нормальному закону розподілу.

Вигляд полігона і гістограми частостей також нагадує нормальну криву (криву Гаусса).

Можна передбачити, що діаметр отвору (випадкова величина X) змінюється під впливом великого числа чинників, приблизно рівнозначних за силою (зміна температури свердла або заготівки, вібрації заготівки, вібрації свердла, зміна механічних або хімічних властивостей заготівки і таке інше).

Тому, виходячи з "технології" побудови (випадкової величини X), тобто механізму утворення відхилень діаметрів отворів від деякого номінального значення, можна припустити, що закон розподілу діаметрів отворів є нормальним.

6. Густина імовірності нормального закону має вигляд

$$f(x) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

Знайдемо точкові оцінки $\hat{a}$ і $\hat{\sigma}$ параметрів a і σ нормального закону розподілу методом моментів:

$$\hat{a} \;=\; \overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{5} x_i m_i \;=\; 40,344\,(\text{мм});$$

$$\hat{\sigma} \;=\; s \;=\; \sqrt{\frac{1}{n}\sum_{i=1}^{5}(x_i-\overline{x})^2 m_i} \;=\; 0,0422\,(\text{мм}).$$

Отже, густина імовірності передбачуваного нормального закону розподілу має вигляд

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\,0,0422}\,\exp\left(-\frac{(x-40,344)^2}{2\cdot 0,0422^2}\right).$$

Функція розподілу передбачуваного нормального закону

$$F(x) \;=\; \frac{1}{\sqrt{2\,\pi}\,0,0422}\int\limits_{-\infty}^{x}\exp\left(-\frac{(x'-40,344)^2}{0,003567}\right)\,dx'.$$

Використовуючи функцію Лапласа $\Phi(x)$, функцію розподілу нормального закону можна записати у вигляді

$$F(x) \;=\; \frac{1}{2}+\frac{1}{2}\,\Phi\left(\frac{x-40,344}{0,0422}\right).$$

7. Проведемо детальну перевірку гіпотези про розподіл ВВ X (діаметра отворів) згідно з нормальним законом за допомогою критерію згоди χ^2.

Для цього інтервали спостережених значень віднормуємо, тобто виразимо їх в одиницях середнього квадратичного відхилення s:

$$u_i \;=\; \frac{x_i-\overline{x}}{s},$$

причому будемо вважати, що найменше значення u_i дорівнює $-\infty$, а найбільше дорівнює $+\infty$.

Далі обчислимо ймовірності влучення ВВ X, розподіленої згідно з нормальним законом з параметрами $a = 40,344$ та $\sigma = 0,042$, в часткові інтервали $[x_{i-1};\, x_i]$ за формулою

$$p_i \;=\; \Pr\left(x_{i-1} < \mathrm{X} < x_i\right) \;=\; \Phi(u_i)-\Phi(u_{i-1}),$$

де $\Phi(z)$ – функція Лапласа

$$\Phi(z) \;=\; \frac{1}{\sqrt{2\,\pi}}\int_0^z \exp(-t^2/2)\,dt.$$

Наприклад, імовірність того, що ВВ X (діаметр отворів) влучить в перший частковий інтервал $[-\infty;\, 40,28]$, дорівнює

$$p_1 \;=\; \Pr(-\infty < \mathrm{X} < 40,28) \;=$$

$$= \Phi\left(\frac{40,28-40,344}{0,0422}\right) - \Phi\left(\frac{-\infty-40,344}{0,0422}\right) =$$

$$= \Phi(\infty) - \Phi(1,5166) = \frac{1}{2} - 0,4352 = 0,0648.$$

Аналогічно

$$p_2 = \Pr(40,28 < \mathrm{X} < 40,32) =$$

$$= \Phi\left(\frac{40,32-40,344}{0,042}\right) - \Phi\left(\frac{40,28-40,344}{0,042}\right) =$$

$$= \Phi(-0,57) - \Phi(-1,5166) = 0,4352 - 0,2152 = 0,2200$$

і так само для p_3, p_4, p_5.

Після цього обчислимо теоретичні частоти нормального закону розподілу $n_{\text{теор}} = np_i$ і значення критерію χ^2

$$\chi^2_{\text{спос}} = \sum_{i=1}^{5} \frac{(m_i - np_i)^2}{np_i}.$$

Потім за таблицею квантилів χ^2-розподілу для рівня значущості $\alpha = 0,05$ і кількості ступенів вільності $\nu = k - r - 1$ ($k = 5$ – кількість інтервалів, $r = 2$ – кількість параметрів передбачуваного закону розподілу ВВ X знайдемо (див. додаток) критичне значення $\chi^2_{0,05;\nu}$.

Якщо $\chi^2_{\text{спос}} \le \chi^2_{0,05;\nu}$, то зробимо висновок, що немає підстав для відхилення гіпотези про розподіл діаметрів отворів згідно з нормальним законом.

У іншому випадку, тобто якщо $\chi^2_{\text{спос}} > \chi^2_{0,05;\nu}$, вважається, що гіпотеза розподілу діаметрів отворів згідно з нормальним законом не узгоджується з експериментальними даними.

Обчислення, необхідні для визначення значення $\chi^2_{\text{спос}}$ вибіркової статистики χ^2, зведені в таблицю, в якій визначено: $Q_{2i} = (m_i - np_i)^2$ та $R_{2i} = (m_i - np_i)^2/(np_i)$.

i	Інтервали спостережених значень X	Частоти m_i	Нормовані інтервали $u_i; u_{i+1}$	p_i	np_i	Q_{2i}	R_{2i}
1	40,24–40,28	4	$-\infty; -1,5152$	0,0648	5,1860	1,4067	0,2712
2	40,28–40,32	19	$-1,5152; -0,5682$	0,2200	17,6031	1,9514	0,1109
3	40,32–40,36	32	$-0,5682; 0,3788$	0,3626	29,0107	8,9356	0,3080
4	40,36–40,40	15	$0,3788; 1,3258$	0,2600	20,8013	33,6550	1,6179
5	40,40–40,44	10	$1,3258; \infty$	0,0925	7,3989	6,7659	0,9144
	Сума	80		1,0000	80,0000		3,2225

Зауваження. Найменше значення стандартизованої змінної, що використовується, $(40,24 - 40,344)/0,042 = -2,48$ замінено на $-\infty$; найбільше значення $(40,44 - 40,344)/0,042 = 2,29$ замінено на ∞.

Ця заміна зроблена для того, щоб сума теоретичних частот np_i дорівнювала обсягу вибірки.

Отже, внаслідок обчислень отримали $\chi^2_{\text{спос}} = 3,2225$.

Знайдемо тепер за таблицею квантилів χ^2-розподілу за рівнем значущості $\alpha = 0{,}05$ і кількістю ступенів вільності $\nu = k - r - 1 = 5 - 2 - 1 = 2$ критичне значення $\chi^2_{0{,}05;\,2} = 5{,}99$ (див. додаток).

Оскільки $\chi^2_{\text{спос}} = 3,2225 < 5,99$, то немає підстав для відхилення гіпотези про нормальний закон розподілу діаметрів отворів.

Далі перевіримо гіпотезу про розподіл діаметрів отворів згідно з нормальним законом за допомогою λ-критерію Колмогорова. З цією метою для кожного значення x_i знайдемо модуль різниці між емпіричною і теоретичною функціями розподілів $|F^*(x) - F(x)|$ і обчислимо значення вибіркової статистики λ Колмогорова, що спостерігається:

$$\lambda_{\text{спос}} \;=\; D\sqrt{n} \;=\; \max_x |F^*(x) - F(x)|\,\sqrt{n}.$$

Значення λ-статистики Колмогорова порівняємо з критичним значенням, що визначається за рівнем значущості $\alpha = 0{,}05$ (див. додаток).

Якщо $\lambda_{\text{спос}} \le \lambda_{0{,}05} = 1{,}358$, то вважається, що гіпотеза нормального розподілу досліджуваної випадкової величини узгоджується з експериментальними даними.

Якщо ж $\lambda_{\text{спос}} > \lambda_{0{,}05} = 1{,}358$, гіпотеза нормального розподілу не узгоджується з експериментальними даними.

Користуючись λ-критерієм згоди Колмогорова, перевіримо гіпотезу нормального розподілу діаметрів отворів. Всі допоміжні розрахунки, необхідні для обчислення вибіркової статистики $\lambda = D\sqrt{n}$, зведемо в таблицю, в якій визначено $m_{\text{н.е.ч}}$ – накопичені емпіричні частоти; $p_{\text{н.й}}$ – ймовірності.

i	Інтервали значень ВВ X	Частості m_i	$m_{\text{н.е.ч}}$	$p_{\text{н.й}}$	$F^*(x)$	$F(x)$	$D(x)$
1	40,24–40,28	4	4	0,064	0,0500	0,0648	0,0148
2	40,28–40,32	19	23	0,220	0,2875	0,2849	0,0026
3	40,32–40,36	32	55	0,364	0,6875	0,6475	0,0400
4	40,36–40,40	15	70	0,260	0,8750	0,9075	0,0325
5	40,40–40,44	10	80	0,092	1,0000	1,0000	0,0000
	Сума	80					

Переглядаючи останній стовпчик таблиці, помічаємо, що найбільший модуль різниці між емпіричною і теоретичною функціями розподілу складає

$$D \;=\; \max_x |F^*(x) - F(x)| \;=\; 0{,}040.$$

Обчислимо значення вибіркової статистики λ Колмогорова:

$$\lambda_{\text{спос}} \;=\; D\sqrt{n} \;=\; 0{,}040\,\sqrt{80} \;=\; 0{,}358.$$

Приймемо рівень значущості $\alpha = 0{,}05$.

З таблиць квантилів λ-розподілу Колмогорова (див. додаток) за рівнем значущості $\alpha = 0{,}05$ знаходимо величину – критичне значення $\lambda_{0{,}05} = 1{,}358$.

Оскільки $\lambda_{\text{спос}} = 0{,}358 < 1{,}358$, то немає підстави для відхилення гіпотези про нормальний закон розподілу діаметрів отворів.

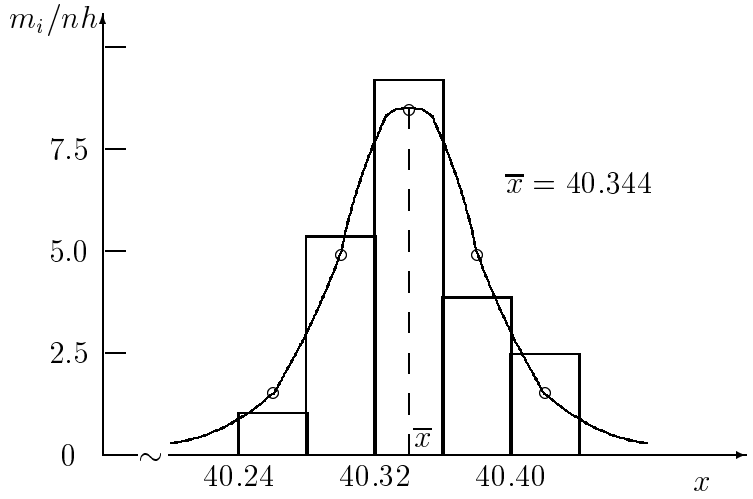


Рисунок 5.3 — Гістограма і нормальна крива

Для побудови нормальної кривої з середин часткових інтервалів відновимо перпендикуляри висотою p_i/h (p_i – ймовірність влучення випадкової величини X в частковий інтервал, h – довжина інтервалу).

На рис. 5.3 кінці цих перпендикулярів відмічені кулями. Отримані точки з'єднані плавною кривою.

Порівняння гістограми і нормальної кривої наочно показує, що нормальна крива добре згладжує гістограму відносних частот.

8. Знайдемо інтервальні оцінки параметрів нормального закону розподілу.

Для знаходження довірчого інтервалу, який містить математичне сподівання діаметрів отвору (ВВ X), знайдемо за таблицею квантилів розподілу Стьюдента (див. додаток) за заданою довірчою ймовірністю $\mathrm{P} = 1 - \alpha = 0,95$ і кількістю ступенів вільності $\nu = n - 1 = 80 - 1 = 79$ значення квантилю $t_{\alpha/2;\,\nu} = t_{0,025;\,79} = 1{,}99$.

Обчислимо межову похибку інтервального оцінювання

$$\varepsilon \;=\; t_{\alpha/2;\,\nu}\,\frac{s}{\sqrt{n}} \;=\; 1,99 \cdot \frac{0,042}{\sqrt{80}} \;=\; 0,009.$$

На основі формули $\overline{x} - \varepsilon \;<\; a \;<\; \overline{x} + \varepsilon$ знайдемо, що шуканий довірчий інтервал для математичного сподівання дорівнює

$$40,344 - 0,009 \;<\; a \;<\; 40,344 + 0,009.$$

Отже, отримуємо довірчий інтервал для математичного сподівання

$$40,335 \;<\; a \;<\; 40,353.$$

Значення отриманого результату: якщо буде зроблено досить велику кількість вибірок по 80 свердлувань отворів, то в 95% з них довірчий інтервал накриє математичне сподівання діаметра отвору і тільки у 5% випадків математичне сподівання може вийти за межі довірчого інтервалу.

Для знаходження довірчого інтервалу, який містить невідоме середнє квадратичне відхилення σ із заданою ймовірністю $\mathrm{P} = 1 - \alpha = 0{,}95$, знайдемо за таблицею

квантилів розподілу Стьюдента (див. додаток) за довірчою ймовірністю P$= 1-\alpha =$ 0,95 і кількістю ступенів вільності $\nu = n-1 = 80-1 = 79$ два числа: $\gamma_1 = 0{,}87$ і γ_2 = 1,18.

На основі формули $\gamma_1 s < \sigma < \gamma_2 s$ знайдемо, що шуканий довірчий інтервал дорівнює

$$0,87 \cdot 0,042 \ < \ \sigma \ < \ 1,18 \cdot 0,042.$$

Отже, отримуємо довірчий інтервал для середнього квадратичного відхилення

$$0,037 \ < \ \sigma \ < \ 0,050.$$

Отриманий результат означає, що якщо буде зроблено досить велику кількість вибірок по 80 свердлувань отворів, то в 95 % з них довірчий інтервал накриє середнє квадратичне відхилення σ і тільки в 5 % середнє квадратичне відхилення σ може вийти за межі довірчого інтервалу.

5.7. Приклади

Приклад 5.1

Результати дослідження міцності на стиснення (випадкова величина X) 200 зразків бетону подані у вигляді згрупованого статистичного ряду.

i	Інтервали міцності, *кг/см*^{2}	Частоти m_i
1	190–200	10
2	200–210	26
3	210–220	56
4	220–230	64
5	230–240	30
6	240–250	14

Потрібно перевірити нульову гіпотезу про нормальний закон розподілу міцності на стиснення. Рівень значущості прийняти $\alpha = 0{,}05$.

Розв'язання

З умови випливає, що точні параметри гіпотетичного нормального закону нам невідомі, тому нульову гіпотезу словесно можна сформулювати таким чином: $\{H_0$: $F(x)$, тобто $F(x)$є*функцією нормального закону розподілу*$\}$ з параметрами M[X]$= \hat{a}$ та D[X]$= \hat{\sigma}^2 = s^2$.

Для перевірки цієї нульової гіпотези визначимо значення x_i^* середин інтервалів і знайдемо точкові оцінки математичного сподівання і середнього квадратичного відхилення нормально розподіленої випадкової величини за формулами ($n = 200$):

$$\hat{a} \ = \ \overline{x} \ = \ \frac{1}{200} \sum_{i=1}^{6} x_i^* m_i \ =$$

$$= \frac{195 \cdot 10 + 205 \cdot 26 + 215 \cdot 56 + 225 \cdot 64 + 235 \cdot 30 + 245 \cdot 14}{200} = 221 \ кг/см^2;$$

$$\hat{\sigma}^2 \;=\; s^2 \;=\; \frac{1}{200}\sum_{i=1}^{6}(x_i^* - \overline{x})^2\, m_i \;=$$

$$=\; \frac{1}{200}[(-26)^2\cdot 10 + (-16)^2\cdot 26 + (-6)^2\cdot 56 + 4^2\cdot 64 + 14^2\cdot 30 + 24^2\cdot 14] \;=\; 152;$$

$$\hat{\sigma} \;=\; s \;=\; \sqrt{152} \;=\; 12,33 \;\textit{кг/см}^2.$$

Обчислимо теоретичні ймовірності p_i влучення випадкової величини X в часткові інтервали $[x_{i-1};\, x_i]$ за формулою

$$p_i \;=\; \Pr\left(x_{i-1} \le \mathrm{X} < x_i\right) \;=\; \Phi(u_i) - \Phi(u_{i-1}); \qquad (i = 1, 2, \ldots, k),$$

де

$$u_i \;=\; \frac{x_i^* - \overline{x}}{s}; \qquad \Phi(u) \;=\; \frac{1}{\sqrt{2\pi}}\int_0^u \exp(-t^2/2)\, dt.$$

Подальші обчислення, необхідні для визначення значення вибіркової статистики χ^2, зведені в таблицю, в якій позначено $Q_i = (m_i - np_i)^2$ та $R_i = (m_i - np_i)^2/np_i$.

i	Інтервали зміни	m_i	p_i	np_i	Q_i	R_i
1	190–200	10	0,0443	8,8507	1,3208	0,1492
2	200–210	26	0,1419	28,3769	5,6496	0,1991
3	210–220	56	0,2815	56,3078	0,0947	0,0017
4	220–230	64	0,2996	59,9254	16,6026	0,2771
5	230–240	30	0,1711	34,2101	17,7248	0,5181
6	240–250	14	0,0617	12,3292	2,7917	0,2264
	Суми:	200	1,0000	200,00	44,1842	1,3716

Внаслідок обчислень знаходимо: $\chi^2_{\text{спос}} = 1,37$.

Зауваження. Оскільки випадкова величина X, що розподілена згідно з нормальним законом, визначена на $(-\infty;\, \infty)$, то найменше значення стандартизованої змінної $(190 - 221)/12,33 \;=\; -2,51$ замінено на $-\infty$; найбільше значення $(250 - 221)/12,33 = 2,35$ замінено на $+\infty$.

За таблицею квантилів χ^2-розподілу за заданим рівнем значущості $\alpha = 0,05$ і кількістю ступенів вільності $\nu = k - r - 1 = 6 - 2 - 1 = 3$ знайдемо критичне значення $\chi^2_{0,05;\,3} \;=\; 7,82$.

Оскільки $\chi^2_{\text{спос}} = 1,37 < 7,82$, то немає підстав для відхилення нульової гіпотези про нормальний закон розподілу межі міцності на стиснення з параметрами $a = 221$ та $\sigma^2 = 152$.

Приклад 5.2

Є дані про розподіл товщини X 12000 бобів (в *мм*):

Номер інтервалу	1	2	3	4	5	6	7	8
Частота	32	103	239	624	1187	1650	1883	1930
Номер інтервалу	9	10	11	12	13	14	15	16
Частота	1638	1130	737	427	221	110	57	32

Тут перший інтервал — значення X, менші 7,00 *мм*, другий — 7,00–7,25, третій — 7,25–7,50 і т.д.

Потрібно перевірити, чи узгодиться товщина бобів у вибірці з припущенням, що ця ознака в генеральній сукупності розподілена згідно з нормальним законом?

Розв'язання

Оскільки в даних до прикладу відсутні відомості про параметри нормального розподілу, то за параметр $\overline{\mathrm{X}}$ використаємо $\tilde{\mathrm{X}}$, значення якого в цьому випадку дорівнює 8,562, а за генеральну дисперсію — її незсунену оцінку

$$s^2 \;=\; \frac{n}{n-1}\,\sigma^2 \;=\; 0,3833,$$

звідки $s = 0{,}6191$.

Знаходимо величини np_i для кожного з інтервалів.

Покажемо, як це робиться на прикладі другого інтервалу:

$$np_i \;=\; 12000 \cdot \Pr(7,00 \le \mathrm{X} \le 7,25) \;=$$

$$=\; 12000 \cdot \left[\Phi\left(\frac{7,25-8,712}{0,6163}\right) - \Phi\left(\frac{7,00-8,712}{0,6163}\right)\right] \;=\; 134,8.$$

Таким чином, отримаємо наступний ряд теоретичних частот:

69,9; 134,8; 313,5; 620,8; 1046,6; 1502,3; 1836,0; 1910,3;

1692,3; 1276,3; 819,6; 448,1; 208,6; 82,6; 27,8; 10,5.

Тому

$$\chi_0^2 \;=\; \frac{(32-69,9)^2}{49,3} + \frac{(103-134,8)^2}{134,8} + \cdots + \frac{(57-27,8)^2}{27,8} + \frac{(32-10,5)^2}{8,1},$$

що дає $\chi_0^2 \;=\; 192,99$.

Два параметри генерального розподілу оцінені за вибіркою, тому $s = 2$, а кількість ступенів вільності дорівнює $k - s - 1 = 16 - 2 - 1 = 13$. При $\alpha = 0{,}01$ межою критичної області є $\chi^2 = 27{,}688$.

Оскільки $\chi^2_{\text{спос}} = 192,99 > 27,688$, то гіпотеза про те, що випадкова величина X (товщина бобів) розподілена згідно з нормальним законом, не підтвердилася.

Приклад 5.3

У перших двох стовпчиках таблиці наведені дані про відмови апаратури за 10000 годин роботи. Загальне число обстежених примірників апаратури $n = 757$; при цьому спостерігалася $0 \cdot 427 + 1 \cdot 235 + 2 \cdot 72 + 3 \cdot 21 + 4 \cdot 1 + 5 \cdot 1 \;=\; 451$ відмова.

Кількість відмов k	Кількість випадків з k відмовами n_k
0	427
1	235
2	72
3	21
4	1
5	1

Прийнявши рівень значущості $\alpha = 0,01$, перевірити гіпотезу про те, що кількість відмов має розподіл Пуассона:

$$p_k = \Pr[\mathrm{X} = k] = \frac{\lambda^k}{k!}\exp(-\lambda), \qquad k = 0, 1, 2, \ldots.$$

Розв'язання

За оцінку $\hat{\lambda}$ параметра λ використовуємо середню кількість відмов:

$$\hat{\lambda} = 451/757 = 0,596.$$

За таблицею розподілу Пуассона з $\lambda = 0{,}596$ знаходимо ймовірності p_k і очікувану кількість випадків з k відмовами (третій і четвертий стовпчики таблиці).

Кількість відмов k	Кількість випадків з k відмовами n_k	Ймовірність $p_k = \frac{0{,}596^k}{k!}\exp(-0,596)$	Очікувана кількість випадків з k відмовами np_k
0	427	0,5511	417
1	235	0,3284	249
2	72	0,0978	74
3	21	0,0194	15
4	1	0,0029	2
5	1	0,0003	0

Для $k = 4$ і 5 значення $np_k < 5$, тому об'єднуємо ці рядки з рядком для $k = 3$. У результаті набуваємо значень, наведених у таблиці:

k	n_k	np_k	$(n_k - np_k)^2/np_k$
0	427	417	0,230
1	235	249	0,740
2	72	74	0,056
≥ 3	23	17	1,991

Маємо

$$\chi^2_{\text{спос}} = \sum_{k=0}^{3} \frac{(n_k - np_k)^2}{np_k} = 3,017.$$

Оскільки за вибіркою оцінювався один параметр λ розподілу, то $l = 1$, і тому кількість ступенів вільності дорівнює $4 - 1 - 1 = 2$. За таблицею квантилів χ^2-розподілу знаходимо $\chi^2_{0,99;\,2} = 9{,}21$.

Отже, маємо $\chi^2_{\text{спос}} < \chi^2_{0,99;\,2}$. Тому при даному рівні значущості гіпотеза про розподіл кількості відмов згідно із законом Пуассона приймається.

Приклад 5.4

За допомогою вимірювального приладу було проведено 200 вимірювань заданої відстані. Випадкова похибка вимірювань записана в мікрометрах. Дійсна вісь була розділена на 9 проміжків, результати (випадкова похибка вимірювань) зведені в таблицю (дробові значення m_i в ній з'явилися через те, що значення, які влучили на межу інтервалу, прийнято записувати порівну як одному, так і іншому інтервалу).

Номер інтервалу	Інтервал, *мкм*	Частота m_i
1	менше –15	6
2	від –15 до –10	11,5
3	від –10 до –5	15,5
4	від –5 до 0	22
5	від 0 до 5	47,5
6	від 5 до 10	42
7	від 10 до 15	28
8	від 15 до 20	17
9	більше 20	10,5

Потрібно перевірити нульову гіпотезу H_0 про те, що *випадкова похибка вимірювання* X *розподілена нормально.*

Розв'язання

Оскільки ширина каналів дорівнює $5\,мкм$, приймем для 1-го та 9-го каналів значення $x_1 = -17,5\,мкм$ та $x_9 = 22,5\,мкм$ відповідно.

Необхідні дії виконаємо в наступному порядку:

а) розглянемо гіпотетичний (нормальний) закон розподілу;

б) за заданою вибіркою отримаємо найбільш правдоподібні значення параметрів розподілу;

в) побудуємо критерій χ^2, який використовуємо для перевірки гіпотези про нормальність випадкової похибки вимірювання X.

Гіпотетичний розподіл має густину

$$f_{\rm X}(x;a,\sigma) \;=\; \frac{1}{\sqrt{2\pi\,\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right),$$

яка містить два невідомих параметри a і σ.

Побудуємо допоміжну таблицю статистичного розподілу випадкової похибки вимірювань. Для цього за вибіркою, яка містить $n = 200$ вимірювань, обчислимо оцінки для a і σ, що дає

$$a \Rightarrow \overline{x}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i m_i \;=\; 4,49мкм, \qquad \sigma^2 \Rightarrow \sigma^{*2} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}^*)^2 m_i \;=\; 90,86мкм^2.$$

Таким чином, функція $f^*(x)$, що використовується для перевірки, має вигляд

$$f^*(x) \;=\; \frac{1}{\sqrt{2\pi\cdot 90,86}}\exp\left(-\frac{(x-4,49)^2}{2\cdot 90,86}\right).$$

Звідси за таблицею функції Лапласа можна знайти ймовірності p_i влучення в i-й інтервал. Наприклад, при $i = 1$ та $i = 2$

$$p_1 \;=\; \Pr(\mathrm{X} < -15) \;=\; \Pr\left(\frac{\mathrm{X}-4,49}{\sqrt{90,86}} < \frac{-15-4,49}{\sqrt{90,86}}\right) \;=$$

$$=\; \frac{1}{2} - \Phi(2,0444) \;=\; \frac{1}{2} - 0,4795 \;=\; 0,0204,$$

аналогічно

$$p_2 \;=\; \Pr(-15 < \mathrm{X} < -10) \;=\; \Pr\left(-2,0447 < \frac{\mathrm{X}-4,49}{\sqrt{90,86}} < -1,5201\right) \;=$$

$$= \;\Phi(-1,5199) - \Phi(-2,0444) \;=\; \Phi(2,0444) - \Phi(1,5199) \;=\; 0,0439$$

і так далі ($i = 3, \ldots, 9$).

У результаті будуємо таблицю числових даних, необхідних для знаходження значення критерію χ^2 (див. таблицю розрахункових даних).

i	Частота m_i	p_i	np_i	$(m_i - np_i)^2/np_i$
1	6	0,0204	4,088	0,894
2	11,5	0,0438	8,763	0,855
3	15,5	0,0955	19,104	0,680
4	22	0,1591	31,822	3,032
5	47,5	0,2025	40,508	1,207
6	42	0,1970	39,406	0,171
7	28	0,1465	29,296	0,057
8	17	0,0832	16,644	0,008
9	10,5	0,0518	10,383	0,002
Сума	200,0	1,0000	200,00	$\chi^2 = 6,905$

Таким чином, обчислення дає $\chi^2 = 6,905$.

Оскільки проводиться оцінка двох параметрів, то маємо в цьому випадку $k = 9 - 2 - 1 = 6$ ступенів вільності. Приймемо рівень значущості $\alpha = 0,05$, тоді з таблиці квантилів χ^2-розподілу (див. додаток) отримуємо $\chi^2_{0,05;\,6} = 12,6$.

Обчислене значення χ^2-критерію $\chi^2 = 6,905$ виявилося менше квантилю $\chi^2_{0,05;\,6}$.

Таким чином, гіпотеза про нормальний розподіл випадкової похибки вимірювання приладу не суперечить результатам спостережень на рівні значущості 0,05.

Приклад 5.5

В цеху з поточної продукції токарного верстата-автомата, що налагоджений на опрацювання заданої деталі, взяті дві вибірки обсягом $n_1 = 150$ та $n_2 = 100$. Перша вибірка вироблена на початку зміни, а друга – після двох годин роботи верстата. Результати вимірювань відхилень від номіналу розміру (випадкова величина X), що контролюється, в мікрометрах наведені в таблиці.

Інтервали зміни ВВ X, *мкм*	Частота у вибірці № 1, m_{1j}	Частота у вибірці № 2, m_{2j}
–15; –10	10	0
–10; –5	27	7
–5; 0	43	17
0; 5	38	30
5; 10	23	29
10; 15	8	15
15; 20	1	1
20; 25	0	1
Сума :	$n_1 = 150$	$n_2 = 100$

Потрібно за допомогою λ-критерію Смірнова-Колмогорова (модель 2) перевірити нульову гіпотезу про те, що розподіл похибок опрацювання верстатом-автоматом протягом 1 години описується однією і тією ж функцією розподілу. Прийняти рівень значущості $\alpha = 0,05$.

Розв'язання

Згідно з умовою, треба перевірити нульову гіпотезу $\{H_0 : F_1(x) = F_2(x)\}$ (*процес виготовлення деталей верстата-автомата є стійким у часі*).

Обчислення вибіркової статистики $D^* = \max\limits_x |F_1^*(x) - F_2^*(x)|$ наведене нижче в таблиці, в якій позначено m_{1j} та m_{2j} – частоти, n_{1j} та n_{2j} — накопичені частоти, $F_1^*(x) = n_1(x)/n_1$ та $F_2^*(x) = n_2(x)/n_2$.

x_{i+1}	m_{1i}	m_{2i}	$n_1(x)$	$n_2(x)$	$F_1^*(x)$	$F_2^*(x)$	$\lvert F_1^*(x) - F_2^*(x)\rvert$
–10	10	0	10	0	0,067	0,000	0,067
–5	27	7	37	7	0,247	0,070	0,177
0	43	17	80	24	0,533	0,240	0,293
5	38	30	118	54	0,787	0,540	0,247
10	23	29	141	83	0,940	0,830	0,110
15	8	15	149	98	0,993	0,980	0,013
20	1	1	150	99	1,000	0,990	0,010
25	0	1	150	100	1,000	1,000	0,000

Аналізуючи останній стовпчик даної таблиці, помічаємо, що найбільший модуль різниці $F_1^*(x) - F_2^*(x)$ між емпіричними функціями розподілу $F_1^*(x)$ і $F_2^*(x)$ дорівнює

$$D^* = \max_x |F_1^*(x) - F_2^*(x)| = 0,293.$$

Оскільки

$$n = \frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{150 \cdot 100}{150 + 100} = 60,$$

то значення вибіркової статистики, що спостерігається,

$$\lambda_{\text{спос}} = D^* \sqrt{n} = 0,293 \sqrt{60} = 2,272.$$

За таблицею квантилів λ-розподілу Смірнова-Колмогорова для заданого рівня значущості α знайдемо критичне значення $\lambda_{0,05} = 1,358$.

Оскільки $\lambda_{\text{спос}} = 2,272 > 1,358$, нульову гіпотезу потрібно відхилити.

Таким чином, не можна стверджувати, що похибки опрацювання за інтервал часу, що досліджується, описуються однією і тією ж функцією розподілу. Іншими словами, отриманий результат говорить про те, що процес виготовлення деталей верстатом-автоматом не є стійким у часі.

Приклад 5.6

При дослідженні межі пластичності 15 зразків визначеного сорту сталі отримані наступні результати (в *кг/см*^{2}):

3540, 3580, 3570, 3560, 3500, 3610, 3720,

3640, 3600, 3650, 3750, 3590, 3600, 3550, 3770.

З метою збільшення межі пластичності був проведений додатковий технологічний процес. Отримано для тих же зразків наступні результати:

$$3580,\ 3570,\ 3680,\ 3880,\ 3530,\ 3680,\ 3730,$$
$$3720,\ 3670,\ 3710,\ 3810,\ 3660,\ 3770,\ 3640,\ 3670.$$

Перевірити за допомогою критерію знаків гіпотезу, що межа пластичності сорту сталі при проведенні додаткового технологічного процесу збільшилася. Рівень значущості прийняти в даній задачі $\alpha = 0,05$.

Розв'язання

Сформулюємо нульову гіпотезу $\{H_0 : F(x) = F(y)\}$ – *межа пластичності сорту сталі, що досліджується, не змінюється при проведенні додаткового технологічного процесу.* Визначимо знаком "+" зростання межі пластичності, а знаком "–" – зменшення межі пластичності, викликане додатковим процесом. Отримуємо наступну послідовність знаків:

$$+\ -\ +\ +\ +\ +\ +\ +\ +\ +\ +\ +\ +\ +\ -$$

Кількість знаків "–" $r = 2$.

Знайдемо за таблицею критичних значень кількості знаків (див. додаток) за заданим рівнем значущості $\alpha = 0,05$ і обсягу вибірки $n = 15$ критичне значення $r_{\alpha,\,n} = r_{0,05;\,15} = 3$.

Оскільки $r_{\text{спос}} = 2 < 3$, то нульова гіпотеза відхиляється. Іншими словами, вважається статистично встановленим, що додатковий процес призводить до збільшення межі пластичності даного сорту сталі.

Приклад 5.7

Протягом 100 днів фіксувалася кількість аварій водогінної мережі в деякому районі міста. Отримані наступні числові дані:

Кількість аварій (ВВ X)	0	1	2	3	4	5
Частоти m_i	8	28	31	18	9	6

$$n = \textstyle\sum_i m_i = 100$$

Потрібно перевірити гіпотезу про те, що розподіл кількості аварій водогінної мережі міста підкоряється закону Пуассона. Рівень значущості прийняти $\alpha = 0,05$.

Розв'язання

Згідно з умовою, необхідно перевірити нульову гіпотезу $\{H_0$: *функція розподілу* $F(x)$ *кількості аварій має вигляд*$\}$

$$F(x;\lambda) \;=\; \sum_{i=0}^{x} \frac{\lambda^i}{i!} \exp(-\lambda)$$

з параметром

$$\hat{\lambda} \;= \overline{x} \;=\; \frac{1}{n} \sum_{i=0}^{5} m_i x_i \;=\; \frac{8 \cdot 0 + 28 \cdot 1 + 31 \cdot 2 + 18 \cdot 3 + 9 \cdot 4 + 6 \cdot 5}{100} \;=\; 2,1.$$

Обчислимо теоретичні ймовірності p_i появи рівно x_i аварій протягом n днів за формулою Пуассона:

$$p_i \;=\; \mathrm{P}_n(x_i) \;=\; \frac{1}{x_i!}\,\lambda^{x_i}\exp(-\lambda), \qquad x_i = 0,1,2,3,4,5.$$

Подальші обчислення зводимо в таблицю.

Кількість аварій x_i	m_i	p_i	np_i	$(m_i - np_i)^2$	$(m_i - np_i)^2/np_i$
0	8	0,1225	12,25	18,03	1,47
1	28	0,2572	25,72	5,22	0,20
2	31	0,2700	27,00	15,99	0,59
3	18	0,1890	18,90	0,81	0,04
4	9	0,0992	9,92	0,85	0,09
≥ 5	6	0,0621	6,21	0,05	0,01
Суми:	100	1,000	100,0	44,25	2,40

Внаслідок обчислень знаходимо $\chi^2_{\text{спос}} = 2,40$.

За таблицею квантилів χ^2-розподілу за заданим рівнем значущості $\alpha = 0{,}05$ та кількістю ступенів вільності $\nu = k - r - 1 = 6 - 1 - 1 = 4$ знайдемо критичне значення $\chi^2_{0{,}05;\,4} = 9,49$.

Оскільки $\chi^2_{\text{спос}} = 2,40 < 9,49$, то немає підстав для відхилення гіпотези про те, що закон розподілу кількості аварій водогінної мережі є законом Пуассона з параметром $\lambda = 2,1$.

Приклад 5.8

Комплектуючі вироби одного найменування надходять з трьох підприємств: A, B і C. Результати перевірки виробів (для постачальників A, B і C) наведені в наступній таблиці:

Результати перевірки	A	B	C	Всього (v_i)
Придатні	29	38	53	120
Непридатні	1	2	7	10
Всього (u_j)	30	40	60	130

Прийнявши рівень значущості $\alpha = 0{,}10$, з'ясувати, чи можна вважати, що якість виробів не залежить від постачальника?

Розв'язання

Перевіримо гіпотезу H_0 про незалежність двох ознак: якості виробу X і місця його виготовлення Y. Для цього за критерієм χ^2 використовуємо статистику

$$\chi^2_{\text{спос}} \;=\; \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \tag{1}$$

Тут n_{ij} – число виходів, в яких реалізовувалася подія $\{\mathrm{X} = x_i \text{ та } \mathrm{Y} = y_j\}$; $\tilde{n}_{ij} = n\tilde{p}_{ij}$ – очікувані частоти, $\tilde{p}_{ij}$ – очікувані частості.

В рамках гіпотези про незалежність ознак X та Y маємо $p_{ij} = p_i p_j$, де p_i – ймовірності влучення X в i-й інтервал $(i = 1,\ldots,k)$, p_j — ймовірності влучення Y в

j-й інтервал ($j = 1, \ldots, l$), відповідно. З виразу (1) маємо

$$\chi^2 \;=\; \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{\tilde{n}_{ij}} - 2\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{l}\tilde{n}_{ij}.$$

В цьому виразі другий доданок дорівнює $2n$, третій доданок дорівнює n, тому

$$\chi^2 \;=\; \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{np_{ij}} - n \;=\; n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{n^2 p_i p_j} - 1\right). \tag{2}$$

Приймемо для очікуваних ймовірностей, що $p_i = v_i/n$ та $p_j = u_j/n$, тоді після перетворень в виразі (2), запишемо

$$\chi^2_{\text{набл}} \;=\; n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}^2}{v_i u_j} - 1\right). \tag{3}$$

За умови, що гіпотеза H_0 вірна і для всіх очікуваних частот виконується $\tilde{n}_{ij} \geq 4$, статистика (1) має розподіл χ^2 з $(k-1)(l-1)$ ступенями вільності. Параметри k і l дорівнюють $k = 3$ і $l = 2$, так що повна кількість ступенів вільності становить $(k-1)(l-1) = 2$.

Користуючись формулой (2), знайдемо

$$\chi^2_{\text{спос}} \;=\; 130\cdot\left(\frac{29^2}{30\cdot 120} + \frac{38^2}{40\cdot 120} + \frac{53^2}{60\cdot 120} + \frac{1^2}{30\cdot 10} + \frac{2^2}{40\cdot 10} + \frac{7^2}{60\cdot 10} - 1\right) \approx 2,546.$$

За таблицею квантилів χ^2-розподілу знаходимо $\chi^2_{0,90;\,2} = 4{,}605$.

Оскільки $\chi^2_{\text{спос}} < \chi^2_{0,90;\,2}$, то при даному рівні значущості потрібно вважати, що якість виробів не залежить від постачальника.

5.8. Задачі для розв'язання

Задача 5.1

За рік на деякий район упало 537 невеликих метеоритів. Вся територія району була розділена на 576 дільниць площею по 0,25 *км*^{2} кожна. Нижче наведені кількості дільниць n_k, на які впало k метеоритів:

k	0	1	2	3	4	5 і більше
n_k	229	211	93	35	7	1

Чи узгоджуються ці дані з гіпотезою про те, що кількість метеоритів, які упали на кожну з дільниць, має розподіл Пуассона? Прийняти $\alpha = 0,10$.

Задача 5.2

Амплітуда коливань визначалася двома лаборантами. Перший лаборант по 10 спостереженнях набув середнього значення амплітуди $\overline{x}_1 = 81$ *мм*, а другий лаборант по 15 спостереженнях отримав $\overline{x}_2 = 84$ *мм*.

У припущенні, що дисперсії вимірювань двома лаборантами відомі і дорівнюють $\sigma_1^2 = 64\,мм^2$ та $\sigma_2^2 = 64\,мм^2$ для першого і другого відповідно, знайти 99 %-й довірчий інтервал для різниці середніх $\overline{X}_1$ і $\overline{X}_2$. Чи можна вважати, що результати лаборантів дійсно різняться ?

Задача 5.3

Отримані наступні дані (див. таблицю).

40,25	40,37	40,33	40,28	40,29	40,41	40,35	40,28	40,29	40,27
40,35	40,35	40,41	40,30	40,33	40,40	40,34	40,46	40,39	40,38
40,45	40,44	40,35	40,40	40,31	40,33	40,34	40,32	40,39	40,37
40,39	40,30	40,33	40,32	40,36	40,34	40,43	40,31	40,37	40,36
40,40	40,34	40,38	40,32	40,34	40,30	40,36	40,31	40,38	40,35
40,42	40,31	40,33	40,42	40,30	40,43	40,34	40,36	40,36	40,32
40,35	40,35	40,30	40,36	40,33	40,37	40,31	40,34	40,37	40,37
40,32	40,32	40,33	40,35	40,30	40,34	40,34	40,34	40,41	40,36

За допомогою статистичного λ–критерію Колмогорова перевірити гіпотезу про те, що наведена вибірка витягнута з генеральної сукупності, яка рівномірно розподілена на інтервалі $(40,238;\ 40,462)$.

Задача 5.4

Отримані наступні дві групи даних (див. таблиці).

40,25	40,37	40,33	40,28	40,29	40,41	40,35	40,28	40,29	40,27
40,35	40,35	40,41	40,30	40,33	40,40	40,34	40,46	40,39	40,38
40,45	40,44	40,35	40,40	40,31	40,33	40,34	40,32	40,39	40,37
40,39	40,30	40,33	40,32	40,36	40,34	40,43	40,31	40,37	40,36

40,40	40,34	40,38	40,32	40,34	40,30	40,36	40,31	40,38	40,35
40,42	40,31	40,33	40,42	40,30	40,43	40,34	40,36	40,36	40,32
40,35	40,35	40,30	40,36	40,33	40,37	40,31	40,34	40,37	40,37
40,32	40,32	40,33	40,35	40,30	40,34	40,34	40,34	40,41	40,36

За допомогою критерію Колмогоров–Смірнова перевірити гіпотезу про те, що обидві ці наведені вибірки витягнуті з однієї і тієї ж генеральної сукупності.

Задача 5.5

Ставлення глядачів до включення даної передачі в програму виразилося наступними даними :

	Позитивне	Негативне	Байдуже
Чоловіки	14	24	2
Жінки	29	36	15

Чи можна вважати, що ставлення до включення даної передачі в програму не залежить від статі глядача ? Прийняти $\alpha = 0,10$.

Задача 5.6

Випробовувалася чутливість 40 приймачів. Дані наведені в таблиці; в першому рядку – інтервали чутливості в мікровольтах, у другому – середні точки цих

інтервалів $f_{\text{ср}}$, в третьому – кількість приймачів n_i, чутливість яких виявилася в цьому інтервалі.

Інтервал	25–75	75–125	125–175	175–225	225–275
$f_{\text{ср}}$	50	100	150	200	250
n_i	0	0	1	5	9
Інтервал	275–325	325–375	375–425	425–475	475–525
$f_{\text{ср}}$	300	350	400	450	500
n_i	6	8	6	2	2
Інтервал	525–575	575–625	625–675	675–725	725–775
$f_{\text{ср}}$	550	600	650	700	750
n_i	0	1	1	0	0

За допомогою χ^2-критерію перевірити гіпотезу про те, що вибірка витягнута з нормальної сукупності.

Задача 5.7

Стверджується, що результат дії ліків залежить від способу їх застосування. Перевірити це твердження при $\alpha = 0{,}05$ за наступними даними.

Результат	Спосіб 1	Спосіб 2	Спосіб 3
Несприятливий	11	17	16
Сприятливий	20	23	19

Задача 5.8

Нижче наводиться час (в секундах) розв'язання контрольних задач учнями до і після спеціальних вправ по усному рахунку.

До вправ	87	61	98	90	74	83	72	81	75	83	85
Після вправ	50	45	79	88	65	52	79	84	61	52	85

Чи можна вважати, що ці вправи поліпшили здатність учнів в розв'язанні задач? Прийняти $\alpha = 0,10$.

Задача 5.9

Передбачається, що застосування нової технології у виробництві мікросхем призведе до збільшення виходу придатної продукції. Результати контролю двох партій продукції, виготовлених за старою і новою технологіями, наведені нижче:

Вироби	Стара технологія	Нова технологія
Придатні	140	185
Непридатні	10	15
Всього	150	200

Чи підтверджують ці результати припущення про збільшення виходу придатної продукції? Прийняти $\alpha = 0,10$.

Задача 5.10

При 50 підкиданнях монети "герб" з'явився 20 разів.

Чи можна вважати монету симетричною? Прийняти $\alpha = 0,10$.

Задача 5.11

Передбачається, що один з двох приладів, які визначають швидкість автомобіля, має систематичну помилку (завищення). Для перевірки цього припущення визначили швидкість 10 автомобілів, причому швидкість кожного з них фіксувалася одночасно двома приладами.

У результаті отримані наступні дані:

v_1, *км/годин*	70	85	63	54	65	80	75	95	52	55
v_2, *км/годин*	72	86	62	55	63	80	78	90	53	57

Чи дозволяють ці дані стверджувати, що другий прилад дійсно дає завищені значення швидкості? Прийняти $\alpha = 0{,}10$.

Задача 5.12

У лабораторії при випробуванні радіоелектронної апаратури фіксувалася деяка кількість відмов. Результати 59 випробувань наводяться нижче:

Кількість відмов	0	1	2	3
Кількість випробувань	42	10	4	3

Перевірити гіпотезу H_0 про те, що кількість відмов має розподіл Пуассона, прийнявши $\alpha = 0{,}10$.

Задача 5.13

Технологія виробництва деякої речовини дає в середньому 1000 *кг* речовини за добу з СКВ середнього $\sigma^* = 80$ *кг*. Нова технологія виробництва в середньому дає 1100 *кг* речовини за добу з тим же СКВ середнього.

Чи можна вважати, що нова технологія забезпечує підвищення продуктивності, якщо: а) $\alpha = 0,05$; б) $\alpha = 0,10$?

Задача 5.14

Метод отримання випадкових чисел був застосований 250 разів. При цьому були отримані наступні результати:

Цифра	0	1	2	3	4	5	6	7	8	9
Частота появи	27	18	23	31	21	23	28	25	22	32

Чи можна вважати, що застосований метод дійсно дає випадкові числа? Прийняти $\alpha = 0{,}05$.

Задача 5.15

Експериментатор при $n = 4040$ киданнях монети отримав $m = 2048$ випадань герба.

Чи сумісне це з гіпотезою про те, що існує стала ймовірність $p = 0{,}5$ випадання герба?

Задача 5.16

Годинники, виставлені у вітринах годинникових магазинів, показують випадковий час. Пропонується гіпотеза, що свідчення цих годинників у вітринах великої кількості магазинів розподілені рівномірно в інтервалі $(0; 12)$. Спостереження 500

вітрин 500 магазинів дали наступну вибірку (весь інтервал $(0;12)$ розбитий на 12 часових інтервалів):

Час	0	1	2	3	4	5	6	7	8	9	10	11
Кількість вибіркових значень	41	34	54	39	49	45	41	33	37	41	47	39

Чи узгодяться ці дані з гіпотезою про рівномірність? Вибрати $\alpha = 0{,}10$ і $\alpha = 0{,}05$.

Задача 5.17

При 600 підкиданнях шестигранної грального кості шістка з'явилася 75 разів.

Чи можна стверджувати, що гральна кость симетрична і однорідна? Прийняти $\alpha = 0{,}05$.

Задача 5.18

При 600 підкиданнях шестигранної грального кості шістка з'явилася 75 разів.

Чи вірна гіпотеза про те, що ймовірність появи шістки менш ніж $1/6$, якщо $\alpha = 0{,}01$?

Задача 5.19

При 120 підкиданнях грального шестигранної кості на верхній грані одиниця випала $m_1 = 25$ разів, двійка $m_2 = 19$, трійка $m_3 = 15$, четвірка $m_4 = 22$, п'ятірка $m_5 = 15$ і, нарешті, шістка $m_6 = 21$ раз.

Чи узгоджується це з тим, що гральна кость правильної форми?

Задача 5.20

При вимірюванні продуктивності двох агрегатів (A і B) отримані наступні результати (в *кг* речовини за годину роботи):

№ виміру	1	2	3	4	5
Агрегат A	14,1	10,1	14,7	13,7	14,0
Агрегат B	14,0	14,5	13,7	12,7	14,1

Чи можна вважати, продуктивність агрегатів A та B однаковою, припускаючи, що обидві вибірки отримані з нормально розподілених генеральних сукупностей? Прийняти $\alpha = 0{,}05$.

Задача 5.21

Відповідно до технічних умов середній час безвідмовної роботи для приладів з великої партії повинен складати не менш 1000 *годин* з середньоквадратичним відхиленням $\sigma = 100$ *годин*. Вибіркове середнє часу безвідмовної роботи для випадково відібраних 25 приладів дорівнює 970 *годин*. Передбачимо, що СКВ часу безвідмовної роботи для приладів у вибірці збігається з СКВ часу безвідмовної роботи всієї партії.

Чи можна вважати, що вся партія приладів не задовольняє технічним умовам, якщо: а) $\alpha = 0{,}10$; б) $\alpha = 0{,}05$?

Задача 5.22

До наладки верстата була перевірена точність виготовлення 10 втулок і знайдено значення оцінки дисперсії діаметра $s^{*2} = 9,6\,мкм^2$. Після наладки зазнало контролю ще 15 втулок і отримано нове значення дисперсії $s^{*2} = 5,7\,мкм^2$.

Чи можна вважати, що внаслідок наладки верстата точність виготовлення втулок збільшилася ? Прийняти $\alpha = 0{,}05$.

Задача 5.23

У міській лікарні спостерігався розподіл червоних кров'яних кульок по 169 відділеннях приладу (гемацитометра). Кількості ν_i відділень, що містять по n_i червоних кров'яних кульок, вказані в таблиці.

ν_i	4	5	6	7	8	9	10	11	12
n_i	1	3	5	8	13	14	15	15	21
ν_i	13	14	15	16	17	18	19	20	21
n_i	18	17	16	9	6	3	2	2	1

Знайти середнє значення кількості червоних кров'яних кульок в одному відділенні. Прийнявши знайдене значення за параметр λ розподілу Пуассона, перевірити за допомогою χ^2-критерію гіпотезу про те, що вибірка узгоджується з розподілом Пуассона.

5.9. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

На протязі 100 діб фіксувалося кількість відвідувачів (X) деякої установи (дані наводяться в таблиці).

Потрібно за допомогою критерія χ^2 Пірсона перевірити нульову гіпотезу про те, що розподіл кількості відвідувачів підпорядковується закону Пуассона. Рівень значущості прийняти $\alpha = 0,05$.

Результати оформіть графічно.

Результат роботи – оцінка параметра λ закона Пуассона, масив амплітуд ймовірностей P_m, прийняття рішення відносно нульової гіпотези.

Варіант 1

Кількість відвідувачів (X)	0	1	2	3	4	5
Частоти m_i	7	27	32	19	11	4

Варіант 2

Кількість відвідувачів (X)	0	1	2	3	4	5
Частоти m_i	16	21	23	17	13	10

Варіант 3

Кількість відвідувачів (X)	0	1	2	3	4	5
Частоти m_i	3	27	39	24	6	1

Завдання 2

Отримано вибірку (дані наводяться в таблиці).

Потрібно за допомогою λ-критерія Колмогорова перевірити нульову гіпотезу про те, що наведену вибірку вибрано з генеральної сукупності, рівномірно розподіленої на інтервалі $(0,0;\ 2,0)$. Рівень значущості прийняти $\alpha = 0,05$.

Результати оформіть графічно.

Результат роботи — прийняття рішення відносно нульової гіпотези.

За яким значенням рівня значущості α нульову гіпотезу буде відхилено?

Варіант 1

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

Варіант 2

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

Варіант 3

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

5.10. Завдання для перевірки

1. Що називається критерієм згоди?

2. Чи є критерії згоди статистичними критеріями значущості?

3. Які характерні особливості властиві для статистичних критеріїв значущості і, зокрема, для критеріїв згоди?

4. Дайте загальну схему перевірки гіпотез про вигляд функції розподілу за допомогою критерію згоди χ^2 Пірсона.

5. Дайте загальну схему перевірки гіпотез про вигляд функції розподілу за допомогою λ-критерію Колмогорова.

6. Вкажіть переваги і недоліки критерію згоди χ^2 Пірсона; λ-критерію Колмогорова.

7. На основі яких ознак або критеріїв можна зробити попередній вибір закону розподілу?

8. Чи можуть дослідні дані одночасно узгоджуватися з декількома законами розподілу?

9. Вкажіть переваги і недоліки непараметричних критеріїв в порівнянні з параметричними критеріями значущості.

10. Для перевірки яких гіпотез застосовується критерій знаків?

11. Чи є критерій знаків статистичним критерієм значущості?

12. Аналогом якого параметричного критерію є критерій знаків?

6. Лінійний регресійний аналіз

6.1. Задачі регресійного і кореляційного аналізу

Однією з основних задач математичної статистики є дослідження залежності між двома або декількома параметрами. Дві змінні X і Y можуть бути або незалежними, або пов'язаними функціональною або іншою статистичною залежністю.

Визначення 1. *Функціональною залежністю* між змінними X і Y називається правило f, яке кожному елементу X з довільної множини Ω ставить у відповідність певний елемент Y множини F, тобто $y = f(x)$. Наприклад, площа кола ($S = \pi R^2$) і довжина кола ($L = 2\pi R$) повністю визначаються радіусом R, площа трикутника – його сторонами і таке інше.

Визначення 2. *Статистичною залежністю* між випадковими величинами X і Y – складовими двовимірної випадкової величини (X, Y) – називається правило f, яке кожному числу x з числової множини R ставить у відповідність умовний закон розподілу складової Y, тобто кожному x відповідає $f(x, y)$.

На практиці функціональні зв'язки між ознаками зустрічаються рідко. Частіше мають місце такі зв'язки між змінними величинами, при яких значенню однієї з них відповідає декілька значень інших. Наприклад, відомо, що врожайність залежить від кількості внесених добрив, але на неї впливають і інші чинники (якість грунту і таке інше). Крім того, одні й ті ж дози добрив, навіть за дуже однакових умов, часто по-різному впливають на врожайність.

Визначення 3. *Випадкові величини X та Y називаються незалежними*, якщо умовний закон розподілу однієї з складових не залежить від того, які значення прийняла інша складова, тобто якщо $f(y|x) = f(y)$ або $f(x|y) = f(x)$.

У курсі теорії ймовірностей показується, що для того, щоб складові X і Y двовимірної випадкової величини були незалежними, необхідно і достатньо, щоб густина розподілу дорівнювалась добутку парціальних густин розподілу складових: $f(x, y) = f(x) \cdot f(y)$.

З визначення статистичної залежності випливає, що для вивчення зміни випадкової величини Y за значеннями випадкової величини X на основі здобутих статистичних даних, тобто спостережених значень двовимірної випадкової величини (X, Y) $(x_i, y_i,\ i = 1, 2, \ldots, n)$, необхідно:

1) підібрати теоретико-ймовірнісну модель, що характеризує частотні закономірності двовимірного статистичного ряду $(x_i, y_i,\ i = 1, 2, \ldots, n)$, тобто вибрати функцію $F(x, y)$ або $f(x, y)$;

2) оцінити параметри цієї функції;

3) знайти умовні закони розподілу

$$f(y|x) = \frac{f(x,y)}{f(x)} \quad \text{або} \quad f(x|y) = \frac{f(x,y)}{f(y)}; \tag{6.1}$$

4) задати ймовірність $\mathrm{P} = 1 - \alpha$ і за значенням випадкової величини $\mathrm{X} = x$ визначити інтервал $[a,\, b]$ зміни випадкової величини Y:

$$\Pr(a < \mathrm{Y} < b) \;=\; \int\limits_a^b f(y|x)\, dy \;=\; 1 - \alpha. \tag{6.2}$$

На практиці при дослідженні залежності між випадковими величинами X і Y часто обмежуються дослідженням залежності між X і умовним математичним сподіванням

$$\mathrm{M}[\mathrm{Y}|\mathrm{X} = x] \;=\; \int\limits_{-\infty}^{\infty} y\, f(y|x)\, dy.$$

Залежність такого роду називається *регресійною залежністю*.

Умовне математичне сподівання $\mathrm{M}[\mathrm{Y}|\mathrm{X} = x]$ залежить від вибраної теоретико-ймовірнісної моделі $f(x,y)$, тобто воно є поняттям модельним.

Визначення 4. *Рівнянням регресії першого роду* Y *на* X (або *модельним рівнянням регресії*) називається математичне сподівання складової Y двовимірної випадкової величини (X, Y), яка розглядається як функція x, обчислене за умови, що складова X набула деякого фіксованого значення X= x:

$$\mathrm{M}[\mathrm{Y}|\mathrm{X} = x] \;=\; \int\limits_{-\infty}^{\infty} y\, f(y|x)\, dy \;=\; \varphi(x). \tag{6.3}$$

Функцію $\varphi(x)$ називають функцією регресії першого роду або модельною функцією регресії Y на X.

Визначення 5. *Рівнянням регресії* X *на* Y *другого роду* називається умовне математичне сподівання складової X двовимірної випадкової величини (X, Y), що розглядається як функція y:

$$\mathrm{M}[\mathrm{X}|\mathrm{Y} = y] \;=\; \int\limits_{-\infty}^{\infty} x\, f(x|y)\, dx \;=\; \psi(y). \tag{6.4}$$

Функцію $\psi(y)$ називають функцією регресії X на Y.

Модельне рівняння регресії вигляду $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \varphi(x)$ дозволяє робити "точковий" прогноз значень умовних математичних сподівань складової Y двовимірної випадкової величини (X, Y) за значеннями складової X= x. Однак, для такого прогнозу необхідно знати закон розподілу двовимірної випадкової величини (X, Y). На практиці при опрацюванні експериментальних даних закон розподілу двовимірної випадкової величини (X, Y), як правило, є невідомим. У розпорядженні

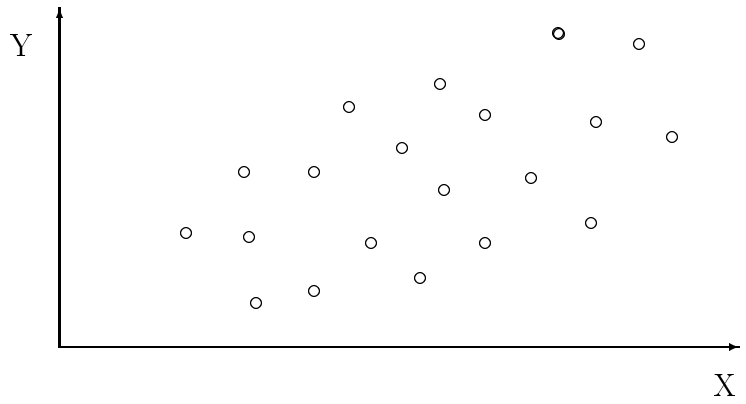


Рисунок 6.1 — Приклад кореляційного поля

експериментатора є тільки спостереженні значення двовимірної величини – точки (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$ або, більш стисло, вибірка обсягом n.

Якщо результати вибірки $\{x_i, y_i\}$, де $i = 1, 2, 3, \ldots, n$, зобразити у вигляді точок у декартовій системі координат, то отримаємо точкову діаграму, звану *кореляційним полем* (рис. 6.1).

Визначення 6. *Емпіричною функцією регресії* Y на X називається функція $\overline{y}_x = f(x; a, b, \ldots, d)$ певного класу, сукупність параметрів якої $a, b, \ldots, d$ знаходиться за здобутими значеннями двовимірної випадкової величини (x_i, y_i) $(i = 1, 2, \ldots, n)$, тобто за результатами вибірки обсягом n.

Для розв'язання задачі знаходження параметрів емпіричних рівнянь регресії $\mathrm{M[Y|X}= x] = f(x; a, b, \ldots, d)$ застосовується *метод найменших квадратів (МНК)*. Цей метод дозволяє, розглянувши задану вибрану залежність $\mathrm{M[Y}|x] = f(x; a, b, \ldots, d)$, так вибрати параметри $a, b, \ldots, d$, що емпірична функція регресії $\overline{y}_x = f(x; a, b, \ldots, d)$ буде найкращою оцінкою дійсної функції регресії. Під "найкращою оцінкою" мається на увазі те, що сума квадратів (*нев'язка*) відхилень ε спостережених значень змінної Y від відповідних ординат емпіричної функції регресії $\overline{y}_x = f(x; a, b, \ldots, d)$ буде *мінімальною* в просторі вказаних параметрів.

Шукані параметри a, b, ..., d заданої функції $f(x; a, b, \ldots, d)$ за МНК знаходяться з умови

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i; a, b, \ldots, d)]^2 \Rightarrow \min. \tag{6.5}$$

З умови екстремальності нев'язки S випливає система рівнянь:

$$\partial S/\partial a = 0; \qquad \partial S/\partial b = 0; \qquad \ldots \qquad \partial S/\partial d = 0. \tag{6.6}$$

Вибір класу емпіричної функції регресії може бути зроблений:

а) з візуальної оцінки характеру розташування точок (x_i, y_i) на кореляційному полі;

б) з досвіду попередніх досліджень;

в) з припущень теоретичного характеру, заснованих на знанні сутності задачі, що розв'язується.

Отже, емпіричне рівняння регресії $\overline{y}_x = f(x; a, b, \ldots, d)$ інтерпретується як оцінка (наближений вираз) модельного рівняння регресії Y на X. Аналогічно інтерпретується емпіричне рівняння регресії X на Y.

Регресійний аналіз – це аналіз функцій регресії першого і другого роду. За його допомогою розв'язуються наступні задачі:

1) знаходяться точкові й інтервальні оцінки параметрів емпіричної функції регресії;

2) проводяться точкове та інтервальне оцінювання умовних математичних сподівань, що необхідно для прогнозу середніх значень однієї випадкової величини, які відповідають певним фіксованим значенням іншої випадкової величини;

3) перевіряється узгодженість знайденої емпіричної функції регресії з експериментальними даними та інше.

Кореляційний аналіз – це аналіз властивостей оцінок $\hat{\rho}$ коефіцієнта кореляції

$$\rho \;=\; \frac{\mathrm{M}\left[(\mathrm{X}-m_{\mathrm{X}})\,(\mathrm{Y}-m_{\mathrm{Y}})\right]}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}. \tag{6.7}$$

Він дозволяє дати відповідь на питання про існування лінійної функціональної залежності між математичними сподіваннями випадкових величин X і Y. У випадку позитивної відповіді метод кореляційного аналізу дозволяє знаходити ступінь тісноти статистичної залежності (ступінь близькості статистичної залежності до функціональної).

6.2. Ймовірнісне введення у регресійний аналіз

Якщо закон розподілу двовимірної випадкової величини (X,Y) є відомим, то можна знайти умовний закон розподілу складової Y за умови, що складова X набула деякого фіксованого значення x, за формулою

$$f(y|x) \;=\; \frac{f(x,y)}{f_1(x)}, \tag{6.8}$$

де

$$f_1(x) = \int\limits_{-\infty}^{\infty} f(x,y)\,dy \tag{6.9}$$

– густина розподілу складової X двовимірної ВВ (X,Y).

Умовне математичне сподівання складової Y (модельне рівняння регресії Y на X) знаходиться за формулою

$$\mathrm{M}[\mathrm{Y}|\mathrm{X}=x] \;=\; \int\limits_{-\infty}^{\infty} y\,f(y|x)\,dy. \tag{6.10}$$

Розглянемо двовимірну випадкову величину (X,Y), розподілену згідно з нормальним законом з густиною розподілу ймовірностей (рис. 6.2 та 6.3)

$$f(x,y) \;=\; \frac{1}{2\pi\,\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}\,\sqrt{1-\rho^2}}\;\times \tag{6.11}$$

$$\times\exp\left\{-\;\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2}-\frac{2\rho\,(x-m_{\mathrm{X}})(y-m_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\,\sigma_{\mathrm{Y}}}+\frac{(y-m_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right\}.$$

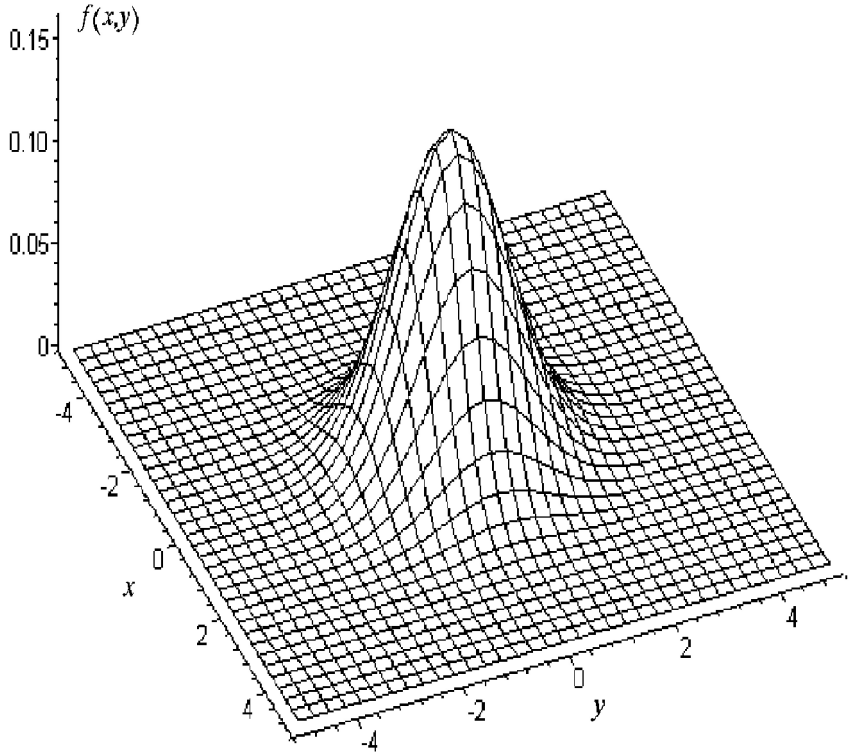


Рисунок 6.2 — Густина розподілу ймовірностей $f(x, y)$ системи з двох випадкових величин X, Y (параметри: $m_{\mathrm{X}} = m_{\mathrm{Y}} = 0$; $\sigma_{\mathrm{X}} = \sigma_{\mathrm{Y}} = 1$; коефіцієнт кореляції $\rho = 0{,}0$)

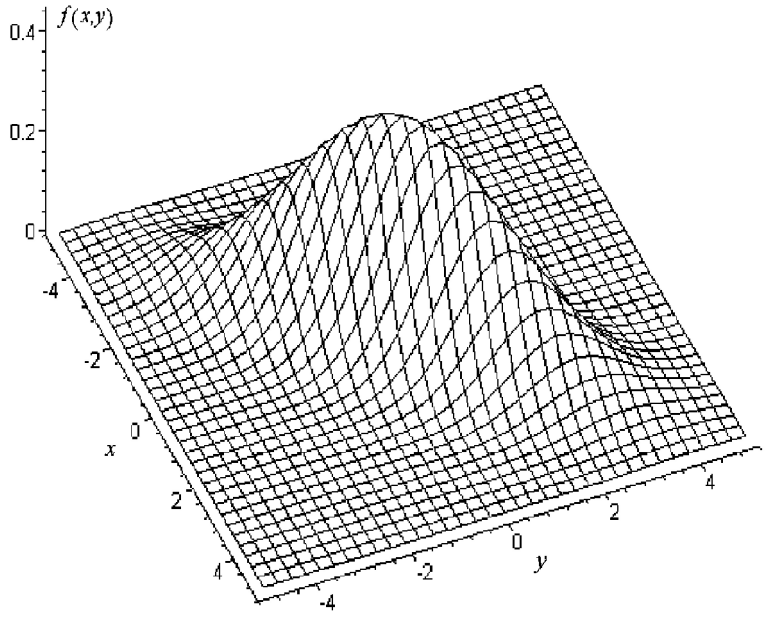


Рисунок 6.3 — Густина розподілу ймовірностей $f(x, y)$ системи з двох випадкових величин X, Y (параметри: $m_{\mathrm{X}} = m_{\mathrm{Y}} = 0$; $\sigma_{\mathrm{X}} = \sigma_{\mathrm{Y}} = 1$; коефіцієнт кореляції $\rho = 0{,}8$)

Знайдемо густину розподілу ймовірностей складових X і Y:

$$f_1(x) = \int\limits_{-\infty}^{\infty} f(x,y)\,dy \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{X}}}\,\exp\left(-\frac{(x-m_{\mathrm{X}})^2}{2\sigma_{\mathrm{X}}^2}\right), \tag{6.12}$$

тому з симетрії густини нормального двовимірного розподілу випливає, що

$$f_1(y) = \int\limits_{-\infty}^{\infty} f(x,y)\,dx \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_{\mathrm{Y}}}\,\exp\left(-\frac{(y-m_{\mathrm{Y}})^2}{2\sigma_{\mathrm{Y}}^2}\right). \tag{6.13}$$

Знайдемо умовні розподіли складових X і Y:

$$f(x|y) \;=\; \frac{f(x,y)}{f_1(y)} \;= \tag{6.14}$$

$$= \;\frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_{\mathrm{X}}}\,\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_{\mathrm{X}}^2}\left[x-m_{\mathrm{X}}-\rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y-m_{\mathrm{Y}})\right]^2\right\};$$

$$f(y|x) \;=\; \frac{f(x,y)}{f_1(x)} \;= \tag{6.15}$$

$$= \;\frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_{\mathrm{Y}}}\,\exp\left\{-\;\frac{1}{2(1-\rho^2)\sigma_{\mathrm{Y}}^2}\left[y-m_{\mathrm{Y}}-\rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x-m_{\mathrm{X}})\right]^2\right\}.$$

Умовні математичні сподівання наступні:

$$\begin{aligned} m_{\mathrm{X|Y}} \;&=\; \mathrm{M}[\mathrm{X}|\mathrm{Y}=y] \;=\; m_{\mathrm{X}} + \rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y-m_{\mathrm{Y}})\,;\\ m_{\mathrm{Y|X}} \;&=\; \mathrm{M}[\mathrm{Y}|\mathrm{X}=x] \;=\; m_{\mathrm{Y}} + \rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x-m_{\mathrm{X}})\,; \end{aligned} \tag{6.16}$$

$$\begin{aligned} \sigma[\mathrm{X}|\mathrm{Y}=y] \;&=\; \sigma_{\mathrm{X}}\sqrt{1-\rho^2};\\ \sigma[\mathrm{Y}|\mathrm{X}=x] \;&=\; \sigma_{\mathrm{Y}}\sqrt{1-\rho^2}. \end{aligned} \tag{6.17}$$

У наведеному умовному розподілі складової Y за умови, що складова X набула деякого значення x, тільки умовне математичне сподівання $m_{\mathrm{Y|X}} = \mathrm{M}[\mathrm{Y}|\mathrm{X}=x]$ залежить від значення x, в той час як умовне середнє квадратичне відхилення не залежить від x.

Таким чином, модельне рівняння регресії Y на X двовимірної випадкової величини (X,Y), розподіленої згідно з нормальним законом, має вигляд

$$m_{\mathrm{Y|X}} - m_{\mathrm{Y}} \;=\; \rho\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}\,(x-m_{\mathrm{X}})\,. \tag{6.18}$$

Графіком цієї функції (*модельна лінія регресії* Y *на* X) є пряма лінія, що проходить крізь центр розподілу – точку $(m_{\mathrm{X}}, m_{\mathrm{Y}})$.

Аналогічні результати отримуємо, аналізуючи умовну густину розподілу ймовірностей $f(x|y)$. У цьому випадку рівняння регресії X на Y має вигляд

$$m_{\mathrm{X|Y}} - m_{\mathrm{X}} \;=\; \rho\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}\,(y-m_{\mathrm{Y}})\,, \tag{6.19}$$

при цьому відповідна модельна лінія регресії X на Y також проходить крізь центр розподілу – точку $(m_{\mathrm{X}}, m_{\mathrm{Y}})$.

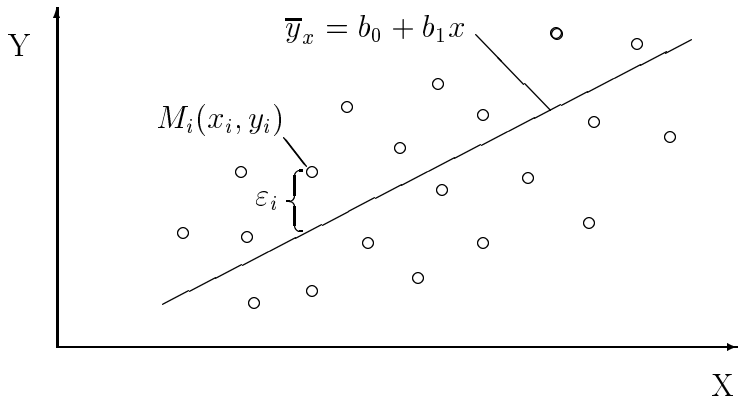


Рисунок 6.4 — Приклад кореляційного поля і лінійного рівняння регресії $\overline{y}_x = b_0 + b_1 x$

6.3. Лінійна регресія

Розглянемо методику розрахунку емпіричних лінійних рівнянь регресії за *незгрупованими* і *згрупованими даними*.

Незгруповані дані. Припустимо, що зроблений експеримент, внаслідок якого зафіксовано n значень змінних X і Y: (x_i, y_i) $(i = 1, 2, \ldots, n)$.

При нанесенні експериментальних даних у вигляді точок у декартовій системі координат отримуємо кореляційне поле (рис. 6.1). Якщо є підстави вважати, що двовимірна випадкова величина (X,Y) розподілена згідно з нормальним законом або якщо точки на кореляційному полі групуються навколо прямої лінії (рис. 6.4), то емпіричне рівняння регресії підбирається у вигляді

$$\overline{y}_x = b_0 + b_1 x. \tag{6.20}$$

Наступна задача – знаходження коефіцієнтів (параметрів) b_0 і b_1 лінійних емпіричних функцій регресії Y на X. Шукатимемо ці параметри методом найменших квадратів, тобто прагнемо виконання умови мінімуму нев'язки

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \overline{y}_x)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \Rightarrow \min. \tag{6.21}$$

Оскільки функціонал нев'язки S в просторі параметрів b_0 та b_1 досягає мінімуму, то для визначення координат цього екстремуму маємо знайти часткові похідні $\partial S/\partial b_0$, $\partial S/\partial b_1$ і прирівняти їх до нуля:

$$\partial S/\partial b_0 = 0 \Rightarrow 2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0; \tag{6.22a}$$

$$\partial S/\partial b_1 = 0 \Rightarrow 2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0. \tag{6.22b}$$

Отримуємо систему з двох лінійних рівнянь (тобто *систему нормальних рівнянь*):

$$\begin{cases} b_0 n + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i; \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{6.23}$$

Розв'язуючи систему, знаходимо шукані коефіцієнти b_0 та b_1.

Якщо потрібно за експериментальними даними розв'язати лінійне рівняння регресії X на Y вигляду

$$\overline{x}_y \;=\; a_0 + a_1 y, \tag{6.24}$$

то його коефіцієнти (параметри) a_0 та a_1 знаходять з розв'язку системи нормальних рівнянь:

$$\begin{cases} a_0 n + a_1 \sum\limits_{i=1}^{n} y_i \;=\; \sum\limits_{i=1}^{n} x_i; \\ a_0 \sum\limits_{i=1}^{n} y_i + a_1 \sum\limits_{i=1}^{n} y_i^2 \;=\; \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{6.25}$$

Згруповані дані. Якщо кількість вимірювань велика, то для спрощення розрахунків експериментальні дані заведено групувати, тобто об'єднувати в зведену таблицю, звану *кореляційною таблицею.*

Таблиця 6.1 — Кореляційна таблиця

X	Y						
	y_1	y_2	$\cdots$	y_j	$\cdots$	y_l	m_x
x_1	m_{11}	m_{12}	$\cdots$	m_{1j}	$\cdots$	m_{1l}	m_{x1}
x_2	m_{21}	m_{22}	$\cdots$	m_{2j}	$\cdots$	m_{2l}	m_{x2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	m_{i1}	m_{i2}	$\cdots$	m_{ij}	$\cdots$	m_{il}	m_{xi}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_k	m_{k1}	m_{k2}	$\cdots$	m_{kj}	$\cdots$	m_{kl}	m_{xk}
m_y	m_{y1}	m_{y2}	$\cdots$	m_{yj}	$\cdots$	m_{yl}	n

У цій таблиці:

$x_1, x_2, \ldots, x_k$ – середини інтервалів або значення ознаки X;

$y_1, y_2, \ldots, y_n$ – середини інтервалів або значення ознаки Y;

$m_{x1}, m_{x2}, \ldots, m_{xi}, \ldots, m_{xk}$ та

$m_{y1}, m_{y2}, \ldots, m_{yj}, \ldots, m_{yl}$ – відповідні частоти;

m_{ij} – частота, з якою зустрічається пара (x_i, y_j);

$n = \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij}$.

Докладні приклади обчислень за допомогою кореляційної таблиці наведені далі та в прикладах до розділу.

Нехай потрібно на основі вибірки обсягом n "оцінити" значення модельної функції регресії і зробити прогноз умовних математичних сподівань випадкової величини Y, що відповідають певним значенням випадкової величини X= x. Для цього використовуються рівняння регресії другого роду (емпіричні рівняння регресії).

Наближений вираз (оцінку) модельної функції регресії називають *емпіричною функцією регресії* $\overline{y}_x = f(x; a, b, \ldots, d)$.

6.4. Коефіцієнт кореляції

З теорії ймовірностей відомо, що основними характеристиками, які описують ступінь зв'язку між складовими X і Y двовимірної випадкової величини (X, Y), є *коваріація* μ_{XY} (кореляційний момент)

$$\mu_{11} = \mu_{\mathrm{XY}} = \mathrm{M}[(\mathrm{X} - m_{\mathrm{X}})(\mathrm{Y} - m_{\mathrm{Y}})] = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}(x - m_{\mathrm{X}})(y - m_{\mathrm{Y}})f(x,y)\,dx\,dy \qquad (6.26)$$

і *коефіцієнт кореляції*

$$\rho \;=\; \frac{\mu_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;=\; \frac{\mathrm{M}[(\mathrm{X} - m_{\mathrm{X}})(\mathrm{Y} - m_{\mathrm{Y}})]}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}. \qquad (6.27)$$

З виразів (6.26) і (6.27) випливає, що для знаходження μ_{XY} та ρ необхідно знати закон розподілу $f(x,y)$ двовимірної випадкової величини. У більшості випадків при опрацюванні експериментальних даних закон розподілу двовимірної випадкової величини є невідомим. Тому для оцінки тісноти зв'язку (рис. 6.5, 6.6 та 6.7) застосовуються точкові оцінки $\hat{\mu}_{\mathrm{XY}}$ та $\hat{\rho}$ величин μ_{XY} та ρ:

емпіричний кореляційний момент

$$\hat{\mu}_{\mathrm{XY}} \;=\; K_{\mathrm{XY}} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}); \qquad (6.28)$$

емпіричний коефіцієнт кореляції

$$\hat{\rho} \;=\; r \;=\; \frac{K_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;= \qquad (6.29)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\,\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{-1/2}.$$

Надалі буде корисною дещо інша форма запису емпіричного кореляційного моменту:

$$K_{\mathrm{XY}} \;=\; \overline{xy} - \overline{x}\cdot\overline{y}. \qquad (6.30)$$

Аналогічно можна записати для вибіркових дисперсій s_x^2 та s_y^2, що

$$s_x^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \;=\; \overline{x^2} - \overline{x}^2, \qquad s_y^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \;=\; \overline{y^2} - \overline{y}^2. \qquad (6.31)$$

З урахуванням цих виразів отримуємо ще одну формулу для обчислення емпіричного коефіцієнта кореляції:

$$\hat{\rho} \;=\; r \;=\; \frac{K_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} \;=\; (\overline{xy} - \overline{x}\cdot\overline{y})\left(\left[\,\overline{x^2} - \overline{x}^2\,\right]\left[\,\overline{y^2} - \overline{y}^2\,\right]\right)^{-1/2}. \qquad (6.32)$$

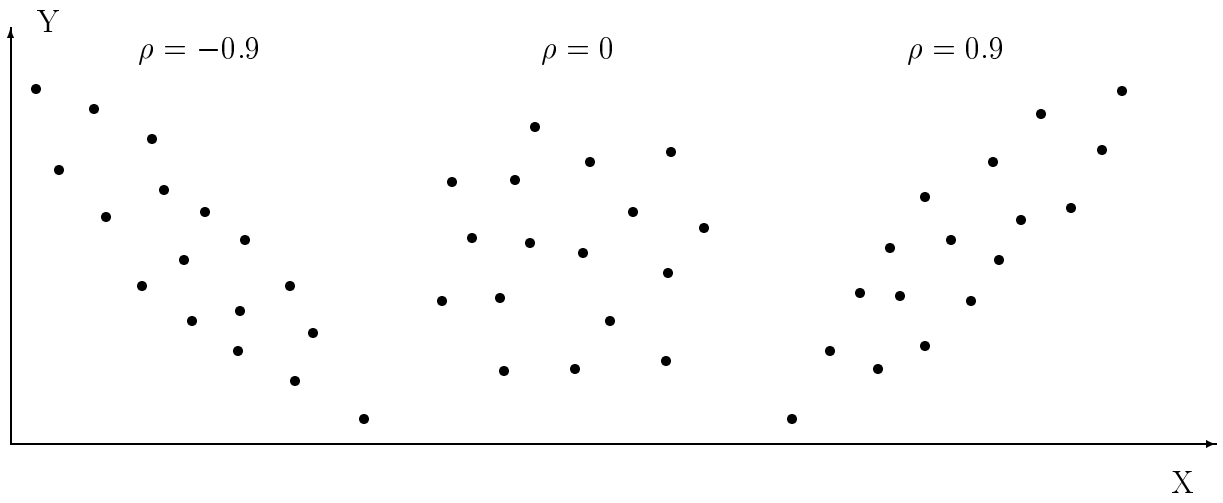


Рисунок 6.5 — Три приклади кореляційного поля ($\rho = -0{,}9;\ 0{,}0;\ 0{,}9;$ лінії регресії відсутні)

Система нормальних рівнянь в цих позначеннях набуває вигляду (регресія Y на X)

$$\begin{cases} b_0 + b_1\overline{x} \;=\; \overline{y}; \\ \\ b_0\overline{x}^2 + b_1\overline{x^2} \;=\; \overline{xy}. \end{cases} \tag{6.33}$$

Розв'язуючи цю систему, знаходимо

$$b_0 \;=\; \overline{y} - b_1\overline{x}, \qquad b_1 \;=\; \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} \;=\; r\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}. \tag{6.34}$$

З системи (6.33) випливає, що пряма регресії Y на X проходить через точку $C = (\overline{x}, \overline{y})$. Отже, рівняння регресії Y на X можна записати у вигляді

$$\overline{y}_x - \overline{y} \;=\; r\,\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}\,(x - \overline{x}). \tag{6.35}$$

Діючи аналогічно, можна показати, що система нормальних рівнянь регресії X на Y має вигляд

$$\begin{cases} a_0 + a_1\overline{y} \;=\; \overline{x}, \\ \\ a_0\overline{y}^2 + a_1\overline{y^2} \;=\; \overline{xy}. \end{cases} \tag{6.36}$$

Розв'язуючи цю систему, отримуємо

$$a_0 \;=\; \overline{x} - a_1\overline{y}, \qquad a_1 \;=\; \frac{\overline{xy} - \overline{y}\cdot\overline{x}}{\overline{y^2} - \overline{y}^2} \;=\; r\,\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}. \tag{6.37}$$

Отже, рівняння прямої регресії X на Y можна записати у вигляді

$$\overline{x}_y - \overline{x} \;=\; r\,\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}\,(y - \overline{y}). \tag{6.38}$$

Знайдені формули зручні з обчислювальної точки зору. Дійсно, щоб записати рівняння регресії змінної Y на X, або змінної X на Y, досить знайти точкові оцінки нормальної двовимірної випадкової величини (X, Y): $\overline{x}$, $\overline{y}$, s_x, s_y, r.

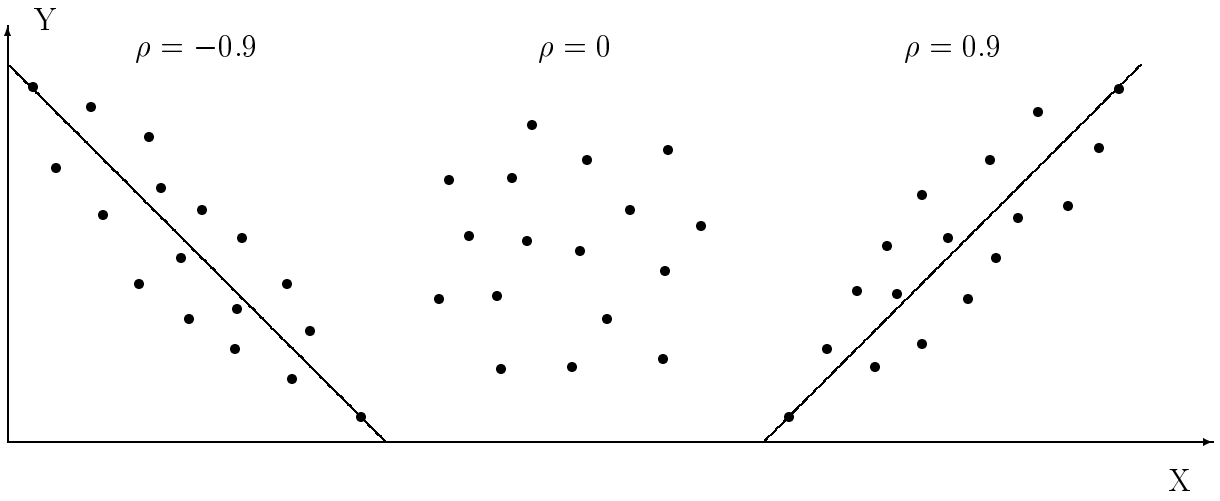


Рисунок 6.6 — Три приклади кореляційного поля ($\rho = $ –0,9; 0,0; 0,9; наведені лінії регресії)

З формул (6.35) і (6.38) випливає важливе співвідношення, яке пов'язує коефіцієнти регресії a_1, b_1 та коефіцієнт кореляції:

$$r^2 = a_1 b_1 \qquad \text{або} \qquad r = \pm\sqrt{a_1 b_1}.$$

Знак коефіцієнта кореляції збігається зі знаком коефіцієнтів регресії. Якщо у формулах (6.35) й (6.38) коефіцієнт регресії Y на X $b_1 = rs_{\mathrm{Y}}/s_{\mathrm{X}}$ та коефіцієнт регресії X на Y $a_1 = rs_{\mathrm{X}}/s_{\mathrm{Y}}$ та вони *додатні* (*від'ємні*), то говорять, що напрямок залежності Y на X *додатний* (*від'ємний*). Це означає, що змінні Y та X одночасно *зростають* (*зменшуються*).

Регресійні коефіцієнти a_1 й b_1 не дозволяють судити про ступінь зв'язку між випадковими величинами Y та X. Ступінь зв'язку залежить від кута між прямими регресії. Чим менше кут між прямими регресії, тим тісніше зв'язок між випадковими величинами Y та X. При злитті двох прямих регресії в одну має місце лінійна функціональна залежність між Y та X.

Кут між лініями регресії визначається рівняннями (6.35) й (6.38). Оскільки $\mathrm{tg}\alpha = b_1 = rs_{\mathrm{Y}}/s_{\mathrm{X}}$ та $\mathrm{tg}\beta = a_1 = rs_{\mathrm{X}}/s_{\mathrm{Y}}$, то з рис. 6.8 випливає, що $\mathrm{tg}\theta = \mathrm{tg}\left(\frac{\pi}{2} - \alpha - \beta\right) = \mathrm{ctg}\,(\alpha + \beta)$, і тому

$$\mathrm{tg}\theta = \frac{1 - \mathrm{tg}\alpha \cdot \mathrm{tg}\beta}{\mathrm{tg}\alpha + \mathrm{tg}\beta} = \frac{1 - r^2}{r} \frac{s_{\mathrm{X}} s_{\mathrm{Y}}}{s_{\mathrm{X}}^2 + s_{\mathrm{Y}}^2}. \tag{6.39}$$

В обчислювальной практиці прийнято користуватися коефіцієнтом кореляції

$$\hat{\rho} = r = \pm\sqrt{a_1 b_1} = K_{\mathrm{XY}}/(s_{\mathrm{X}} s_{\mathrm{Y}})$$

як показника ступеня звя'зку.

Можливі числові значення коефіцієнта кореляції $\hat{\rho}$ обмежені інтервалом $[-1;\ 1]$. Всередині цього інтервалу величина $\hat{\rho}$ визначається з наступних умов:

1) якщо між змінними X та Y існує лінійний позитивний зв'язок, то коефіцієнт кореляції $\hat{\rho} = 1$;

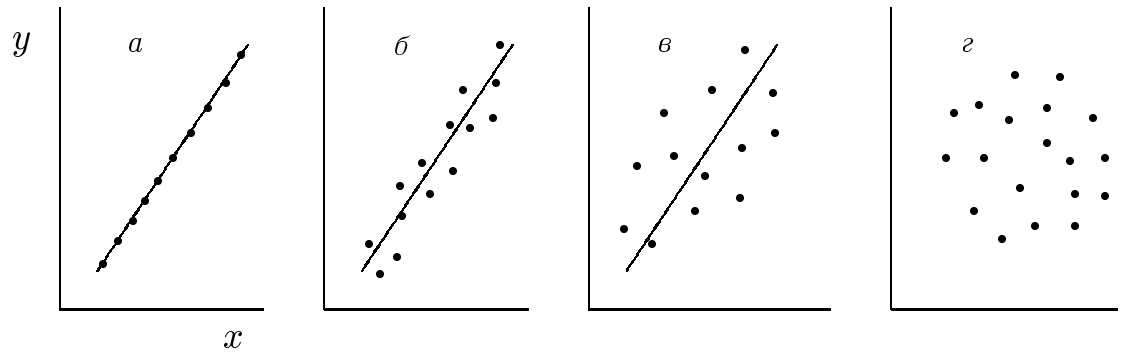


Рисунок 6.7 — Чотири приклада позитивної кореляційної залежності даних: *а*) практично повна кореляція ($\rho = 1$); *б*) висока ступінь кореляції ($\rho \approx 0,8$); *в*) помірна кореляція ($\rho \approx 0,4$); *г*) відсутність кореляції ($\rho \approx 0,0$)

2) якщо між змінними X та Y існує лінійний негативний зв'язок, то коефіцієнт кореляції $\hat{\rho} = -1$;

3) при відсутності залежності між змінними X та Y коефіцієнт кореляції $\hat{\rho} = 0$;

4) у всіх інших випадках $-1 < \hat{\rho} < 1$.

Справедливе і зворотне твердження: чим ближче за модулем коефіцієнт кореляції $\hat{\rho}$ до одиниці, тим сильніша *лінійна залежність* між випадковими величинами X і Y, а чим ближче $\hat{\rho}$ до нуля, тим ця залежність слабша.

Нагадаємо, що коефіцієнт кореляції $\hat{\rho}$ характеризує міру лінійної залежності, тому навіть при $\hat{\rho} = 0$ може виявитися, що між X і Y існує зв'язок функціонального нелінійного вигляду.

Кутовий коефіцієнт прямої регресії Y на X називають *коефіцієнтом регресії* Y на X і позначають через $\rho_{\mathrm{Y|X}}$. Аналогічно кутовий коефіцієнт прямої регресії X на Y називають *коефіцієнтом регресії* X на Y і позначають через $\rho_{\mathrm{X|Y}}$. Ці коефіцієнти можна обчислити за формулами:

$$\rho_{\mathrm{Y|X}} = \frac{1}{\sigma_{\mathrm{X}}^2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right] = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_{\mathrm{X}}^2}; \tag{6.40a}$$

$$\rho_{\mathrm{X|Y}} = \frac{1}{\sigma_{\mathrm{Y}}^2}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right] = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sigma_{\mathrm{Y}}^2}. \tag{6.40b}$$

В цих позначеннях рівняння прямих регресій приймають вигляд

$$y - \overline{y} = \rho_{\mathrm{Y|X}}(x - \overline{x}), \qquad x - \overline{x} = \rho_{\mathrm{X|Y}}(y - \overline{y}). \tag{6.41}$$

Коефіцієнтом лінійної кореляції ознак X і Y називають величину

$$r = r(\mathrm{X},\mathrm{Y}) = \pm\sqrt{\rho_{\mathrm{X|Y}}\,\rho_{\mathrm{Y|X}}} = \tag{6.42}$$

$$= \frac{1}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right].$$

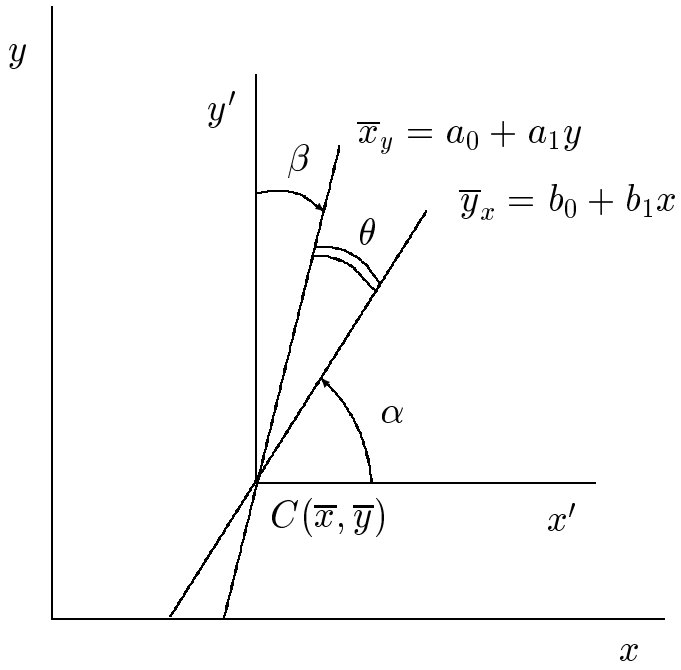


Рисунок 6.8 — До обчислення кута θ між лініями регресії

Квадрат коефіцієнта лінійної кореляції дає *коефіцієнт детермінації* k_{det}, який вимірює долю варіації компоненти X, що пояснюється впливом Y, і навпаки.

$$k_{\text{det}} \;=\; r^2 \;=\; \rho_{\text{X|Y}}\,\rho_{\text{Y|X}}. \tag{6.43}$$

Якщо лінії регресії відмінні від прямих, то коефіцієнт лінійної кореляції не дає повного уявлення про силу зв'язку між відповідними ознаками X і Y. У цьому випадку за ступінь залежності ознаки Y від ознаки X приймають *кореляційне відношення* $\eta_{\text{Y|X}}$, яке є відношенням середнього квадратичного відхилення умовних середніх $\overline{y}_x$ до середнього квадратичного відхилення ознаки Y,

$$\eta_{\text{Y|X}} \;=\; \frac{\sigma\left(\overline{\text{Y}}_x\right)}{\sigma_{\text{Y}}} \;=\; \frac{\left(\sum\limits_{j=1}^{k} m_{x_j}\left(\overline{\text{Y}}_{x_j} - \overline{\text{Y}}\right)^2\right)^{1/2}}{\left(\sum\limits_{j=1}^{n} m_{y_j}\left(y_j - \overline{\text{Y}}\right)^2\right)^{1/2}}. \tag{6.44a}$$

Аналогічно впроваджується *кореляційне відношення* $\eta_{x|y}$ ознаки X від ознаки Y

$$\eta_{\text{X|Y}} \;=\; \frac{\sigma\left(\overline{\text{X}}_y\right)}{\sigma_{\text{X}}} \;=\; \frac{\left(\sum\limits_{j=1}^{n} m_{y_j}\left(\overline{\text{X}}_{y_j} - \overline{\text{X}}\right)^2\right)^{1/2}}{\left(\sum\limits_{j=1}^{k} m_{x_j}\left(x_j - \overline{\text{X}}\right)^2\right)^{1/2}}. \tag{6.44b}$$

Можливі значення кореляційних відношень $\eta_{\text{Y|X}}$ та $\eta_{\text{X|Y}}$ обмежені інтервалом $[0;\,1]$. Всередені цього інтервалу величини $\eta_{\text{Y|X}}$ та $\eta_{\text{X|Y}}$ визначаються з наступних умов:

1) якщо $\eta_{Y|X} = 0$ (або $\eta_{X|Y} = 0$), то ознаки Y та X не корелюють;

2) якщо $\eta_{Y|X} = 1$, то ознака Y пов'язана з ознакою X функціональним зв'язком, $y = f(x)$;

3) виконуються нерівності $\eta_{Y|X} \geq |r(X,Y)|$ та $\eta_{X|Y} \geq |r(X,Y)|$.

6.5. Перевірка гіпотез про значущість коефіцієнта кореляції

Головною метою кореляційного аналізу є виявлення зв'язку між випадковими величинами X та Y і, якщо виявиться, що між ознаками цей зв'язок має місце, – визначення ступеня близькості цього зв'язку до функціонального. На практиці при опрацюванні експериментальних даних коефіцієнт кореляції генеральної сукупності ρ є невідомим. За результатами експерименту на підставі вибірки отримується його точкова оцінка (наближене значення) – *вибірковий коефіцієнт кореляції* $r = \hat{\rho}$.

Якщо виявиться, що вибірковий коефіцієнт кореляції r дорівнює нулю, то це ще не значить, що відповідний коефіцієнт кореляції ρ теж дорівнює нулю, а випадкові величини X та Y незалежні (у випадку їх нормальності), і навпаки, якщо $r \neq 0$, то це ще не значить, що $\rho \neq 0$, а випадкові величини можуть виявитися незалежними.

Для того, щоб на підставі статистичного аналізу відповісти на питання, чи знаходяться випадкові величини в кореляційної залежності, необхідно перевіріти нульову гіпотезу $\{H_0 : \rho = 0\}$ проти однієї з альтернативних гіпотез $\{H_a : \rho \neq 0\}$, або $\{H_a : \rho > 0\}$, або $\{H_a : \rho < 0\}$.

Для перевірки нульової гіпотези необхідно знати закон розподілу вибіркового коефіцієнта кореляції. Густина розподілу ймовірностей $f(\hat{r})$ емпіричного (вибіркового) коефіцієнта кореляції r у випадку, якщо вибірка отримана з сукупності з двовимірним нормальним розподілом, залежить від обсягу вибірки n та коефіцієнта кореляції ρ генеральної сукупності і має вигляд:

$$f(\hat{r}) = \frac{n-2}{\pi}(1-\rho^2)^{(n-1)/2}(1-\hat{r}^2)^{(n-4)/2}\int\limits_0^1 \frac{x^{n-2}}{(1-\rho\hat{r}x)^{n-1}}\frac{dx}{\sqrt{1-x^2}}. \qquad (6.45)$$

Для вибірок з генеральної сукупності, у якій величини X та Y нормальні та незалежні, тобто $\rho = 0$, густина розподілу ймовірностей емпіричного коефіцієнта кореляції така:

$$f(\hat{r}) = \frac{1}{\pi}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}(1-\hat{r}^2)^{(n-4)/2}. \qquad (6.46)$$

Формули (6.45) й (6.46) називають *розподілом Крамера*.

На рис. 6.9 і 6.10 наведено графіки диференціальної та інтегральної функцій розподілу вибіркового коефіцієнта кореляції при значеннях параметрів $n = 10$ та $\rho =$ –0,8; –0,4; 0,0; 0,4; 0,8.

Побудова інтервальних оцінок для коефіцієнта кореляції генеральної сукупності методами, що безпосередньо використовують формули (6.45)–(6.46), має місце достатньо рідко.

В практиці для побудови інтервальних оцінок для ρ користуються або номограмами (в спеціалізованих математичних або статистичних пакетах), що основані на

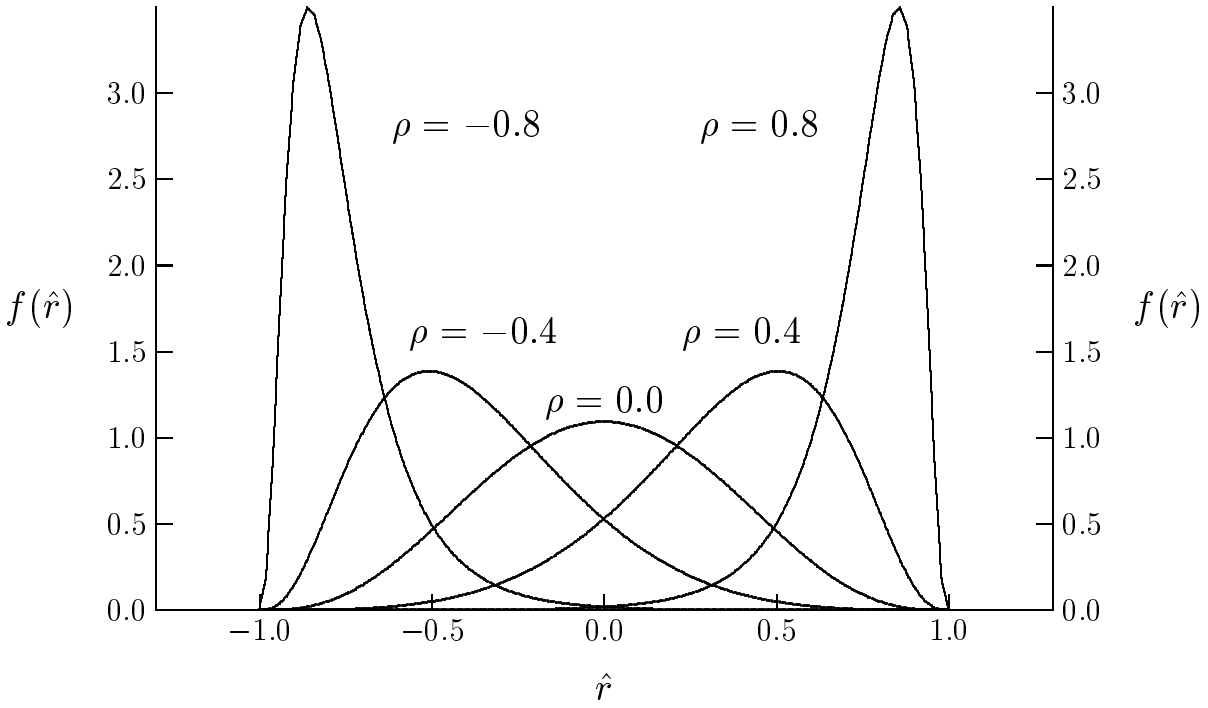


Рисунок 6.9 — Сімейство з 5 густин розподілу ймовірностей $f(\hat{r}) = f(\hat{r}; \rho, n)$; (обсяг вибірки дорівнює $n = 10$; залежності приведено для ідеальних коефіцієнтів кореляції, що дорівнюють $\rho =$ –0,8; –0,4; 0,0; 0,4; 0,8)

цих формулах, або використовують спеціальні перетворення емпіричного коефіцієнта кореляції, які дозволюють звести розподіл деякої функції від вибіркового коефіцієнта кореляції до добре знайомих розподілів, наприклад, нормального або розподілу Стьюдента.

Перевірка значущості вибіркового коефіцієнта кореляції r, тобто перевірка того, яку величину вибіркового коефіціента кореляції потрібно вважати достатньою для статистично обгрунтованого висновка про наявність кореляційного зв'язку між змінними X та Y, що досліджуються, грунтується на наступних трьох математичних моделях, які є узагальненням вихідних передумов про двовимірну генеральну сукупність (X, Y).

Модель 1.

Використовуються для перевірки нульової гіпотези про відсутність кореляційного зв'язку між змінними, що досліджуються, тобто для перевірки гіпотези $\{H_0 : \ \rho = 0\}$.

Вихідні передумови та обмеження:

а) двовимірний закон розподілу змінних (X, Y) в генеральній сукупності передбачається нормальним;

б) обсяг вибірки n може бути будь-яким.

Для перевірки значущості вибіркового коефіцієнта кореляціїї обчислюється статистика

$$t \ = \ r\sqrt{(n-2)(1-r^2)}, \tag{6.47}$$

яка має розподіл Стьюдента з $\nu = n - 2$ ступенями вільності.

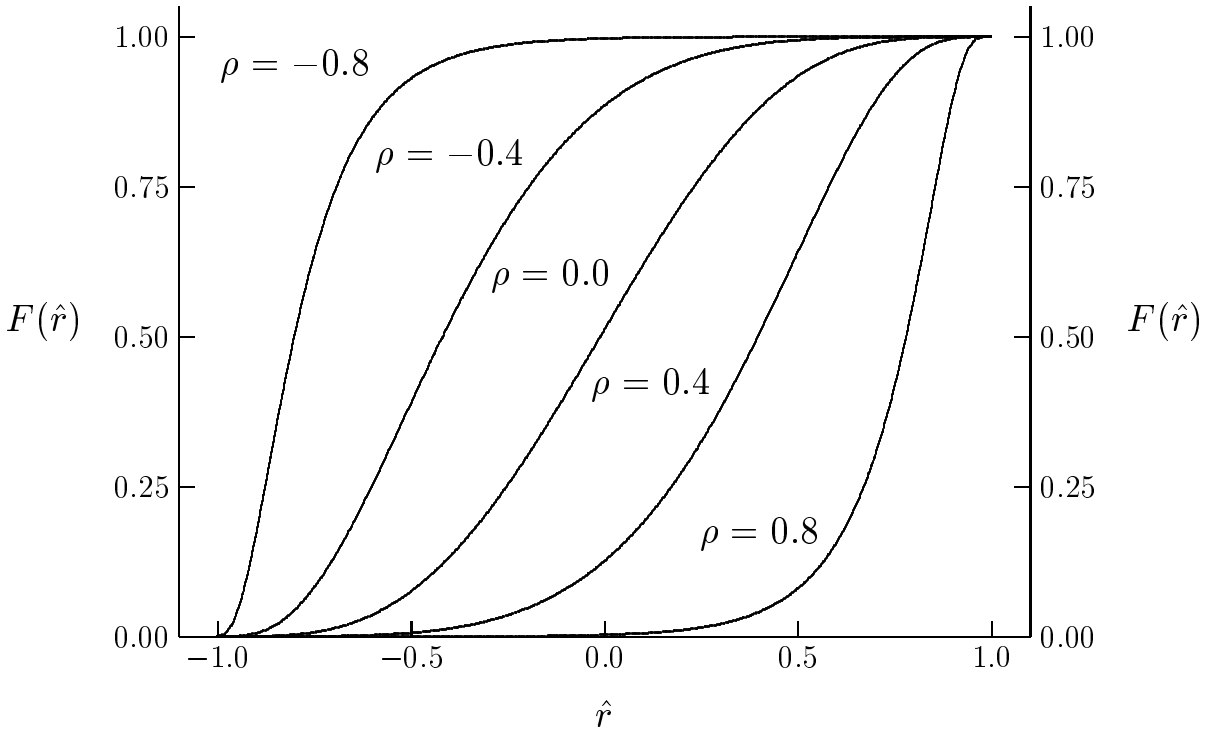


Рисунок 6.10 — Сімейство з 5 інтегральних розподілів імовірностей $F(\hat{r}) = F(\hat{r}; \rho, n)$; (обсяг вибірки дорівнює $n = 10$; залежності приведено для ідеальних коефіцієнтів кореляції, що дорівнюють $\rho =$ –0,8; –0,4; 0,0; 0,4; 0,8)

Для перевірки нульової гіпотези $\{H_0 : \ \rho = 0\}$ знаходять за таблицею розподілу Стьюдента за фіксованим рівнем значущості α та кількістю ступенів вільності $\nu = n - 2$ критичне значення $t_{\alpha/2;\, n-2}$, яке задовольняє умові $\Pr(|t| \geq t_{\alpha/2;\, n-2}) = \alpha$.

Якщо

$$|t_{\text{спос}}| \ \geq \ t_{\alpha/2;\, n-2}, \tag{6.48a}$$

то нульову гіпотезу про відсутність лінійної залежності між змінними X та Y потрібно відхилити.

Якщо ж

$$|t_{\text{спос}}| \ < \ t_{\alpha/2;\, n-2}, \tag{6.48b}$$

то немає причин відхиляти нульову гіпотезу про некорельованість змінних X та Y.

Модель 2.

Використовується для перевірки гіпотези про силу кореляційного зв'язку між змінними X та Y, інакше, для перевірки нульової гіпотези про те, що коефіцієнт кореляції ρ генеральної сукупності дорівнює деякому фіксованому числу, тобто $\{H_0 : \ \rho = \rho_0\}$. (Якщо $\rho_0 = 0$, то перевіряється гіпотеза про некорельованість змінних X і Y.)

Вихідні передумови та обмеження:

а) двовимірний закон розподілу змінних (X, Y), що досліджуються, в генеральній сукупності передбачається нормальним;

б) обсяг вибірки n достатньо великий ($n \geq 50$);

в) обчислене значення вибіркового коефіцієнта кореляції невелике ($|\rho| \leq 0,5$).

Якщо ця передумова має місце, то вибірковий коефіцієнт кореляції r має приблизно нормальний розподіл з математичним сподіванням, що дорівнює коефіцієнту кореляції генеральної сукупності ρ, та дисперсією $\sigma_r = (1-r^2)/\sqrt{n}$. Це безпосередньо можна помітити, аналізуючи графік густини розподілу вибіркового коефіцієнта кореляції (рис. 6.9). Звідси випливає: якщо нульова гіпотеза $\{H_0 : \ \rho = \rho_0\}$ вірна, то статистика

$$u \ = \ \frac{r - \rho_0}{\sqrt{(1-\rho_0^2)/\sqrt{n}}} \tag{6.49}$$

має приблизно нормальний розподіл з нульовим математичним сподіванням і дисперсією, що дорівнює одиниці. Вибиремо $u_{\alpha/2}$ – критичне значення стандартизованої нормальної випадкової величини, яка задовольняє умові $\Pr(|u| \geq u_{\alpha/2}) = \alpha$.

Якщо

$$|u_{\text{спос}}| \ \geq \ u_{\alpha/2}, \tag{6.50a}$$

то гіпотеза $\{H_0 : \rho = \rho_0\}$ відхиляється.

Якщо ж

$$|u_{\text{спос}}| \ < \ u_{\alpha/2}, \tag{6.50b}$$

то відсутні причини відхилити нульову гіпотезу.

Застосування цієї моделі дозволяє також знаходити наближені довірчі інтервали для коефіцієнта кореляції ρ генеральної сукупності за формулою

$$\left(r - u_{\alpha/2}\sqrt{(1-\rho_0^2)/\sqrt{n}}\right) \ < \ \rho \ < \ \left(r + u_{\alpha/2}\sqrt{(1-\rho_0^2)/\sqrt{n}}\right). \tag{6.51}$$

Якщо виявиться, що обчислений довірчий інтервал не накриває значення $\rho = \rho_0$, то гіпотеза $\{H_0 : \ \rho = \rho_0\}$ відхиляється, в протилежному випадку дані експерименту не дозволяють відхилити нульову гіпотезу.

Модель 3.

Вихідні передумови та обмеження:

а) двовимірний закон розподілу змінних (X, Y), що досліджуються, в генеральній сукупності передбачається нормальним;

б) обсяг вибірки n достатньо великий ($n \geq 50$).

Для перевірки гіпотези про силу кореляційного зв'язку обчислюється статистика (перетворення Фішера)

$$z \ = \ \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \qquad \text{тобто} \qquad r \ = \ \text{th}z \ = \ \frac{\text{e}^{2z}-1}{\text{e}^{2z}+1}. \tag{6.52}$$

Можна показати, що розподіл статистики z достатньо добре апроксимується нормальним розподілом з математичним сподіванням

$$\text{M[Z]} \ = \ \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{2(n-1)} \tag{6.53}$$

і дисперсією, яка не залежить від ρ,

$$\text{D[Z]} \ = \ \frac{1}{n-3}. \tag{6.54}$$

Статистика

$$u \;=\; \left(1,1513\lg\frac{1+r}{1-r} - 1,1513\lg\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\right)\sqrt{n-3} \qquad (6.55)$$

має асимптотично нормальний розподіл з нульовим математичним сподіванням і дисперсією, що дорівнює одиниці.

Перевірка гіпотези $\{H_0 : \rho = \rho_0\}$ проводиться за тією ж схемою, що і в моделі 2, тобто за формулою (6.55) обчислюється емпіричне значення $u_{\text{спос}}$ статистики u. Користуючись таблицею функції Лапласа, за фіксованим рівнем значущості α знаходять критичне значення $u_{\alpha/2}$, яке задовольняє умові $\Pr(|u| \geq u_{\alpha/2}) = \alpha$.

Якщо обчислене значення статистики задовольняє нерівності

$$|u_{\text{спос}}| \;\geq\; u_{\alpha/2}, \qquad (6.56a)$$

то гіпотеза $\{H_0 :\; \rho = \rho_0\}$ відхиляється.

Якщо ж

$$|u_{\text{спос}}| \;<\; u_{\alpha/2}, \qquad (6.56b)$$

то відсутні причини відхилити нульову гіпотезу.

Зауваження.

Модель 3 є найбільш загальною, оскільки може використовуватися за будь-яких значень n, ρ та r. Але коли перевіряється нульова гіпотеза $\{H_0 :\; \rho = 0\}$, то доцільніше використати модель 1. Якщо ж обчислене значення вибіркового коефіцієнта кореляції не є дуже великим $(|r| < 0,5)$ та обсяг вибірки достатньо великий $(n \geq 50)$, то слід обирати модель 2.

6.6. Оцінка точності знаходження точкових оцінок коефіцієнтів лінійного рівняння регресії

Після знаходження емпіричного рівняння $\overline{y}_x = b_0 + b_1 x$ регресії Y на X обчислюють середню квадратичну похибку $s_e \equiv s_{y.x}$. Отримана величина $s_{y.x}$ буде характеризувати ступінь розсіювання експериментальних точок навколо лінії регресії, тобто ступінь розсіювання залежної змінної Y, яка звільнена від впливу іншої змінної X :

$$s_e \;\equiv\; s_{y.x} \;=\; \left(\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \overline{y}_x)^2\right)^{1/2}. \qquad (6.57)$$

Середні квадратичні похибки σ_{b_0} та σ_{b_1} визначення коефіцієнтів b_0 та b_1 наступні :

$$\sigma^2_{b_0} \;=\; \sigma^2_{y.x}\left(\frac{1}{n} + \overline{x^2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{-1}\right), \qquad (6.58a)$$

$$\sigma^2_{b_1} \;=\; \sigma^2_{y.x}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{-1}. \qquad (6.58b)$$

Якщо замінити в цих формулах невідому дисперсію $\sigma^2_{y.x}$ її незсуненою оцінкою $s^2_{y.x}$, то після перетворень отримаємо емпіричні дисперсії коефіцієнтів β_0 та β_1 лінійного рівняння регресії $\overline{y}_x = b_0 + b_1 x$:

$$\sigma^2_{b_0} = \left(s^2_{y.x}\sum_{i=1}^{n} x_i^2\right)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{-1}, \tag{6.59}$$

$$\sigma^2_{b_1} = \left(ns^2_{y.x}\right)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{-1}. \tag{6.60}$$

Квадратний корінь з цих значень називається *середньою квадратичною помилкою знаходження оцінок* b_0 *та* b_1.

Середня квадратична помилка вказує, наскільки в середньому оцінки коефіцієнтів b_0 та b_1 відрізняються від модельних коефіцієнтів регресії β_0 та β_1.

Методи регресійного аналізу будуть змістовними і коректними лише за виконання деяких вимог, які ставляться до збору статистичного матеріалу в експерименті.

Якщо гіпотеза про нормальний закон розподілу системи випадкових величин (X, Y) справедлива, то:

а) математичне сподівання $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \beta_0 + \beta_1 x$;

б) умовна дисперсія $\mathrm{D}[\mathrm{Y}|\mathrm{X}= x]$ постійна для всіх значень x і дорівнює $\sigma^2_{y.x} = \sigma^2_y(1-\rho^2)$;

в) розподіл $f_{\mathrm{Y}}(y|x)$ нормальний;

г) спостереження $\{(x_i, y_i)\}$, $(i = 1, 2, \ldots, n)$ незалежні.

Якщо виявляється можливим прийняти, що нормальна модель ряду спостережень $\{(x_i, y_i)\}$ добре віддзеркалює закономірності двовимірної величини (X, Y), що досліджується, тобто перелічені умови а) — г) виконуються, то в такому випадку оцінки коефіцієнтів регресії b_0 та b_1 в практичних розрахунках описують (приблизно) нормально розподіленими випадковими величинами з математичними сподіваннями β_0 й β_1 та дисперсіями $\sigma^2_{b_0}$ й $\sigma^2_{b_1}$.

Якщо при цьому величина $\sigma^2_{y.x}$ є відомою, то статистики $u_0 = (b_0 - \beta_0)/\sigma^2_{y.x}$ та $u_1 = (b_1 - \beta_1)/\sigma^2_{y.x}$ розподілені за нормальним законом з нульовим математичним сподіванням і дисперсією, що дорівнює одиниці.

Статистики u_j, де $j = 0,\ 1,$ можна використовувати для побудови інтервальних оцінок коефіцієнтів. У цьому випадку довірчі інтервали, що вишукуються, для коефіцієнтів істинного рівняння регресії будуть знаходитися за заданою довірчою ймовірністю $p = 1 - \alpha$ (або заданим рівнем значущості α) за формулами

$$\Pr\left(-u_{\alpha/2} < \frac{b_j - \beta_j}{\sigma_{y.x}\sqrt{c_{jj}}} < u_{\alpha/2}\right) = 1 - \alpha, \qquad j = 0, 1, \tag{6.61}$$

де $u_{\alpha/2}$ – квантиль нормального розподілу $N(0;\ 1)$;

$$c_{00} = \frac{\sum\limits_{i=1}^{n} x_i^2}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}, \qquad c_{11} = \frac{n}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}. \tag{6.62}$$

Перетворюючи вираз (6.61), отримаємо

$$\Pr\left(b_j - u_{\alpha/2}\sigma_{y.x}\sqrt{c_{jj}} < \beta_j < b_j + u_{\alpha/2}\sigma_{y.x}\sqrt{c_{jj}}\right) \;=\; 1-\alpha, \qquad j=0,1. \tag{6.63}$$

При проведенні експериментальних робіт величина $\sigma_{y.x}$ невідома. Припустимо, що за допомогою метода найменших квадратів була знайдена незсунена оцінка $\hat{\sigma}_{y.x}$ величини $\sigma_{y.x}$ за формулою

$$\hat{\sigma}_{y.x} \;=\; s_{y.x} \;=\; \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x)^2. \tag{6.64}$$

Тоді статистика ($j=0,\,1$)

$$t \;=\; \frac{b_j-\beta_j}{s_{b_j}} \;=\; \frac{b_j-\beta_j}{s_{y.x}\sqrt{c_{jj}}} \tag{6.65}$$

має розподіл Стьюдента з $\nu = n-2$ ступенями вільності. Цю статистику можна використати для побудови інтервальних оцінок коефіцієнтів лінійного рівняння регресії.

Розглянемо $100(1-\alpha)\%$-й довірчий інтервал для коефіцієнтів β_0 і β_1. Для його побудови за таблицею t-розподілу Стьюдента за кількістю ступенів вільності $\nu = n-2$ та довірчою ймовірністю $p = 1-\alpha$ знаходимо значення $t_{\alpha/2;\,n-2}$, що задовольняє умові

$$\Pr\left(|t| < t_{\alpha/2;\,n-2}\right) \;=\; 1-\alpha. \tag{6.66}$$

Після простих перетворень отримаємо

$$\Pr\left(-t_{\alpha/2;\,n-2} < \frac{b_0-\beta_0}{s_{y.x}\sqrt{c_{00}}} < t_{\alpha/2;\,n-2}\right) \;=\; 1-\alpha, \tag{6.67}$$

$$\Pr\left(-t_{\alpha/2;\,n-2} < \frac{b_1-\beta_1}{s_{y.x}\sqrt{c_{11}}} < t_{\alpha/2;\,n-2}\right) = 1-\alpha. \tag{6.68}$$

В результаті перетворень двох нерівностей в круглих дужках знайдемо для $100(1-\alpha)\%$-го довірчого інтервалу для коефіцієнтів лінійної регресії β_0 та β_1:

$$\left(b_0 - t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{00}}\right) \;<\; \beta_0 \;<\; \left(b_0 + t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{00}}\right), \tag{6.69}$$

$$\left(b_1 - t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{11}}\right) \;<\; \beta_1 \;<\; \left(b_1 + t_{\alpha/2;\,n-2}s_{y.x}\sqrt{c_{11}}\right). \tag{6.70}$$

Знайденими виразами часто користуються при розв'язанні проблем прогнозування часових залежностей (разом з пред'явленням оцінки прогнозних похибок).

6.7. Лінійний регресійний аналіз між двома змінними

У цьому розділі розглянемо два приклада застосування регресійного аналіза.

В процесі виконання практичних робіт і розв'язання задач лінійного регресійного аналізу між двома змінними потрібно пам'ятати про відповідність між характеристиками, що стосуються теоретико-ймовірнісних моделей, та відповідно статистичними харатеристиками до їх вибіркових аналогів.

Головні з вказаних величин наведені в таблиці-зведенні.

Зведення чисельних характеристик генеральної та вибіркової сукупностей

Генеральна сукупність	Вибіркова сукупність
$\mathrm{M[X]} = m_{\mathrm{X}} = \int\int x f(x,y)\,dx\,dy = \int x f(x)\,dx$	$\overline{x} = \frac{1}{n}\sum_i x_i$
$\mathrm{M[Y]} = m_{\mathrm{Y}} = \int\int y f(x,y)\,dx\,dy = \int y f(y)\,dy$	$\overline{y} = \frac{1}{n}\sum_j y_j$
$\sigma_{\mathrm{X}}^2 = \int\int (x-m_{\mathrm{X}})^2 f(x,y)\,dx\,dy = \int (x-m_{\mathrm{X}})^2 f(x)\,dx$	$s_{\mathrm{X}}^2 = \frac{1}{n}\sum_i (x_i-\overline{x})^2 = \overline{x^2}-\overline{x}^2$
$\sigma_{\mathrm{Y}}^2 = \int\int (y-m_{\mathrm{Y}})^2 f(x,y)\,dx\,dy = \int (y-m_{\mathrm{Y}})^2 f(y)\,dy$	$s_{\mathrm{Y}}^2 = \frac{1}{n}\sum_j (y_j-\overline{y})^2 = \overline{y^2}-\overline{y}^2$
$\mu_{\mathrm{XY}} = \int\int xy f(x,y)\,dx\,dy - \mathrm{M[X]M[Y]}$	$K_{\mathrm{XY}} = \frac{1}{n}\sum_i\sum_j (x_i-\overline{x})(y_j-\overline{y})$
$\rho = \frac{\mu_{\mathrm{XY}}}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}}$	$r = \frac{K_{\mathrm{XY}}}{s_{\mathrm{X}}s_{\mathrm{Y}}}$
Модельне рівняння регресії Y на X $\mathrm{M[Y\|X}=x] = m_{\mathrm{Y}} + \rho\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}}(x-m_{\mathrm{X}})$	Емпіричне рівняння регресії Y на X $\overline{y}_x = \overline{y} + r\frac{s_{\mathrm{Y}}}{s_{\mathrm{X}}}(x-\overline{x})$
Модельне рівняння регресії X на Y $\mathrm{M[X\|Y}=y] = m_{\mathrm{X}} + \rho\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}}(y-m_{\mathrm{Y}})$	Емпіричне рівняння регресії X на Y $\overline{x}_y = \overline{x} + r\frac{s_{\mathrm{X}}}{s_{\mathrm{Y}}}(y-\overline{y})$

Приклад 1

У таблиці наведено результати $n = 11$ вимірювань значень x_i номіналів довжин моделей (ознака X) і ширин моделей y_i (ознака Y).

Для даних, що наведено, передбачається, що між ознаками X та Y є лінійна регресійна залежність $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \beta_0 + \beta_1 x$.

Таблиця 6.2 — Дані до прикладу 1

i	1	2	3	4	5	6
x_i, *мм*	0,90	1,22	1,32	0,77	1,30	1,20
y_i, *мм*	–0,30	0,10	0,70	–0,28	–0,25	0,02
i	7	8	9	10	11	
x_i, *мм*	1,32	0,95	1,45	1,30	1,20	
y_i, *мм*	0,37	–0,70	0,55	0,35	0,32	

Вимагається:

1) оцінити параметри β_0 та β_1 модельного рівняння регресії методом найменших квадратів;

2) знайти середні квадратичні помилки коефіцієнтів знайденого емпіричного рівняння регресії;

3) побудувати 95%-ві довірчі інтервали для коефіцієнтів лінійного рівняння регресії β_0 та β_1;

4) користуючись емпіричним рівнянням регресії, знайти точкову оцінку відхилення від номінального розміру довжини моделі, якщо ширина моделі відхиляється від номінального розміру на величину $x_* = 1,1$ *мм*;

5) обчислити коефіцієнт кореляції та коефіцієнт детермінації; пояснити сенс коефіцієнта детермінації.

Розв'язання

1. Спочатку обчислимо допоміжні суми:

$$\sum_{i=1}^{n} x_i = 12,93; \qquad \sum_{i=1}^{n} y_i = 0,88;$$

$$\sum_{i=1}^{n} x_i y_i = 1,7193; \qquad \sum_{i=1}^{n} x_i^2 = 15,6411;$$

Знайдемо вибіркові середні:

$$\overline{x} = \frac{1}{11} \cdot 12,93 = 1,175; \qquad \overline{y} = \frac{1}{11} \cdot 0,88 = 0,080.$$

Знайдемо значення оцінок коефіцієнтів b_0 та b_1 істинного рівняння регресії за формулами

$$b_1 = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2} = \frac{11 \cdot 1,7193 - 12,93 \cdot 0,88}{11 \cdot 15,6411 - 12,93^2} = 1,5479;$$

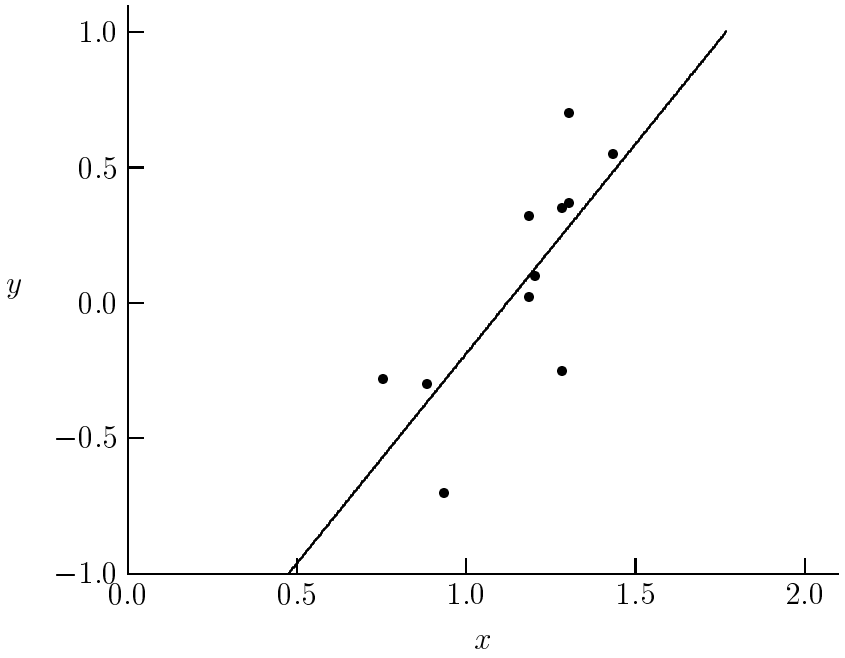


Рисунок 6.11 — Діаграма розсіювання експериментальних точок; вказано емпіричне рівняння регресії $\overline{y}_x = -1,7388 + 1,5479\overline{x}$

$$b_0 = \overline{y} - b_1\overline{x} = 0,08 - 1,5479 \cdot 1,175 = -1,7388.$$

Отже, емпіричне рівняння регресії має такий вигляд:

$$\overline{y}_x = -1,7388 + 1,5479\overline{x}.$$

Графік отриманого емпіричного рівняння регресії наведено на рис. 6.11.

2. Для знаходження середніх квадратичних помилок σ_{b_0} та σ_{b_1}, які характеризують точність здобутих коефіцієнтів b_0 та b_1 емпіричного рівняння регресії, обчислимо спочатку середню квадратичну помилку, що характеризує розсіювання емпіричних точок навколо лінії регресії. Для цього виконаємо оцінку значень залежної змінної за формулою $\overline{y}_x = -1,7388 + 1,5479x$ та обчислимо відхилення

$$e_i = \overline{y}_x + 1,7388 - 1,5479x_i.$$

В результаті знайдемо $\sum_{i=1}^{11} e_i^2 = 0,7561$.

Знайдемо незсунену оцінку дисперсії залежної змінної Y, звільненої від впливу змінної X, за формулою

$$\hat{\sigma}^2_{y.x} = s^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(\overline{y}_x + 1,7388 - 1,5479x_i)^2 = \frac{0,7561}{9} = 0,08401.$$

Обчислимо емпіричні дисперсії точкових оцінок коефіцієнтів регресії:

$$s^2_{b_0} = \left(s^2_{y.x}\sum_{i=1}^{n} x_i^2\right)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{-1} = \frac{15,6411 \cdot 0,08401}{11 \cdot 15,6411 - 12,93^2} = 0,26997;$$

$$s^2_{b_1} \;=\; \left(ns^2_{y.x}\right)\left[n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]^{-1} \;=\; \frac{11\cdot 0,08401}{11\cdot 15,6411-12,93^2} \;=\; 0,18986.$$

Тому одержуємо: $s_{b_0}=0,520;\;\; s_{b_1}=0,436.$

3. Для побудови 95%–х довірчих інтервалів для коефіцієнтів лінійного рівняння регресії β_0 та β_1 за таблицею t–розподілу Стьюдента за кількістю ступенів вільності $\nu=n-2=9$ та довірчою ймовірністю $p=1-\alpha=0{,}95$ знаходимо критичне значення статистики (квантиль) $t_{\alpha/2;\,n-2}=t_{0{,}025;\,9}=2{,}262$.

Користуючись формулою

$$\left(b_0-t_{\alpha/2;\,n-2}s_{b_0}\right) \;<\; \beta_0 \;<\; \left(b_0+t_{\alpha/2;\,n-2}s_{b_0}\right),$$

знаходимо 95%-й довірчий інтервал для коефіцієнта b_0

$$b_0\pm t_{0{,}025;\,9}s_{b_0} \;=\; -1,7388\pm(2,262\cdot 0,520) \;=\; [-2,9150;\;-0,5626].$$

Тому одержуємо: $-2,9150<b_0<-0,5626.$

Аналогічно, користуючись формулою

$$\left(b_1-t_{\alpha/2;\,n-2}s_{b_1}\right) \;<\; \beta_1 \;<\; \left(b_1+t_{\alpha/2;\,n-2}s_{b_1}\right),$$

знаходимо 95%-й довірчий інтервал для коефіцієнта b_1

$$b_1\pm t_{0{,}025;\,9}s_{b_1} \;=\; 1,5479\pm(2,262\cdot 0,436) \;=\; 1,5479\pm 0,9885.$$

Тому одержуємо: $0,5594 \;<\; b_1 \;<\; 2,5364.$

4. Знайдемо точкову оцінку $\overline{y}_{x^*}$ відхилення від номінального розміру довжини моделі, якщо ширина моделі відхиляється від номінального розміру на величину $x_*=1,1\,мм$:

$$\overline{y}_{x^*} \;=\; 1,7394-1,5479\cdot 1,10 \;=\; 0,036\;мм.$$

5. Обчислимо коефіцієнт кореляції

$$r \;=\; \frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}} \;=$$

$$=\; \frac{11\cdot 1,7193-12,93\cdot 0,88}{\sqrt{11\cdot 15,6411-12,93^2}\sqrt{11\cdot 1,8856-0,88^2}} \;=\; 0,7644.$$

Тому маємо, що коефіцієнт детермінації дорівнює

$$k_{\rm det} \;=\; r^2 \;=\; 0,584.$$

Цей отриманий результат означає, що 58,4% розсіювання залежної змінної Y пояснюється лінійною регресією Y на X, а 41,5% розсіювання Y можуть бути обумовлені або випадковими помилками експерименту, або тим, що лінійна регресійна модель недобре погоджується з експериментальними даними.

Таблиця 6.3 — Дані до прикладу 2

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
81	77	54	81	100	129	94	104	84	96
77	96	40	57	95	145	84	108	94	112
76	86	61	86	106	142	73	93	152	136
86	92	68	87	118	120	107	124	98	104
53	98	53	98	109	95	94	112	77	103
47	53	88	87	107	107	107	113	88	115
36	63	136	153	120	133	99	95	94	123
40	80	129	133	114	140	100	112	76	111
49	64	126	159	113	149	104	116	84	127
60	66	96	134	123	147	88	93	73	129

Приклад 2

В таблиці наведено результати $n = 50$ вимірювань значень x_i ознаки X та значень y_i ознаки Y.

Користуючись значеннями, що наведені в таблиці, потрібно:

1) скласти кореляційну таблицю;

2) знайти за допомогою кореляційної таблиці числові характеристики вибірки $\overline{x}$, $\overline{y}$, s_{X}, s_{Y}, K_{XY}, r;

3) побудувати кореляційне поле; за характером розташування точок на кореляційному полі підібрати загальний вид функції регресії;

4) знайти параметри емпіричної лінійної функції регресії Y на X й X на Y і навести їх графіки.

Розв'язання

1. Складемо кореляційну таблицю.

Візьмемо для ознаки X такі межі інтервалів: (30–50), (50–70), ..., (130–150), а для ознаки Y — (50–70), (70–90), ..., (150–170). Таким чином, довжини інтервалів містять $h_x = h_y = 20$. Після цього подраховуємо кількість експериментальних точок, які влучили в прямокутники, утворені межами інтервалів.

Таблиця 6.4 — Кореляційна таблиця

Y	X						
	40	60	80	100	120	140	n_y
160					1	1	2
140				3	5	1	9
120			4	8	1		13
100		2	7	5			14
80	1	3	3				7
60	4	1					5
n_x	5	6	14	16	7	2	$n = 50$

В результаті одержуємо кореляційну таблицю, в якій відмічені середини

відповідних інтервалів.

2. Для знаходження емпіричних рівнянь регресії обчислимо:
середні арифметичні

$$\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; 88,62; \qquad \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \;=\; 107,66;$$

вибіркові дисперсії

$$s_x^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; 678,796; \qquad s_y^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; 668,184;$$

вибіркові середні квадратичні відхилення

$$s_x \;=\; \sqrt{678,796} \;=\; 26,054; \qquad s_y \;=\; \sqrt{668,184} \;=\; 25,849;$$

вибірковий кореляційний момент

$$K_{xy} \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \;=\; \overline{xy}-\overline{x}\,\overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} \;=\; 546,191;$$

вибірковий коефіцієнт кореляції

$$r \;=\; \frac{K_{xy}}{s_x s_y} \;=\; \frac{546,191}{26,054\cdot 25,849} \;=\; 0,811.$$

Перевіримо значущість отриманого вибіркового коефіцієнта кореляції. Для цього обчислимо статистику:

$$t_{\text{спос}} \;=\; \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \;=\; \frac{0,811\,\sqrt{50-2}}{\sqrt{1-0,811^2}} \;=\; 9,604.$$

Знайдемо за таблицею квантилів розподілу Стьюдента з рівнем значущості $\alpha \;=\; 0,05$ та кількістю ступенів вільності $\nu \;=\; n-2 \;=\; 48$ значення квантиля $t_{\alpha/2;\,n-2} \;=\; t_{0,025;\,48} = 2{,}02$.

Порівнюючи отримане значення $t_{\text{спос}} = 9{,}604$ з квантилем $t_{0,025;\,48} = 2{,}02$, маємо $t_{\text{спос}} > t_{0,025;\,48}$. Отже, можна зробити висновок, що лінійна регресійна модель вигляду $\mathrm{M}[\mathrm{Y}|\mathrm{X}= x] = \beta_0 + \beta_1 x$ обрана вдало, тобто вона погоджується з експериментальними даними.

3. Для підтвердження існування лінійної регресійної залежності між змінними X та Y, що досліджуються, побудуємо кореляційне поле.

Результати вимірювань $\{(x_i, y_i)\}$, $(i = 1, 2, \ldots, 50)$ наведемо у вигляді точок у декартовій системі координат (рис. 6.12). Візуальна оцінка розташування точок на кореляційному полі дозволяє прийняти гіпотезу про лінійну регресійну залежність між ознаками X та Y.

4. Знайдемо значення параметрів емпіричного рівняння регресії Y на X

$$\overline{y}_x \;=\; \overline{y} + r\frac{s_y}{s_x}\,(x-\overline{x}) \;=\; 107,66 + 0,811\,\frac{25,849}{26,054}\,(x-88,62) \;=\; 36,352 + 0,805x$$

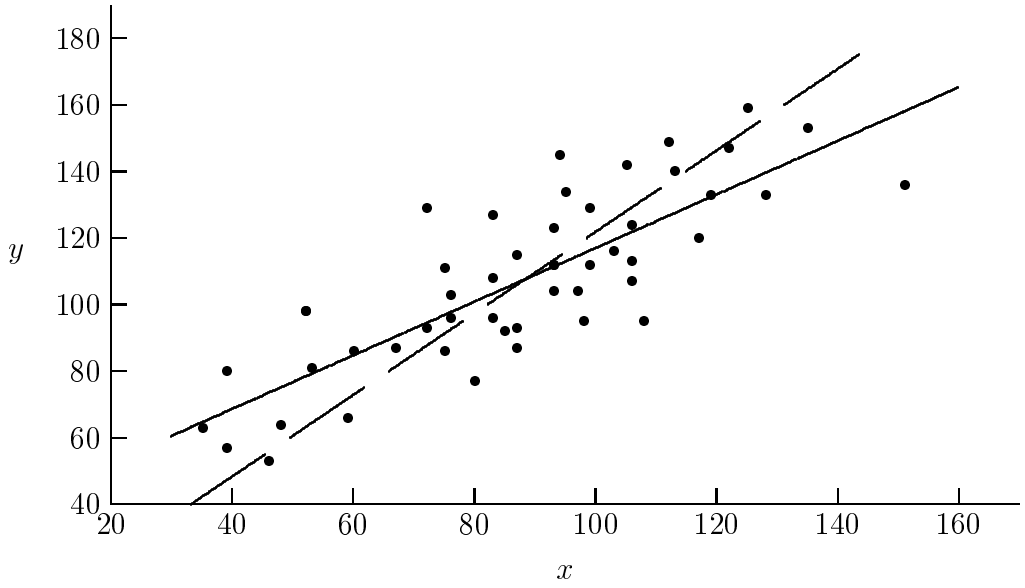


Рисунок 6.12 — Кореляційне поле експериментальних точок; суцільною лінією вказано рівняння регресії $\overline{y}_x = 36,352 + 0,805x$; пунктирною лінією вказано рівняння регресії $\overline{x}_y = 0,616 + 0,817y$

і значення параметрів емпіричного рівняння регресії X на Y

$$\overline{x}_y = \overline{x} + r\frac{s_x}{s_y}(y - \overline{y}) = 88,62 + 0,811\,\frac{26,054}{25,849}(y - 107,66) = 0,616 + 0,817y.$$

Контроль обчислень: $a_1 b_1 = 0,805 \cdot 0,817 \approx 0,65 \approx r^2$.

Графіки отриманих емпіричних функцій лінійної регресії наведено на рис. 6.12.

6.8. Приклади

Приклад 6.1

Знайти рівняння регресії Y на X і X на Y за чотирма парами наведених значень випадкової величини (X, Y):

x_i	1	2	3	4
y_i	2	4	5	7

Розв'язання

На рис. 6.13 нанесено точки (x_i, y_i) $i = 1, 2, 3, 4$. Аналізуючи їх розташування, помічаємо, що вони групуються навколо прямої лінії. Тому маємо підібрати рівняння регресії лінійного вигляду. Необхідні обчислення розташуємо в таблиці.

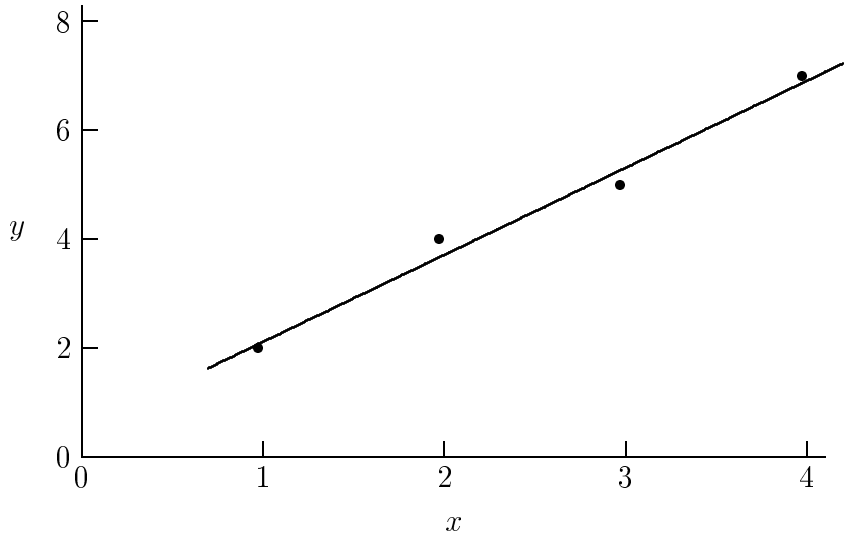


Рисунок 6.13 — До побудови лінійної регресії

i	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	1	2	2	1	4
2	2	4	8	4	16
3	3	5	15	9	25
4	4	7	28	16	49
Суми	10	18	53	30	94

Підставляючи знайдені суми в систему нормальних рівнянь (6.23), маємо

$$\begin{cases} 4b_0 + 10b_1 = 18; \\ 10b_0 + 30b_1 = 53. \end{cases}$$

Розв'язуючи цю систему, знаходимо: $b_0 = 0{,}5$; $b_1 = 1{,}6$.
Отже, рівняння регресії Y на X має вигляд

$$\overline{y}_x = 0,5 + 1,6x.$$

Знайдемо тепер рівняння регресії X на Y вигляду (6.25). Використовуючи значення вже знайдених сум, маємо

$$\begin{cases} 4a_0 + 18a_1 = 10, \\ 18a_0 + 94a_1 = 53. \end{cases}$$

Розв'язуючи систему, знаходимо: $a_0 = -7/26$; $a_1 = 8/13$.
Отже, рівняння регресії X на Y має вигляд

$$\overline{x}_y = -0,269 + 0,615y.$$

Приклад 6.2
Розподіл ознак X і Y наводиться в наступній кореляційній таблиці:

X	Y									
	0	10	20	30	40	50	60	70	80	m_x
–2				1	2	1				4
–1			1	3		3	1			8
0		2	4				4	2		12
1	1	5						5	1	12
2	3								3	6
m_y	4	7	5	4	2	4	5	7	4	42

Знайти кореляційні відношення $\eta_{\mathrm{Y|X}}$ та $\eta_{\mathrm{X|Y}}$ і порівняти їх з відповідним коефіцієнтом лінійної регресії.

Розв'язання

Розрахунок кореляційних відношень $\eta_{\mathrm{y|x}}$ та $\eta_{\mathrm{y|x}}$ проведемо за наступною схемою:

$$\overline{\mathrm{X}} = \frac{\sum_x m_x x}{\sum_x m_x} = \frac{8}{42} = 0,190;$$

$$\overline{\mathrm{X^2}} = \frac{\sum_x m_x x^2}{\sum_x m_x} = \frac{60}{42} = 1,429; \qquad \sigma^2_{\mathrm{X}} = \overline{\mathrm{X^2}} - (\overline{\mathrm{X}})^2 = 1,392.$$

З початкових даних випливає, що $y_0 = 40$ і $\beta = 0{,}1$.

Скористаємося центрованою величиною $\mathrm{V} = (\mathrm{Y} - 40)/10$. Для неї

$$\overline{\mathrm{V}} = \frac{\sum_y m_y v_y}{\sum_y m_y} = \frac{0}{42} = 0; \qquad \overline{\mathrm{V^2}} = \frac{\sum_y m_y v_y^2}{\sum_y m_y} = \frac{302}{42} = 7,190;$$

та

$$\sigma^2_{\mathrm{V}} = \overline{\mathrm{V^2}} - (\overline{\mathrm{V}})^2 = 7,190.$$

Далі

$$\overline{(\overline{\mathrm{V}}_x)^2} = \frac{\sum_x \frac{1}{m_x}\left(\sum_y m_{xy} v_y\right)^2}{\sum_x m_x} = \frac{0}{42} = 0,$$

$$\sigma^2(\overline{\mathrm{V}}_x) = \overline{(\overline{\mathrm{V}}_x)^2} - (\overline{\mathrm{V}})^2 = 0,$$

а також

$$\overline{(\overline{\mathrm{X}}_v)^2} = \frac{\sum_y \frac{1}{m_y}\left(\sum_x m_{xy} x\right)^2}{\sum_y m_y} = \frac{52,542}{42} = 1,251;$$

$$\sigma^2(\overline{\mathrm{X}}_v) = \overline{(\overline{\mathrm{X}}_v)^2} - (\overline{\mathrm{X}})^2 = 1,215.$$

Звідси

$$\eta_{\mathrm{Y|X}} = \eta_{\mathrm{V|X}} = \sqrt{\frac{\sigma^2\left(\overline{\mathrm{V}}_x\right)}{\sigma^2_{\mathrm{Y}}}} = \sqrt{\frac{0}{7,190}} = 0,$$

а також

$$\eta_{\mathrm{X|Y}} = \eta_{\mathrm{X|V}} = \sqrt{\frac{\sigma^2\left(\overline{\mathrm{X}}_v\right)}{\sigma^2_{\mathrm{X}}}} = \sqrt{\frac{1,215}{1,393}} = \sqrt{0,872} = 0,934.$$

Очевидно, що $\rho_{Y|X} = \rho_{X|Y} = 0$ та $r(X, Y) = 0$.

Приклад 6.3

Нехай потрібно знайти лінійні рівняння регресії Y на X і X на Y за 10 парами наведених значень випадкової величини (X, Y) – точками (x_i, y_i) $(i = 1, 2, \ldots, 10)$.

Номер вимірювання i	1	2	3	4	5	6	7	8	9	10
x_i	1	1	1	2	2	2	3	3	3	4
y_i	3	3	3	4	4	5	5	5	6	7

Розв'язання

Якщо побудувати кореляційне поле, то візуально можна погодитися з тим, що точки (x_i, y_i) $(i = 1, 2, \ldots, 10)$ групуються біля прямої лінії. Тому маємо підібрати рівняння регресії лінійного вигляду. Обчислимо суми, що входять до системи нормальних рівнянь (6.23) і (6.25), не виконуючи угрупування експериментальних даних.

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	1	3	3	1	9
2	1	3	3	1	9
3	1	3	3	1	9
4	2	4	8	4	16
5	2	4	8	4	16
6	2	5	10	4	25
7	3	5	15	9	25
8	3	5	15	9	25
9	3	6	18	9	36
10	4	7	28	16	49
Суми	22	45	111	58	219

Представимо ці ж дані у вигляді кореляційної таблиці.

x_i	y_i					n_x
1	3					3
2		2	1			3
3			2	1		3
4					1	1
n_y	3	2	3	1	1	$n = 10$

Знайдемо рівняння регресії Y на X. Підставляючи знайдені суми в систему (6.23), маємо

$$\begin{cases} 10b_0 + 22b_1 = 45; \\ 22b_0 + 58b_1 = 111. \end{cases}$$

Розв'язуючи систему, знаходимо $b_0 = 1{,}75$ та $b_1 = 1{,}25$.

Отже, рівняння регресії Y на X має вигляд

$$\overline{y}_x = 1,75 + 1,25x.$$

Знайдемо рівняння регресії X на Y. Підставляючи знайдені суми в систему (6.25), маємо

$$\begin{cases} 10a_0 + 45a_1 = 22, \\ 45a_0 + 219a_1 = 111. \end{cases}$$

Розв'язуючи систему, знаходимо a_0 = -1,073 та a_1 = 1,727.

Отже, рівняння регресії X на Y має вигляд

$$\overline{x}_y = -1,073 + 0,727y.$$

Приклад 6.4

Є дві вибірки $(x_1, x_2, \ldots, x_n)$ та $(y_1, y_2 \ldots, y_n)$ обсягом n кожна. Побудувати систему нормальних рівнянь, вважаючи, що кореляційна залежність ознаки Y на ознаку X має криволінійний параболічний вигляд $\overline{y}_x = a + bx + cx^2$ (рис. 6.14).

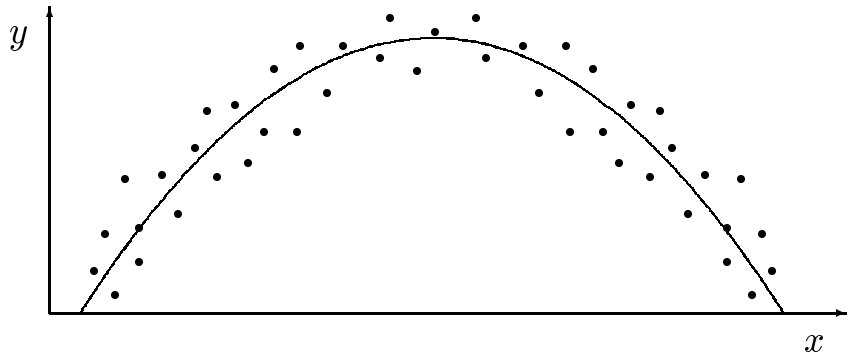


Рисунок 6.14 — Приклад параболічної регресії

Розв'язання

На основі методу найменших квадратів параметри a, b, c будемо визначати виходячи з мінімуму функціонала нев'язки

$$\sum_{i=1}^{n} (y_i - \overline{y}_x)^2 \Rightarrow \min.$$

Якщо підставити в це співвідношення параболічну залежність $\overline{y}_x = a + bx + cx^2$, то після диференціювання за параметрами a, b, c отримаємо шукану систему

$$\begin{cases} an + b\sum\limits_{i=1}^{n} x_i + c\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i; \\ a\sum_i x_i + b\sum\limits_{i=1}^{n} x_i^2 + c\sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} y_i x_i; \\ a\sum_i x_i^2 + b\sum\limits_{i=1}^{n} x_i^3 + c\sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} y_i x_i^2. \end{cases}$$

Аналогічно можна побудувати систему нормальних рівнянь і в тому випадку, коли передбачувана кореляційна залежність Y на X має криволінійний параболічний вигляд порядку, вищого ніж 2.

Приклад 6.5

Є дані про розчинність азотнокислого натрію $NaNO_3$ залежно від температури води. У 100 частинах води розчиняється наступне число умовних частин $NaNO_3$ при відповідних температурах:

$t°C$	0	4	10	15	21	29	36	51	68
$NaNO_3$	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	125,1

Передбачаючи, що кількість $NaNO_3$ (випадкова величина Y), яка розчиняється в 100 частинах води, залежить лінійно від температури (випадкова величина X) розчину, знайти параметри a та b у формулі $y = ax + b$ за методом найменших квадратів.

Розв'язання

Для знаходження параметрів a та b за методом найменших квадратів необхідно розв'язати систему

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i; \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i, \end{cases}$$

в якій в цьому випадку x_i – температура розчину; y_i – кількість $NaNO_3$, що розчиняється в 100 частинах води при даній температурі.

Після знаходження коефіцієнтів системи отримаємо

$$\begin{cases} 10144a + 234b = 24628,6; \\ 234a + 9b = 811,3. \end{cases}$$

Звідси $a = 0,87$; $b = 67,5$.

Отже, залежність Y від X має вигляд $y = 0,87\,x + 67,5$.

Приклад 6.6

З генеральної сукупності, розподіл ознак X і Y в якій нормальний, зроблена вибірка обсягом в $N = 530$ одиниць. Результати вимірювання ознак X та Y у компонент вибірки наводяться в наступній таблиці:

X	Y							
	15–25	25–35	35–45	45–55	55–65	65–75	75–85	m_y
200–300	19	5						24
300–400	23	116	11					150
400–500	1	41	98	9				149
500–600		4	32	65	7			108
600–700		1	4	21	36	3		65
700–800			1	2	11	13	1	28
800–900					1	3	2	6
m_x	43	167	146	97	55	19	3	530

Знайти вибірковий коефіцієнт лінійної кореляції і його середнє квадратичне відхилення. Написати рівняння прямих регресій Y на X й X на Y.

Розв'язання

Спочатку інтервали значень ознак X і Y замінимо їх серединами.

Оскільки коефіцієнт лінійної кореляції не змінюється від зміни початків координат і масштабів ознак (властивість 2), то для спрощення розрахунків замінимо значення ознак X і Y на значення ознак U і V, пов'язаних з ознаками X і Y наступними співвідношеннями :

$$\mathrm{U}= \alpha(\mathrm{X} - x_0), \qquad \mathrm{V}= \beta(\mathrm{Y} - y_0),$$

в яких використаємо $x_0 = 550$; $y_0 = 50$; $\alpha = 0{,}01$; $\beta = 0{,}1$. Отримані значення α і β запишемо над відповідними інтервалами.

У результаті отримаємо кореляційну таблицю.

U	V							
	–3	–2	–1	0	1	2	3	m_v
–3	19	5						24
–2	23	116	11					150
–1	1	41	98	9				149
0		4	32	65	7			108
1		1	4	21	36	3		65
2			1	2	11	13	1	28
3					1	3	2	6
m_u	43	167	146	97	55	19	3	530

Всі подальші обчислення проводяться над значеннями ознак U і V :

$$\overline{\mathrm{U}} = \frac{1}{N}\sum_i m_{x_i} u_i = \frac{-382}{530} = -0,721;$$

$$\overline{\mathrm{V}} = \frac{1}{N}\sum_j m_{y_j} v_j = \frac{-507}{530} = -0,957;$$

$$\overline{\mathrm{UV}} = \frac{1}{N}\sum_i \sum_j m_{ij} u_i v_j = \frac{1161}{530} = 2,191.$$

Далі

$$\overline{\mathrm{V}^2} = \frac{1}{N}\sum_j m_{y_j} v_j^2 = \frac{1359}{530} = 2,564;$$

$$\overline{\mathrm{U}^2} = \frac{1}{N}\sum_i m_{x_i} u_i^2 = \frac{1196}{530} = 2,257,$$

що дає

$$\sigma_{\mathrm{U}}^2 = \overline{\mathrm{U}^2} - \overline{\mathrm{U}}^2 = 2,257 - 0,520 = 1,737;$$
$$\sigma_{\mathrm{V}}^2 = \overline{\mathrm{V}^2} - \overline{\mathrm{V}}^2 = 2,564 - 0,916 = 1,648,$$

звідки отримуємо

$$\sigma_{\mathrm{U}} = 1,318; \qquad \sigma_{\mathrm{V}} = 1,284.$$

Для коефіцієнта кореляції отримаємо

$$r \;=\; \frac{\overline{\mathrm{UV}} - \overline{\mathrm{U}}\cdot\overline{\mathrm{V}}}{\sigma_{\mathrm{U}}\,\sigma_{\mathrm{V}}} \;=\; \frac{2,191 - 0,721\cdot 0.957}{1,318\cdot 1,284} \;=\; \frac{1,501}{1,692} \;=\; 0,887.$$

Далі

$$\overline{\mathrm{X}} \;=\; x_0 + \frac{\overline{\mathrm{U}}}{\alpha} \;=\; 550 + \frac{-0,721}{0,01} \;=\; 550 - 72,1 \;=\; 477,9;$$

$$\overline{\mathrm{Y}} \;=\; y_0 + \frac{\overline{\mathrm{V}}}{\beta} \;=\; 50 + \frac{-0,957}{0,1} \;=\; 50 - 9,57 \;=\; 40,43.$$

Для коефіцієнтів регресії знайдемо

$$\rho_{\mathrm{Y|X}} \;=\; r\,\frac{\sigma_{\mathrm{Y}}}{\sigma_{\mathrm{X}}} \;=\; 0,887\cdot\frac{(1,318/0,1)}{(1,284/0,01)} \;=\; 0,887\cdot\frac{13,18}{128,4} \;=\; 0,086;$$

аналогічно

$$\rho_{\mathrm{X|Y}} \;=\; r\,\frac{\sigma_{\mathrm{X}}}{\sigma_{\mathrm{Y}}} \;=\; 0,887\cdot\frac{(1,284/0,01)}{(1,318/0,1)} \;=\; 0,887\cdot\frac{128,4}{13,18} \;=\; 9,103.$$

Тому рівняння регресій Y на X та X на Y такі:

$$y - 40,43 \;=\; 0,086\,(x - 477,9);$$

$$x - 477,9 \;=\; 9,103\,(y - 40,43),$$

а для середнього квадратичного відхилення вибіркового коефіцієнта лінійної кореляції отримаємо:

$$\sigma_r \;\approx\; \frac{1 - r^2}{\sqrt{N}} \;=\; \frac{1 - 0,887^2}{\sqrt{530}} \;=\; 0,009.$$

Приклад 6.7

Отримана вибірка обсягу $n = 11$, яка здобута з двовимірної нормальної сукупності. Для неї обчислено вибірковий (емпіричний) коефіцієнт кореляції $r = 0{,}76$.

Потрібно: а) побудувати 95%-й довірчий інтервал для коефіцієнта кореляції отриманої вибіркової сукупності; б) перевірити нульову гіпотезу $\{H_0 :\ \rho = 0\}$ проти альтернативної гіпотези $\{H_{\mathrm{a}} :\ \rho \neq 0\}$. Рівень значущості α прийняти 0,01.

Розв'язання

Знайдемо 95%-й довірчий інтервал для коефіцієнта кореляції вибіркової сукупності ρ. Виконавши розрахунки розподілу Крамера, отримаємо для 95%–го довірчого інтервалу $+0,27 < \rho < +0,92$. Оскільки даний довірчий інтервал не накриває значення $\rho = 0$, то на рівні значущості $\alpha = 0{,}01$ відхиляється нульова гіпотеза $\{H_0 :\ \rho = 0\}$ некорельованості змінних X та Y.

Перевірку цієї ж нульової гіпотези можна виконати, використовуючи модель 1. Для цього обчислимо статистику

$$t_{\text{спос}} \;=\; r\sqrt{\frac{n-2}{1-r^2}} \;=\; 0,76\sqrt{\frac{11-2}{1-0,76^2}} \;=\; 3,508.$$

За таблицею стандартизованого нормального розподілу з $\alpha = 0,01$ знаходимо $u_{\alpha/2} = u_{0,005} = 2{,}58$. Тому, використовуючи модель 3, також приходимо до висновку, що вибірковий коефіцієнт кореляції відмінний від нуля.

За таблицею розподілу Стьюдента з $\alpha = 0{,}01$ і кількістю ступенів вільності $\nu = n - 2 = 9$ знаходимо $t_{\alpha/2;\,n-2} = t_{0,005;\,9} = 3{,}25$.

Оскільки $t_{\text{спос}} = 3,50 > 3,25$, то вибірковий коефіцієнт кореляції значуще відрізняється від нуля, тобто змінні X та Y є корельованими.

Для перевірки цієї ж нульової гіпотези H_0, тобто $\rho = 0$, використаємо модель 3. Для цього обчислимо статистику u за формулою

$$u \;=\; \left(1,1513\lg\frac{1+r}{1-r} - 1,1513\lg\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2(n-1)}\right)\sqrt{n-3}.$$

Припускаючи $\rho_0 = 0$, одержимо

$$u \;=\; 1,1513\lg\frac{1,76}{0,24}\sqrt{11-3} \;=\; 2,82.$$

Порівнюючи з $u_{\alpha/2} = u_{0,005} = 2{,}58$, приходимо до висновку про то, що вибірковий коефіцієнт кореляції відрізняється від нуля.

Приклад 6.8

Випуск деяким підприємством промислової продукції (Y) за сім років (X) характеризується наступними даними:

X, *рік*	1	2	3	4	5	6	7
Y, *ум. од.*	0,5	0,5	1,5	3,5	6,5	10,5	15,5

Вирівняти залежність Y від X за параболою $y = ax^2 + bx + c$.

Розв'язання

Для знаходження параметрів a, b та c за методом найменших квадратів необхідно розв'язати систему

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^4 + b\sum\limits_{i=1}^{n} x_i^3 + c\sum\limits_{i=1}^{n} x_i^2 \;=\; \sum\limits_{i=1}^{n} x_i^2\, y_i; \\ a\sum\limits_{i=1}^{n} x_i^3 + b\sum\limits_{i=1}^{n} x_i^2 + c\sum\limits_{i=1}^{n} x_i \;=\; \sum\limits_{i=1}^{n} x_i\, y_i; \\ a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i + c\sum\limits_{i=1}^{n} 1 \;=\; \sum\limits_{i=1}^{n} y_i, \end{cases}$$

в якій в цьому випадку $n = 7$, x_i – це поточний рік семирічки, y_i – відповідна продукція, а $\sum\limits_{i=1}^{n} 1 = n$.

Складаємо систему для визначення параметрів a, b, c:

$$\begin{cases} 4676\,a + 784\,b + 140\,c \;=\; 1372; \\ 784\,a + 140\,b + 28\,c \;=\; 224; \\ 140\,a + 28\,b + 7\,c \;=\; 38,5. \end{cases}$$

Звідси $a = 0{,}5$; $b = -1{,}5$; $c = 1{,}5$.

Приклад 6.9 (Розподіл вибіркового коефіцієнта кореляції)

Розглянемо нормальний розподіл $f_{\mathrm{XY}}(x,y)$ системи двох величин (X,Y). Без обмеження спільності можна передбачати, що моменти першого порядку дорівнюють нулю, так що густина розподілу ймовірності має вигляд

$$f_{\mathrm{XY}}(x,y) \;=\; \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2}-\frac{2\rho xy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right)\right] \;\equiv$$

$$\equiv\; \frac{1}{2\pi\sqrt{\det G}}\exp\left[-\frac{1}{2D}\left(\mu_{02}x^2+2\mu_{11}xy+\mu_{20}y^2\right)\right]. \qquad (1)$$

Запишемо густину нормального розподілу $f_{\mathrm{XY}}(x,y)$ у вигляді

$$f_{\mathrm{XY}}(x,y) \;=\; \frac{1}{2\pi\sqrt{G}}\exp\left[-\frac{1}{2D}\left(\mu_{02}x^2+2\mu_{11}xy+\mu_{20}y^2\right)\right], \qquad (2)$$

де $D=\mu_{20}\mu_{02}-\mu_{11}^2=\sigma_1^2\sigma_2^2(1-\rho^2)$ – визначник матриці других моментів;

$$G \;=\; \begin{pmatrix}\mu_{20} & \mu_{11}\\ \mu_{11} & \mu_{02}\end{pmatrix} \;=\; \begin{pmatrix}\sigma_2^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}. \qquad (3)$$

За вибіркою з n спостережених пар значень $(x_1,y_1),\ldots,(x_n,y_n)$ обчислимо моментні характеристики першого та другого порядку:

$$\overline{x} \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n}y_i,$$

$$m_{20} \;=\; s_1^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2, \qquad (4)$$

$$m_{02} \;=\; s_2^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{y}^2,$$

$$m_{11} \;=\; rs_1s_2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \;=\; \frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\cdot\overline{y}.$$

Коефіцієнт кореляції дорівнює $r=m_{11}/(s_1s_2)$.

Тепер розглянемо характеристичну функцію розподілу п'яти випадкових величин $\overline{x}$, $\overline{y}$, m_{20}, m_{11}, m_{02}, яка є функцією від п'яти аргументів t_1, t_2, t_{20}, t_{11}, t_{02}:

$$Q(t_1,t_2,t_{20},t_{11},t_{02}) \;=\; \mathrm{M}\left[\mathrm{e}^{i(t_1\overline{x}+t_2\overline{y}+t_{20}m_{20}+t_{11}m_{11}+t_{02}m_{02})}\right] \;=$$

$$=\; \frac{1}{(2\pi)^nD^{n/2}}\int\ldots\int\mathrm{e}^{\Omega}\,dx_1\ldots dx_n\,dy_1\ldots dy_n, \qquad (5)$$

де

$$\Omega \;=\; -\frac{1}{2D}\sum_{i=1}^{n}\left(\mu_{02}x_i^2-2\mu_{11}x_iy_i+\mu_{20}y_i^2\right)+$$

$$+i(t_1\overline{x}+t_2\overline{y}+t_{20}m_{20}+t_{11}m_{11}+t_{02}m_{02}),$$

а інтегрування розповсюджується на $2n$-вимірний простір змінних x_1, x_2, …, x_n, y_1, y_2, …, y_n.

Замінимо $x_1, \ldots, x_n$ новими змінними $\xi_1, \ldots, \xi_n$ за допомогою ортогонального перетворення, при якому $\xi_1 = \sqrt{n}\overline{x}$, і застосуємо перетворення з такою ж матрицею до величин $y_1, \ldots, y_n$, які при цьому замінюються новими величинами $\eta_1, \ldots, \eta_n$, причому $\eta_1 = \sqrt{n}\overline{y}$. Тоді маємо

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \xi_i^2, \qquad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \eta_i^2, \qquad \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \xi_i \eta_i,$$

$$nm_{20} = \sum_{i=2}^{n} \eta_i^2, \qquad nm_{02} = \sum_{i=2}^{n} \eta_i^2, \qquad nm_{11} = \sum_{i=2}^{n} \xi_i \eta_i,$$

звідси

$$\Omega = i\frac{t_1\xi_1 + t_2\eta_1}{\sqrt{n}} - \frac{1}{2D}\left(\mu_{02}\xi_1^2 - 2\mu_{11}\xi_1\eta_1 + \mu_{20}\eta_1^2\right) -$$

$$- \frac{1}{n}\sum_{i=2}^{n}\left[\left(\frac{n\mu_{02}}{2M} - it_{20}\right)\xi_i^2 + 2\left(-\frac{n\mu_{11}}{2D} - \frac{1}{2}it_{11}\right)\xi_i\eta_i + \left(\frac{n\mu_{20}}{2D} - it_{02}\right)\eta_i^2\right].$$

Підставляючи цей вираз для Ω у формулу (5), приведемо $2n$-вимірний інтеграл до добутку n подвійних інтегралів. Тоді спільна характеристична функція $Q(t_1, t_2, t_{20}, t_{11}, t_{02})$ набуде вигляду

$$Q(t_1, t_2, t_{20}, t_{11}, t_{02}) = \exp\left[-\frac{1}{2n}\left(\mu_{20}t_1^2 + 2\mu_{11}t_1t_2 + \mu_{02}t_2^2\right)\right](A/A^*)^{(n-1)/2}, \qquad (6)$$

де

$$A = \det\begin{pmatrix} (n\mu_{02})/2D & -(n\mu_{11})/2D \\ -(n\mu_{11})/2D & (n\mu_{20})/2D \end{pmatrix},$$

$$A^* = \det\begin{pmatrix} (n\mu_{02})/2m - it_{20} & -(n\mu_{11})/2D - it_{11}/2 \\ -(n\mu_{11})/2D - it_{11}/2 & (n\mu_{20})/2D - it_{02} \end{pmatrix}.$$

Спільна характеристична функція є добутком двох множників, перший з яких містить лише змінні t_1 й t_2, а другий – лише t_{20}, t_{11} й t_{02}. Перший множник є характеристичною функцією деякого нормального розподілу з нульовим середнім значенням і матрицею других моментів $n^{-1}G$. Другий множник є частковим випадком характеристичної функції

$$q_n(t_{20}, t_{11}, t_{02}) = (A/A^*)^{(n-1)/2}.$$

Відповідний розподіл є частковим випадком розподілу

$$f_n(x_{11}, x_{12}, x_{22}) = C_{2n}\left(a_{11}a_{22} - a_{12}^2\right)^{(n-1)/2} \times \qquad (7)$$

$$\times \left(x_{11}x_{22} - x_{12}^2\right)^{(n-4)/2}\exp\left[-a_{11}x_{11} - a_{22}x_{22} - 2a_{12}x_{12}\right],$$

де

$$C_{2n} = \frac{1}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} = \frac{2^{n-3}}{\pi\Gamma(n-2)}$$

з заміною x_{11}, x_{12}, x_{22} на m_{20}, m_{11}, m_{02}.

Таким чином, складені випадкові величини $(\overline{x},\,\overline{y})$ та $(m_{20},\,m_{11},\,m_{02})$ незалежні. Спільний розподіл величин $\overline{x}$ й $\overline{y}$ нормальний і має ті ж самі моменти першого порядку, що й розподіл генеральної сукупності, та матрицю других моментів $n^{-1}G$.

Спільний розподіл випадкових величин $m_{20},\ m_{11},\ m_{02}$ має густину розподілу ймовірностей $f_n,$ яка задається формулою

$$f_n\left(m_{20},m_{11},m_{02}\right) \;=\; \frac{n^{n-1}}{4\pi\Gamma(n-2)}\,\frac{\left(m_{20}m_{02}-m_{11}^2\right)^{(n-4)/2}}{M^{(n-1)/2}}\;\times \tag{8}$$

$$\times\;\exp\left[-\frac{n}{2M}\left(\mu_{02}m_{20}-2\mu_{11}m_{11}+\mu_{20}m_{02}\right)\right]$$

в області $m_{20}>0,\ m_{02}>0$ та $m_{11}^2<m_{20}m_{02},$ причому $f_n=0$ поза цієї області.

Введемо нову змінну r в спільний розподіл (8) величин $m_{20},\ m_{11}$ й $m_{02},$ беручи $m_{11}=r\sqrt{m_{20}m_{02}},$ так що r – коефіцієнт кореляції вибірки. Тоді отримаємо наступний вираз для спільної густини розподілу ймовірностей величин $m_{20},\ m_{02}$ й r :

$$\sqrt{m_{20}m_{02}}\,f_n(m_{20},r\sqrt{m_{20}m_{02}},m_{02}) \;=\; \tag{9}$$

$$=\;\frac{n^{n-1}}{4\pi\Gamma(n-2)D^{(n-1)/2}}\left(m_{20}m_{02}\right)^{(n-3)/2}\left(1-r^2\right)^{(n-4)/2}\times$$

$$\times\exp\left[-\frac{n}{2D}(\mu_{02}m_{20}-2\mu_{11}r\sqrt{m_{20}m_{02}}+\mu_{20}m_{02})\right],$$

де $m_{20}>0,\ m_{02}>0,\ r^2<1.$

Часткову густину розподілу ймовірності для r отримаємо інтегруванням спільної густини розподілу ймовірності за m_{20} й m_{02} в межах від 0 до ∞. Якщо розкласти в степеневий ряд множник $\exp\left(-n\mu_{11}r\sqrt{m_{20}m_{02}}/D\right),$ то можна здобути густину розподілу ймовірностей для вибіркового коефіцієнта кореляції r :

$$f_n(r) \;=\; \frac{2^{n-3}}{\pi(n-3)!}\left(1-\rho^2\right)^{(n-1)/2}\left(1-r^2\right)^{(n-4)/2}\sum_{\nu=0}^{\infty}\Gamma^2\left(\frac{n+\nu-1}{2}\right)\frac{(2\rho r)^\nu}{\nu!}, \tag{10}$$

при цьому $-1<r<1.$

Присутній в цьому виразі степеневий ряд можна перетворити. Оскільки справедливо

$$\frac{2^{n-3}}{(n-2)!}\sum_{\nu=0}^{\infty}\Gamma^2\left(\frac{n+\nu-1}{2}\right)\frac{(2\rho r)^\nu}{\nu!} \;=\; \int_0^1\frac{x^{n-2}}{(1-\rho r x)^{n-1}}\frac{dx}{\sqrt{1-x^2}}, \tag{11}$$

то з цього маємо наступний вираз для густини розподілу ймовірностей величини r :

$$f_n(r) \;=\; \frac{n-2}{\pi}\left(1-\rho^2\right)^{(n-1)/2}\left(1-r^2\right)^{(n-4)/2}\int_0^1\frac{x^{n-2}}{(1-\rho r x)^{n-1}}\,\frac{dx}{\sqrt{1-x^2}}. \tag{12}$$

Отриманий розподіл залежить тільки від обсягу вибірки n та від коефіцієнта кореляції генеральної сукупності ρ.

При $n=2$ густина розподілу ймовірностей $f_n(r)$ дорівнює нулю. Це відповідає тому факту, що коефіцієнт кореляції, обчислений за вибіркою, яка містить тільки два

значення, необхідно дорівнює ± 1, так що в цьому випадку розподіл відноситься до дискретного типу.

При $n = 3$ густина розподілу ймовірностей U-подібна, з нескінченними ординатами в точках $r = \pm 1$.

При $n = 4$ приходимо до прямокутного розподілу, якщо $\rho = 0$, та J-подібного розподілу у випадку $\rho \neq 0$.

При $n > 4$ розподіл ймовірностей унімодальний, з модою в точці $r = 0$, якщо $\rho = 0$, і біля точки $r = \rho$, якщо $\rho \neq 0$.

6.9. Задачі для розв'язання

Задача 6.1

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–2	–1	0	1	2	3
Y	–2	–3	–3	–1	3	7

Припускаючи, що $y = ax+b$, знайти параметри цієї залежності, користуючись методом найменших квадратів.

Задача 6.2

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–2	–1	0	1	2	3
Y	2	3	3	1	–3	–7

Припускаючи, що $y = ax+b$, знайти параметри цієї залежності, користуючись методом найменших квадратів.

Задача 6.3

Нижче наведена таблиця значень ознаки Y при різних значеннях ознаки X:

X	12	13	14	16	17	18	20	21	21	24	25	26	28
Y	54	59	67	76	85	97	107	118	127	139	153	160	178

Вирівняти залежність Y від X вздовж прямої $y = ax+b$, користуючись методом найменших квадратів. Вирівняти залежність Y від X вздовж прямої $x = cy + d$, користуючись методом найменших квадратів. Здобуті залежності порівняти.

Задача 6.4

З 125 дослідних дільниць 55 знаходилися на неудобреному масиві і 70 на удобреному. При цьому на всіх 125 дільницях мала місце наступна врожайність:

Середній урожай, *г/м*	95	105	115	125	135	145	155	165	175
Кількість дільниць на I масиві	2	5	12	15	10	7	3	1	0
Кількість дільниць на II масиві	0	0	1	2	8	24	19	11	5

Знайти кореляційне відношення врожайності від добрив.

Задача 6.5

Обчислити коефіцієнт лінійної кореляції і отримати рівняння прямої регресії X на Y за даними нижченаведеної таблиці розподілу 100 рослин житняка за загальною вагою всієї рослини Y і за вагою насіння X:

X, *г*	Y, *г*										
	20--30	30--40	40--50	50--60	60--70	70--80	80--90	90--100	100--110	110--120	120--130
10–15	3										
15–20	2	6	1	1							
20–25		4	13	2	1						
25–30			5	4							
30–35				8	4	2					
35–40				1	4	6					
40–45					2	6	1				
45–50		1				1	5	1			
50–55								4	2		
55–60								1	4	1	
60–65									1		
65–70									1	1	1

Задача 6.6

У десяти містах є аптеки. Кількість аптек і кількість населення (в десятках тисяч осіб) в цих 10 містах наведено в таблиці.

Номер міста	1	2	3	4	5	6	7	8	9	10
Населення	45	45	47	48	51	58	59	65	67	80
Кількість аптек	12	12	29	25	38	35	16	43	22	34

Нанести дані на діаграму розсіювання, обчислити коефіцієнти кореляції: а) для перших дев'яти міст; б) для всіх десяти міст. Порівняти результати.

Задача 6.7

Середня температура повітря у вересні в двох містах (X) і (Y) вимірювалася протягом 40 років. Дані наведені в таблиці.

X	Y	X	Y	X	Y	X	Y	X	Y
12,0	10,8	13,9	10,1	14,9	13,0	16,0	16,0	18,0	14,0
12,0	11,3	14,2	10,0	14,9	14,2	16,9	12,9	18,0	14,9
12,0	12,0	14,0	10,0	15,1	13,8	17,2	13,9	18,1	16,0
12,0	13,0	14,0	12,0	15,0	16,0	16,9	14,8	18,4	17,8
12,8	10,9	13,9	12,4	15,5	13,9	16,9	15,0	19,2	15,0
13,8	10,0	15,0	11,0	15,9	14,7	17,0	16,0	19,3	16,1
13,1	13,0	14,0	14,8	16,0	13,0	16,8	17,0	20,0	17,0
13,0	13,0	14,0	15,2	15,9	15,0	17,5	16,0	20,1	17,7

Знайти вибіркові середньомісячні температури в обох населених пунктах і їх середньоквадратичні відхилення. Знайти вибірковий коефіцієнт кореляції X і Y, написати вибіркове рівняння лінійної регресії Y на X.

Задача 6.8

Залежність національного прибутку (Y) від року (X) наведена в таблиці.

X	1932	1933	1934	1935	1936	1937	1938
Y, *млрд. крб.*	45,5	48,5	55,8	65,7	86,0	96,3	105,0

Передбачаючи, що регресійна залежність має вигляд $y = a\,b^{x-1932}$, знайти значення параметрів a і b, якщо відомий національний прибуток за вказані роки.

Задача 6.9

Вісім разів при різних значеннях ознаки X було виміряно значення ознаки Y. Отримані результати наводяться в наступній таблиці:

X	0,30	0,91	1,50	2,00	2,20	2,62	3,00	3,30
Y	0,20	0,43	0,31	0,52	0,81	0,68	1,15	0,85

Передбачаючи, що регресійна залежність має вигляд $y = ax+b$, знайти значення параметрів a і b.

Задача 6.10

Дослідження залежності тривалості t розв'язання систем лінійних рівнянь однакової складності від порядку системи n подано у вигляді таблиці:

n	2	3	4	5	6	7	8	9	10
t, *хвилин*	12	35	75	130	210	315	445	600	800

Передбачаючи, що $t = An^{\alpha}$, знайти значення параметрів A та α методом найменших квадратів.

Задача 6.11

У дослідах були отримані наступні результати при вимірюваннях діаметра (Y) пилинки фуксії залежно від кількості спор (X), розташованих в екваторіальній площині пилинки:

X, *мк*	Y					
	0	1	2	3	4	m_y
10	3					3
15	7	3				10
20		6				6
25		6	1			7
30			4			4
35			5			5
40			1	3		4
45				4		4
50				3	3	6
55					4	4
60					3	3

Знайти коефіцієнт лінійної кореляції між ознаками X та Y і рівняння лінійної регресії y на x.

Задача 6.12

Є наступні дані про залежність двох ознак X і Y від ознаки T:

T	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961
X	37	47	49	51	61	75	80	92	102	117	120	122
Y	53	42	30	24	22	22	26	31	35	38	38	36

Вирівняти залежність X та Y від ознаки T. Побудувати відповідні прямі.

Задача 6.13

Наступна кореляційна таблиця дає розподіл двох ознак X і Y:

X	Y					
	10	20	30	40	50	m_x
5	2				2	4
6	1	1		1	1	4
7		2	1	2		5
8			1			1
m_y	3	3	2	3	3	14

Знайти вибірковий коефіцієнт кореляції X і Y. Побудувати залежність лінійної регресії X на Y й Y на X.

Задача 6.14

Розподіл ознак X й Y наводиться в наступній таблиці:

X	Y											
	0	1	2	3	4	5	6	7	8	9	10	m_y
0						2						2
1					3	2	3					8
2			2	6	4	3	4	6	2			27
3			5	4	6	5	6	4	5			35
4	1	3	4	6	7	8	7	6	4	3	1	50
5			5	4	6	5	6	4	5			35
6			2	6	4	3	4	6	2			27
7					3	2	3					8
8						2						2
m_x	1	3	18	26	33	32	33	26	18	3	1	194

Знайти коефіцієнт лінійної кореляції між ознаками X та Y і написати рівняння прямих регресій X на Y.

Задача 6.15

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–2	–1	0	1	2	3
Y	–2	–3	–3	–1	3	7

Передбачаючи, що $y = ax^2 + bx + c$, знайти параметри цієї залежності, користуючись методом найменших квадратів.

6.10. Завдання на практичну роботу

Практична робота розрахована на дві години і містить два завдання. Завдання повинно виконуватись у обраному програмному середовищі.

Завдання 1

Побудуйте програму знаходження значень параметрів a та b лінійної залежності $y = ax + b$, а також значень параметрів c та d лінійної залежності $x = cy + d$. Скористуйтесь методом найменших квадратів.

Варіант 1

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	–5	–3	–3	–1	–1	3	3

Результат роботы – масив, що містить значення шуканих параметрів. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також лінії прямих регресій $y = ax + b$ та $x = cy + d$).

Варіант 2

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	–4	–2	–3	–1	–1	3	2

Результат роботы – масив, що містить значення шуканих параметрів. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також лінії прямих регресій $y = ax + b$ та $x = cy + d$).

Варіант 3

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	–4	–2	–2	–1	–1	2	3

Результат роботи – масив, що містить значення шуканих параметрів. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також лінії прямих регресій $y = ax + b$ та $x = cy + d$).

Завдання 2

Побудуйте програму знаходження значень параметрів a, b, c параболічної залежності $y = ax^2 + bx + c$. Скористуйтесь методом найменших квадратів.

Варіант 1

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	7	5	3	1	–1	3	7

Результат роботи – масив, що містить значення шуканих параметрів a, b, c параболічної апроксимації. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також параболу $y = ax^2 + bx + c$).

Варіант 2

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	7	5	3	0	3	5	7

Результат роботи – масив, що містить значення шуканих параметрів a, b, c параболічної апроксимації. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також параболу $y = ax^2 + bx + c$).

Варіант 3

Залежність ознаки Y від ознаки X характеризується наступною таблицею:

X	–3	–2	–1	0	1	2	3
Y	–6	–5	–3	–1	–1	–3	–7

Результат роботи – масив, що містить значення шуканих параметрів a, b, c параболічної апроксимації. Необхідно передбачити візуалізацію даних (побудувати програмно точки заданих значень, а також параболу $y = ax^2 + bx + c$).

6.11. Завдання для перевірки

1. У чому полягає відмінність між функціональною і статистичною залежністю між двома змінними?

2. Що називається модельним рівнянням регресії Y на X?

3. Що називається емпіричним рівнянням регресії Y на X?

4. Сформулюйте загальну ідею підбору емпіричних рівнянь методом найменших квадратів.

5. Сформулюйте задачі кореляційного аналізу, регресійного аналізу.

6. Запишіть загальний вигляд модельних функцій регресії Y на X й X на Y, якщо відомо, що двовимірна випадкова величина (X, Y) розподілена згідно з нормальним законом.

7. У чому полягає властивість мінімальності модельних функцій регресії?

8. Виведіть формули для обчислення коефіцієнтів емпіричного рівняння регресії за згрупованими і незгрупованими дослідними даними.

9. Що називається коваріацією (коваріаційним моментом) генеральної сукупності?

10. Що називається коваріацією (коваріаційним моментом) вибіркової сукупності?

12. Сформулюйте основні початкові припущення, що лежать в обгрунтуванні лінійної моделі регресії.

13. Який вигляд має графік емпіричної функції регресії, якщо відомо, що випадкова вибірка витягнута з генеральної нормальної сукупності?

14. Запишіть систему рівнянь для визначення коефіцієнтів b_0 і b_1 лінійних рівнянь регресії Y на X вигляду $\overline{y}_x = b_0 + b_1 x$ за методом найменших квадратів.

15. Запишіть систему рівнянь для визначення коефіцієнтів a_0 і a_1 лінійних рівнянь регресії X на Y вигляду $\overline{x}_y = a_0 + a_1 y$ за методом найменших квадратів.

16. Порівняйте два рівняння лінійної регресії : Y на X та X на Y. Поясніть зв'язок між коефіцієнтами a_0, a_1 та b_0, b_1.

Додаток

Д.1. Довідкові таблиці

У цьому додатку до посібника наведено таблиці величин, що часто використовуються при розв'язанні задач математичної статистики:

1. Функція Гаусса $\exp(-x^2)$, функція помилок $\mathrm{erf}(x)$ і додаткова функція помилок $\mathrm{erfc}(x)$;

2. Функція Лапласа $\Phi(x)$ і функції $d\Phi(x)/dx$, $2\Phi(x)$, $1-2\Phi(x)$;

3. Квантилі u_α нормального розподілу;

4. Нижні γ_1 та верхні γ_2 межі довірчих інтервалів СКВ нормальної випадкової величини, що відповідають рівності: $p = \Pr(\gamma_1 S < \sigma < \gamma_2 S) = 1-\alpha$;

5. Квантилі $\chi^2_{\alpha,\nu}$ розподілу χ^2 з ν ступенями вільності, що відповідають рівності $\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) = \alpha$;

6. Правосторонні квантилі $\chi^2_{1-\alpha,\nu}$ розподілу χ^2 з ν ступенями вільності, що відповідають рівності $\Pr\left(\chi^2 > \chi^2_{1-\alpha,\nu}\right) = 1-\alpha$;

7. Квантилі $t_{\alpha,\nu}$ розподілу Стьюдента з ν ступенями вільності, що визначаються ймовірністю $\Pr(|t| > t_{\alpha,\nu}) = \alpha$ (двостороння критична область);

8. Лівосторонні квантилі розподілу Стьюдента $t_p(k)$;

9. Критичні точки розподілу $F_{\alpha;k_1,k_2}$ Фішера–Снедекора з кількістю ступенів вільності k_1 та k_2;

10. Граничний розподіл Колмогорова $K(\lambda)$;

11. Критичні значення λ_α розподілу Колмогорова: $\Pr(\lambda \geq \lambda_\alpha) = \alpha$;

12. Критичні значення λ_α розподілу Смірнова–Колмогорова: $\Pr(\lambda \geq \lambda_\alpha) = \alpha$;

13. Критичні значення r_α кількості знаків: $\Pr(r \geq r_\alpha) = \alpha$.

x	$\exp(-x^2)$	$(2\pi)^{-1/2}\exp(-x^2/2)$	$\operatorname{erf}(x)$	$\operatorname{erfc}(x)$
0,00	1,00000	0,39894	0,00000	1,00000
0,10	0,99005	0,39695	0,11246	0,88754
0,20	0,96079	0,39104	0,22270	0,77730
0,30	0,91393	0,38139	0,32863	0,67137
0,40	0,85214	0,36827	0,42839	0,57161
0,50	0,77880	0,35207	0,52050	0,47950
0,60	0,69768	0,33322	0,60386	0,39614
0,70	0,61263	0,31225	0,67780	0,32220
0,80	0,52729	0,28969	0,74210	0,25790
0,90	0,44486	0,26609	0,79691	0,20309
1,00	0,36788	0,24197	0,84270	0,15730
1,10	0,29820	0,21785	0,88021	0,11979
1,20	0,23693	0,19419	0,91031	0,08969
1,30	0,18452	0,17137	0,93401	0,06599
1,40	0,14086	0,14973	0,95229	0,04771
1,50	0,10540	0,12952	0,96611	0,03389
1,60	0,07730	0,11092	0,97635	0,02365
1,70	0,05558	0,09405	0,98379	0,01621
1,80	0,03916	0,07895	0,98909	0,01091
1,90	0,02705	0,06562	0,99279	0,00721
2,00	0,01832	0,05399	0,99532	0,00468
2,10	0,01216	0,04398	0,99702	0,00298
2,20	0,00791	0,03547	0,99814	0,00186
2,30	0,00504	0,02833	0,99886	0,00114
2,40	0,00315	0,02239	0,99931	0,00069
2,50	0,00193	0,01753	0,99959	0,00041
2,60	0,00116	0,01358	0,99976	0,00024
2,70	0,00068	0,01042	0,99987	0,00013
2,80	0,00039	0,00792	0,99992	0,00008
2,90	0,00022	0,00595	0,99996	0,00004
3,00	0,00012	0,00443	0,99998	0,00002

Тут

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_0^x \exp(-t^2)\,dt, \qquad \operatorname{erfc}(x) = 1 - \operatorname{erf}(x).$$

x	$d\Phi(x)/dx$	$\Phi(x)$	$2\,\Phi(x)$	$1-\Phi(x)$
0,00	0,39894	0,00000	0,00000	1,00000
0,10	0,39695	0,03983	0,07966	0,96017
0,20	0,39104	0,07926	0,15852	0,92074
0,30	0,38139	0,11791	0,23582	0,88209
0,40	0,36827	0,15542	0,31084	0,84458
0,50	0,35207	0,19146	0,38292	0,80854
0,60	0,33322	0,22575	0,45149	0,77425
0,70	0,31225	0,25804	0,51607	0,74196
0,80	0,28969	0,28814	0,57629	0,71186
0,90	0,26609	0,31594	0,63188	0,68406
1,00	0,24197	0,34134	0,68269	0,65866
1,10	0,21785	0,36433	0,72867	0,63567
1,20	0,19419	0,38493	0,76986	0,61507
1,30	0,17137	0,40320	0,80640	0,59680
1,40	0,14973	0,41924	0,83849	0,58076
1,50	0,12952	0,43319	0,86639	0,56681
1,60	0,11092	0,44520	0,89040	0,55480
1,70	0,09405	0,45543	0,91087	0,54457
1,80	0,07895	0,46407	0,92814	0,53593
1,90	0,06562	0,47128	0,94257	0,52872
2,00	0,05399	0,47725	0,95450	0,52275
2,10	0,04398	0,48214	0,96427	0,51786
2,20	0,03547	0,48610	0,97219	0,51390
2,30	0,02833	0,48928	0,97855	0,51072
2,40	0,02239	0,49180	0,98360	0,50820
2,50	0,01753	0,49379	0,98758	0,50621
2,60	0,01358	0,49534	0,99068	0,50466
2,70	0,01042	0,49653	0,99307	0,50347
2,80	0,00792	0,49744	0,99489	0,50256
2,90	0,00595	0,49813	0,99627	0,50187
3,00	0,00443	0,49865	0,99730	0,50135

Тут

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_0^x \exp(-t^2/2)\,dt.$$

$$\Phi(x) = \frac{1}{2}\,\mathrm{erf}(x/\sqrt{2}), \qquad\qquad \mathrm{erf}(x) = 2\,\Phi(\sqrt{2}\,x).$$

Квантиль u_α визначається рівністю

$$\alpha \;=\; 1-\Phi(u_\alpha) \;=\; 1-\frac{1}{\sqrt{2\pi}}\int\limits_{u_\alpha}^{\infty}\exp\left(-t^2/2\right)\,dt \;=\; 1-\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{u_\alpha}\exp\left(-t^2/2\right)\,dt.$$

Рівень значущості α	Квантиль u_α
0,0001	3,0902
0,005	2,5758
0,010	2,3263
0,015	2,1701
0,020	2,0537
0,025	1,9600
0,030	1,8808
0,035	1,8119
0,040	1,7507
0,045	1,6954
0,050	1,6449

При чисельному знаходженні значень $\Phi(x)$ і u_α можна також користуватися наступними наближеними виразами:

$$q_x \;=\; 2[1-\Phi(x)] \;=\; \frac{2}{\sqrt{2\pi}}\int\limits_{x}^{\infty}\exp\left(-u^2/2\right)\,du \;=$$

$$=\; \exp\left(-\frac{(83x+35)x+562}{165+703/x}\right), \qquad 0<x\le 5,5;$$

$$x \;=\; \left[\frac{((4y+100)y+205)y}{(2y+50)(y+192)y+13}\right]^{-1/2}, \qquad 2\cdot 10^{-7}\le q_x\le 1,$$

де $y=-\ln q_x$.

4. Нижні γ_1 і верхні γ_2 межі довірчого інтервалу, що відповідають рівності $p = \Pr(\gamma_1 S < \sigma < \gamma_2 S) = 1 - \alpha$

Кількість ступенів вільності ν	Рівень значущості					
	$p = 1 - \alpha = 0,99$		$p = 1 - \alpha = 0,95$		$p = 1 - \alpha = 0,90$	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
1	0,356	159	0,446	31,9	0,510	15,9
2	0,434	14,1	0,521	6,28	0,578	4,40
3	0,483	6,47	0,566	3,73	0,620	2,92
4	0,519	4,39	0,599	2,87	0,649	2,37
5	0,546	3,48	0,624	2,45	0,672	2,090
6	0,569	2,98	0,644	2,202	0,690	1,916
7	0,588	2,66	0,661	2,035	0,705	1,797
8	0,604	2,440	0,675	1,916	0,718	1,711
9	0,618	2,277	0,688	1,826	0,729	1,645
10	0,630	2,154	0,699	1,755	0,739	1,593
11	0,641	2,056	0,708	1,698	0,748	1,550
12	0,651	1,976	0,717	1,651	0,755	1,515
13	0,660	1,910	0,725	1,611	0,762	1,485
14	0,669	1,854	0,732	1,577	0,769	1,460
15	0,676	1,806	0,739	1,548	0,775	1,437
16	0,683	1,764	0,745	1,522	0,780	1,418
17	0,690	1,727	0,750	1,499	0,785	1,400
18	0,696	1,695	0,756	1,479	0,790	1,385
19	0,702	1,666	0,760	1,460	0,794	1,370
20	0,707	1,640	0,765	1,444	0,798	1,358
21	0,712	1,617	0,769	1,429	0,802	1,346
22	0,717	1,595	0,773	1,416	0,805	1,335
23	0,722	1,576	0,777	1,402	0,809	1,326
24	0,726	1,558	0,781	1,391	0,812	1,316
25	0,730	1,541	0,784	1,380	0,815	1,308
26	0,734	1,526	0,788	1,371	0,818	1,300
27	0,737	1,512	0,791	1,361	0,820	1,293
28	0,741	1,499	0,794	1,352	0,823	1,286
29	0,744	1,487	0,796	1,344	0,825	1,279
30	0,748	1,475	0,799	1,337	0,828	1,274

5. Критичні точки $\chi^2_{\alpha,\nu}$ розподілу χ^2, що відповідають рівності $\Pr\left(\chi^2 > \chi^2_{\alpha,\nu}\right) = \alpha$

Кількість ступенів вільності ν	Рівень значущості α (двостороння критична область)					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,001	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

6. Правосторонні квантилі $\chi^2_{1-\alpha,\nu}$ розподілу χ^2 з ν ступенями вільності, що відповідають рівності $\Pr\left(\chi^2 > \chi^2_{1-\alpha,\nu}\right) = 1-\alpha$

ν	$1-\alpha$							
	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999
1	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,8
2	2,41	3,22	4,61	5,99	7,38	9,21	10,6	13,8
3	3,67	4,64	6,25	7,81	9,35	11,3	12,8	16,3
4	4,88	5,99	7,78	9,49	11,1	13,3	14,9	18,5
5	6,06	7,29	9,24	11,1	12,8	15,1	16,7	20,5
6	7,23	8,56	10,6	12,6	14,4	16,8	18,5	22,5
7	8,38	9,80	12,0	14,1	16,0	18,5	20,3	24,3
8	9,52	11,0	13,4	15,5	17,5	20,1	22,0	26,1
9	10,7	12,2	14,7	16,9	19,0	21,7	23,6	27,9
10	11,8	13,4	16,0	18,3	20,5	23,2	25,2	29,6
11	12,9	14,6	17,3	19,7	21,9	24,7	26,8	31,3
12	14,0	15,8	18,5	21,0	23,3	26,2	28,3	32,9
13	15,1	17,0	19,8	22,4	24,7	27,7	29,8	34,5
14	16,2	18,2	21,1	23,7	26,1	29,1	31,3	36,1
15	17,3	19,3	22,3	25,0	27,5	30,6	32,8	37,7
16	18,4	20,5	23,5	26,3	28,8	32,0	34,3	39,3
17	19,5	21,6	24,8	27,6	30,2	33,4	35,7	40,8
18	20,6	22,8	26,0	28,9	31,5	34,8	37,2	42,3
19	21,7	23,9	27,2	30,1	32,9	36,2	38,6	43,8
20	22,8	25,0	28,4	31,4	34,2	37,6	40,0	45,3
21	23,9	26,9	29,6	32,7	35,5	38,9	41,4	46,8
22	24,9	27,3	30,8	33,9	36,8	40,3	42,8	48,3
23	26,0	28,4	32,0	35,2	38,1	41,6	44,2	49,7
24	27,1	29,6	33,2	36,4	39,4	43,0	45,6	51,2
25	28,2	30,7	34,4	37,7	40,6	44,3	46,9	52,6
26	29,2	31,8	35,6	38,9	41,9	45,6	48,3	54,1
27	30,3	32,9	36,7	40,1	43,2	47,0	49,6	55,5
28	31,4	34,0	37,9	41,3	44,5	48,3	51,0	56,9
29	32,5	35,1	39,1	42,6	45,7	49,6	52,3	58,3
30	33,5	36,3	40,3	43,8	47,0	50,9	53,7	59,7

7. Квантилі $t_{\alpha,\nu}$ розподілу Стьюдента з ν ступенями вільності, що визначаються ймовірністю $\Pr(|t| > t_{\alpha,\nu}) = \alpha$ (двостороння критична область)

Кількість ступенів вільності ν	Рівень значущості α					
	0,050	0,025	0,020	0,010	0,005	0,002
1	12,706	25,452	31,821	63,657	127,3	318,3
2	4,303	6,205	6,965	9,925	14,09	22,33
3	3,182	4,177	4,541	5,841	7,453	10,21
4	2,776	3,495	3,747	4,604	5,598	7,173
5	2,571	3,163	3,365	4,032	4,773	5,893
6	2,447	2,969	3,143	3,707	4,317	5,208
7	2,365	2,841	2,998	3,499	4,029	4,785
8	2,306	2,752	2,896	3,355	3,833	4,501
9	2,262	2,685	2,821	3,250	3,690	4,297
10	2,228	2,634	2,764	3,169	3,581	4,144
11	2,201	2,593	2,718	3,106	3,497	4,025
12	2,179	2,560	2,681	3,055	3,428	3,930
13	2,160	2,533	2,650	3,012	3,372	3,852
14	2,145	2,510	2,624	2,977	3,326	3,787
15	2,131	2,490	2,602	2,947	3,286	3,733
16	2,120	2,473	2,583	2,921	3,252	3,686
17	2,110	2,458	2,567	2,898	3,222	3,646
18	2,101	2,445	2,552	2,878	3,197	3,610
19	2,093	2,433	2,539	2,861	3,174	3,579
20	2,086	2,423	2,528	2,845	3,153	3,552
21	2,080	2,414	2,518	2,831	3,135	3,527
22	2,074	2,405	2,508	2,819	3,119	3,505
23	2,069	2,398	2,500	2,807	3,104	3,485
24	2,064	2,391	2,492	2,797	3,091	3,467
25	2,060	2,385	2,485	2,787	3,078	3,450
26	2,056	2,379	2,479	2,779	3,067	3,435
27	2,052	2,373	2,473	2,771	3,057	3,421
28	2,048	2,368	2,467	2,763	3,047	3,408
29	2,045	2,364	2,462	2,756	3,038	3,396
30	2,042	2,360	2,457	2,750	3,030	3,385
40	2,021	2,329	2,423	2,705	2,971	3,307
60	2,000	2,300	2,390	2,660	2,915	3,232
120	1,980	2,980	2,358	2,617	2,860	3,107
∞	1,960	2,241	2,326	2,576	2,807	3,090

8. Лівосторонні квантилі розподілу Стьюдента $t_p(k)$ з k ступенями вільності, які відповідають рівності $\Pr(t < t_p(k)) = p$

	p						
k	0,750	0,900	0,950	0,975	0,990	0,995	0,999
1	1,000	3,078	6,314	12,706	31,821	63,657	318,3
2	0,817	1,886	2,920	4,303	6,965	9,925	22,33
3	0,765	1,638	2,353	3,182	4,541	5,841	10,21
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025
12	0,696	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485
24	0,686	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435
27	0,686	1,314	1,703	2,052	2,473	2,771	3,421
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408
29	0,683	1,311	1,699	2,045	2,462	2,756	3,390
30	0,686	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160
∞	0,675	1,282	1,645	1,960	2,326	2,576	3,070

9. Квантилі розподілу $F_{\alpha;\,k_1,k_2}$ Фішера-Снедекора

k_1 – кількість ступенів вільності більшої дисперсії;
k_2 – кількість ступенів вільності меншої дисперсії

Рівень значущості $\alpha = 0{,}05$						
	k_1					
k_2	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18,51	19,00	19,16	19,25	19,30	19,33
3	10,13	9,55	9,28	9,12	9,01	8,94
4	7,71	6,94	6,59	6,39	6,26	6,16
5	6,61	5,79	5,41	5,19	5,05	4,95
6	5,99	5,14	4,76	4,53	4,39	4,28
7	5,59	4,74	4,35	4,12	3,97	3,87
8	5,32	4,46	4,07	3,84	3,69	3,58
9	5,12	4,26	3,86	3,63	3,48	3,37
10	4,96	4,10	3,71	3,48	3,33	3,22
11	4,84	3,98	3,59	3,36	3,20	3,09
12	4,75	3,88	3,49	3,26	3,11	3,00
13	4,67	3,80	3,41	3,18	3,02	2,92
14	4,60	3,74	3,34	3,11	2,96	2,85
15	4,54	3,68	3,29	3,06	2,90	2,79
16	4,49	3,63	3,24	3,01	2,85	2,74

Рівень значущості $\alpha = 0{,}05$						
	k_1					
k_2	7	8	9	10	11	12
1	237	239	241	242	243	244
2	19,36	19,37	19,38	19,39	19,40	19,41
3	8,88	8,84	8,81	8,78	8,76	8,74
4	6,09	6,04	6,00	5,96	5,93	5,91
5	4,88	4,82	4,78	4,74	4,70	4,68
6	4,21	4,15	4,10	4,06	4,03	4,00
7	3,79	3,73	3,68	3,63	3,60	3,57
8	3,50	3,44	3,39	3,34	3,31	3,28
9	3,29	3,23	3,18	3,13	3,10	3,07
10	3,14	3,07	3,02	2,97	2,94	2,91
11	3,01	2,95	2,90	2,86	2,82	2,79
12	2,92	2,85	2,80	2,76	2,72	2,69
13	2,84	2,77	2,72	2,67	2,63	2,60
14	2,77	2,70	2,65	2,60	2,56	2,53
15	2,70	2,64	2,59	2,55	2,51	2,48
16	2,66	2,59	2,54	2,49	2,45	2,42

10. Граничний розподіл Колмогорова

$$K(\lambda) \;=\; \sum_{\nu=-\infty}^{\infty} (-1)^{\nu} \exp\left(-2\nu^2\lambda^2\right)$$

λ	$K(\lambda)$		λ	$K(\lambda)$		λ	$K(\lambda)$
0,30	0,0000		0,90	0,6073		1,50	0,9778
0,35	0,0003		0,95	0,6725		1,55	0,9836
0,40	0,0028		1,00	0,7300		1,60	0,9880
0,45	0,0126		1,05	0,7798		1,65	0,9914
0,50	0,0361		1,10	0,8223		1,70	0,9938
0,55	0,0772		1,15	0,8580		1,75	0,9956
0,60	0,1357		1,20	0,8878		1,80	0,9969
0,65	0,2080		1,25	0,9121		1,85	0,9979
0,70	0,2888		1,30	0,9319		1,90	0,9985
0,75	0,3728		1,35	0,9478		1,95	0,9990
0,80	0,4559		1,40	0,9603		2,00	0,99933
0,85	0,5347		1,45	0,9702		2,05	0,99999

11. Критичні значення λ_α розподілу Колмогорова

$$\Pr\left(\lambda \geq \lambda_\alpha\right) \;=\; \alpha$$

Рівень значущості α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	1,070	1,224	1,358	1,520	1,627	1,950

12. Критичні значення λ_α розподілу Смірнова-Колмогорова

$$\Pr\left(\lambda \geq \lambda_\alpha\right) \;=\; \alpha$$

Рівень значущості α	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	0,892	1,075	1,223	1,395	1,518	1,860

13. Критичні значення r_α кількості знаків

$$\Pr\left(r \geq r_\alpha\right) \;=\; \alpha$$

n	Рівень значущості α				n	Рівень значущості α		
	0,01	0,05	0,10			0,01	0,05	0,10
6		0	0		21	4	5	6
7		0	0		22	4	5	6
8	0	0	1		23	4	6	7
9	0	1	1		23	4	6	7
10	0	1	1		24	5	6	7
11	0	1	2		25	5	7	7
12	1	2	2		26	6	7	8
13	1	2	2		27	6	7	8
14	1	2	3		28	6	8	9
15	2	3	3		29	7	8	9
16	2	3	4		30	7	9	10
17	2	4	4		31	7	9	10
18	3	4	4		32	8	9	10
19	3	4	5		33	8	10	11
20	3	5	5		34	8	10	11

Список літератури

Підручники

1. Вентцель Е.С. *Теория вероятностей.* — М.: Наука, 1964.

2. Гнеденко Б.В. *Курс теории вероятностей.* — М.: Наука, 1961.

3. Коваленко И.М., Филиппова А.А. *Теория вероятностей и математическая статистика.* — М.: Высшая школа, 1973.

4. Пугачев В.С. *Теория вероятностей и математическая статистика.* — М.: Наука, 1979.

5. Гихман И.И., Скороход А.В., Ядренко М.И. *Теория вероятностей и математическая статистика.* — Киев: Вищ. шк., 1979.

6. Ивченко Г.И., Медведев Ю.И. *Математическая статистика: Учебное пособие для втузов.* — М.: Высш. шк., 1984.

7. Четыркин Е.М., Калихман И.Л. *Вероятность и статистика.* — М.: Финансы и статистика, 1982.

8. Гмурман В.Е. *Теория вероятностей и математическая статистика.* — М.: Высш. шк., 2000.

Задачники і посібники

9. Емельянов Г.В., Скитович В.П. *Задачник по теории вероятностей и математической статистике.* — Л.: Изд–во ЛГУ, 1967.

10. Герасимович А.И., Матвеева Я.И. *Математическая статистика.* — Минск: Вышэйшая шк., 1978.

11. Гмурман В.Е. *Руководство к решению задач по теории вероятностей и математической статистике.* — М.: Высш. шк., 2000.

12. Зубков А.М., Севастьянов Б.А., Чистяков В.П. *Сборник задач по теории вероятностей.* — М.: Наука, 1989.

13. Сборник задач по математике. Теория вероятностей и математическая статистика / Под ред. А.В. Ефимова. — М.: Наука, 1990.

14. Харин Ю.С., Степанова М.Д. *Практикум на ЭВМ по математической статистике.* — Минск: "Университетское", 1987.

15. Турчин В.М. *Математична статистика в прикладах і задачах: Навчальний посібник.* — К.: НМК ВО, 1993.

16. Мазманишвили А.С. *Теория вероятностей: Учебное пособие.* — Харьков: ХГПУ, 1994.

17. Боровиков В.П., Боровиков. *STATISTIKA. — Статистический анализ и обработка данных в среде Windows.* — М.: "Филин", 1997.

18. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. *Справочник по теории вероятностей и математической статистике.* — М.: Наука, 1985.

19. Корн Г., Корн Т. *Справочник по математике.* — М.: Наука, 1983.

20. Абрамович М., Стиган И. *Справочник по специальным функциям.* — М.: Наука, 1979.

21. Большев Л.Н., Смирнов Н.В. *Таблицы математической статистики.* — М.: Наука, 1983.

Додаткова література

22. Чебышев П.Л. *Теория вероятностей.* — М.–Л.: Изд–во АН СССР, 1936.

23. Дюге Д. *Теоретическая и прикладная статистика.* — М.: Наука, 1972.

24. Крамер Г. *Математические методы статистики.* — М.: Мир, 1975.

25. Боровков А.А. *Математическая статистика. Оценка параметров. Проверка гипотез.* — М.: Наука, 1984.

26. Феллер В. *Введение в теорию вероятностей и ее применение.* — М.: Мир, 1984. – Т.1; 1984. – Т.2.

27. Митропольский А.К. *Техника статистических вычислений.* — М.: Наука, 1971.

28. Смирнов Н.В., Дунин–Барковский И.В. *Курс теории вероятностей и математической статистики для технических приложений.* — М.: Наука, 1969.

29. Яноши Л. *Теория и практика обработки результатов измерений.* — М.: Мир, 1968.

30. Анго А. *Математика для электро– и радиоинженеров.* — М.: Наука, 1965.

31. Воинов В.Г., Никулин М.С. *Несмещенные оценки и их применения.* — М.: Наука, 1989.

32. Розанов Ю.А. *Теория вероятностей, случайные процессы и математическая статистика.* — М.: Наука, 1985.

33. Мартынов Г.В. *Критерий омега–квадрат.* — М.: Наука, 1978.

34. Шеффе Г. *Дисперсионный анализ.* — М.: Наука, 1980.

35. Кокс Д., Хинкли Д. *Теоретическая статистика.* — М.: Мир, 1984.

36. Бикел П., Доксам К. *Математическая статистика.* — М.: Финансы и статистика, 1983.

37. Ван дер Варден Б.Л. *Математическая статистика.* — М.: Изд–во иностр. лит., 1960.

38. Себер Дж. *Линейный регрессионный анализ.* — М.: Мир, 1982.

39. Хьюбер П. *Робастность в статистике.* — М.: Мир, 1984.

40. Налимов В.В. *Теория эксперимента.* — М.: Наука, 1971.

41. Секей Г. *Парадоксы в теории вероятностей и математической статистике.* — М.: Мир, 1990.

42. Уилкс С.С. *Математическая статистика.* — М.: Наука, 1967.

Навчальне видання

МАЗМАНІШВІЛІ Олександр Сергійович

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник до практичних занять

Українською мовою

Роботу до видання рекомендував О.В. Горілий

Редактор О.С. Самініна

План 2010, поз. 63

Підп. до друку 20.10.2009 Формат 60×94 1/16 Папір офсетн.
Друк – ризографія. Гарнітура Times New Roman. Ум. друк. арк. 12,5.
Обл.–вид. арк. 15,5. Наклад 200 прим. Зам. № Ціна договірна

Видавничий центр НТУ "ХПІ", 61002, Харків, вул. Фрунзе, 21
Свідоцтво про реєстрацію ДК № 116 від 10.07.2000

Друкарня НТУ "ХПІ", 61002, Харків, вул. Фрунзе, 21

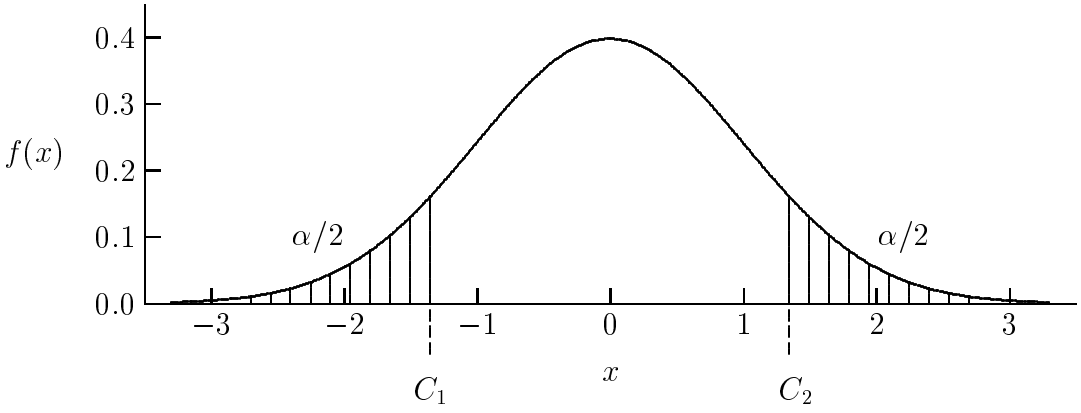
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

О. С. МАЗМАНІШВІЛІ

МАТЕМАТИЧНА СТАТИСТИКА

НАВЧАЛЬНИЙ ПОСІБНИК ДО ПРАКТИЧНИХ ЗАНЯТЬ



Харків НТУ "ХПІ" 2010