МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Державний вищий навчальний заклад
«КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
імені ВАДИМА ГЕТЬМАНА»

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

*Рекомендовано
Міністерством освіти і науки,
молоді та спорту України*

1906 КНЕУ
КИЇВ 2015

УДК 519.22/.25:33(075.8)
ББК 22.172
М 34

Колектив авторів
С. М. ГРИГУЛИЧ, В. П. ЛІСОВСЬКА, О. І. МАКАРЕНКО, Т. В. МАНЖОС, І. І. ПАХОМОВ, В. Д. СТАСЮК, Г. М. ЧЕРНІС

Рецензенти
З. П. Бараннік, д.е.н., проф.
(Київський національний економічний університет ім. Вадима Гетьмана)
Л. Г. Лобас, канд.фіз.-мат.наук, доц.
(Державний економіко-технологічний університет транспорту)
А. С. Багатирчук, канд. фіз.-мат.наук, доц.
(Національний університет харчових технологій)

Гриф надано Міністерством освіти і науки, молоді та спорту України
Лист № 1/11-20283 від 29.12.12

М 34 **Математична** статистика : навч. посіб. [Електронне видання] / [С. М. Григулич, В. П. Лісовська, О. І. Макаренко та ін.]. — К. : КНЕУ, 2015. — 203 с.
ISBN 978-966-483-997-3

Посібник містить теоретичний і практичний матеріал, відповідає програмі з математичної статистики, чинній у Київському національному економічному університеті ім. В. Гетьмана. Розв’язано багато задач економічного характеру та прикладів, за допомогою яких студенти зможуть набути відповідних практичних навиків з використання методів математичної статистики для обробки економічної інформації.

Призначено для студентів економічних спеціальностей усіх форм навчання, фахівців, які використовують у своїй діяльності методи математичної статистики.

УДК 519.22/.25:33(075.8)
ББК 22.172

ISBN 978-966-483-997-3

ПЕРЕДМОВА

Нагальною потребою сучасних економічної, соціально-політичної та інших суспільних наук є вивчення закономірностей, які виникають у масових випадкових явищах. Це завдання виконує математична статистика, базою й обґрунтуванням методів якої є теорія ймовірностей. Знання закономірностей, яким підпорядковуються масові випадкові події, дає можливість передбачати, як ці події відбуватимуться, що особливо важливо для галузей економіки, фінансів, маркетингу, управління виробництвом.

Пропонований навчальний посібник присвячений розгляду методів, способів обробки статистичної інформації і призначений для студентів усіх форм навчання Київського національного економічного університету імені Вадима Гетьмана. Посібник охоплює практично всі розділи математичної статистики, а саме:

— методи одержання і систематизації статистичного матеріалу та його графічне подання;
— методи розрахунку характеристик вибірки;
— статистична оцінка параметрів розподілу;
— перевірка статистичних гіпотез;
— елементи теорії кореляції.

У кожному параграфі посібника подано відповідний теоретичний матеріал з формулами, застосування яких ілюструється великою кількістю прикладів економічного характеру. Приклади розв’язано, роз’яснено економічний зміст параметрів, показано, як використовувати одержані результати в майбутній практичній діяльності.

Зміст посібника відповідає програмі з математичної статистики, чинній у Київському національному економічному університеті імені Вадима Гетьмана.

Автори сподіваються, що викладений матеріал стане в пригоді всім, хто використовує у своїй діяльності методи математичної статистики.

§ 1. ПРЕДМЕТ І ЗАВДАННЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ. СТАТИСТИЧНИЙ РОЗПОДІЛ, ЕМПІРИЧНА ФУНКЦІЯ, ПОЛІГОН, ГІСТОГРАМА

1.1. ПРЕДМЕТ І ЗАВДАННЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Математична статистика — це розділ математики, що вивчає закономірності, характерні для масових явищ і статистичних сукупностей. Тобто вивчення й установлення закономірностей, яким підпорядковуються масові випадкові явища, є предметом математичної статистики. Методологія математичної статистики, застосовувана для обробки статистичних даних, результатів спостережень, даних реальних або штучно поставлених дослідів, базується на методах теорії ймовірностей.

Особливу роль відіграє вибірковий метод математичної статистики, який широко використовується в управлінні, фінансах, маркетингу, організації виробництва тощо. Аналізуючи, здавалося б, випадкові явища, удається встановити відповідні закономірності та зв'язки між ними, зробити прогнози їхнього розвитку, оцінити параметри, які характеризують ці явища. Наявність такої інформації дає змогу прийняти обґрунтовані рішення за умов невизначеності.

Класична математична статистика має на меті в основному розв'язання двох завдань. Перше з них — це виявити й обґрунтувати способи збору та групування статистичних відомостей, одержаних у результаті спостережень за об'єктом дослідження або за результатами спеціально поставлених експериментів. Так, наприклад, організація збору інформації під час екзитпулу може суттєво вплинути на його результати і привести до грубих помилок.

Друге завдання математичної статистики полягає в розробленні методів аналізу статистичних даних, котре ґрунтується на теоремах і методах теорії ймовірностей. Зокрема, закон великих чисел, теорема Чебишова дають можливість обґрунтувати вибірковий метод.

Доречно виокремити завдання математичної статистики:

— перевірка статистичних гіпотез про закон невідомого розподілу або про значення параметрів відомого розподілу;

— побудова регресійних залежностей і оцінки рівня значущості параметрів цих залежностей;

— визначення кількості необхідних випробувань у процесі планування експерименту і т. ін.

Підсумовуючи, зазначимо, що ключовим завданням математичної статистики є створення і обґрунтування методів збору й обробки статистичних даних для одержання наукових та практичних висновків і рекомендацій з метою дальшого їх використання для аналізу закономірностей у масових випадкових явищах.

1.2. ГЕНЕРАЛЬНА І ВИБІРКОВА СУКУПНОСТІ

Під ***генеральною сукупністю*** розумітимемо всю множину статистичних одиниць, якісні або кількісні характеристики котрих потрібно дослідити.

На практиці найчастіше розглядаються скінченні генеральні сукупності. Кількість статистичних одиниць у них є скінченною величиною, хоч і може бути достатньо великою. Проте якщо кількість одиниць генеральної сукупності велика, з метою спрощення обчислень або полегшення теоретичних висновків припускається, що кількість одиниць сукупності нескінченна або прямує до нескінченності.

Розглянуте припущення дозволяє переносити результати, одержані в теоремах теорії ймовірностей, які називаються законами великих чисел, на скінченні генеральні сукупності. Основним результатом цих теорем є те, що за виконання певних умов сумарна поведінка достатньо великої кількості випадкових величин утрачає випадковий характер і стає практично закономірною.

Під ***вибірковою сукупністю*** будемо розуміти скінченну множину статистичних одиниць, випадково вибраних з генеральної сукупності.

Звичайно, точніші результати статистичного дослідження можна одержати за дослідження генеральної сукупності. Але в цьому разі можуть виникнути проблеми, пов'язані з великою кількістю одиниць генеральної сукупності або зі знищенням статистичної одиниці, або з великими матеріальними витратами і т. ін. Скажімо, з метою дослідження якості завезеного на елеватор зерна неможливо перевірити кожну зернину, що є в кузові автомобіля. Тому в таких випадках утворюють вибіркову сукупність, до складу якої входить скінченна кількість (як правило, не дуже велика) випадково вибраних одиниць з генеральної сукупності. У прикладі з якістю зерна навмання вибирають певну кількість проб з усієї маси зерна, утворюючи вибіркову сукупність.

Спираючись на теорему Чебишова, можна стверджувати, що на основі обстеження відносно невеликої випадкової вибірки, одержані результати й закономірності можна переносити з відповідною точністю на всю генеральну сукупність. У цьому полягає сутність вибіркового методу математичної статистики.

Кількість статистичних одиниць генеральної сукупності N і вибіркової сукупності n називають їх обсягами. Значення ознаки X, які попали у вибіркову сукупність (вибірку), називають ***варіантами*** і позначають $x_i,\ i=\overline{1,n}$.

Вибірки бувають повторними та безповторними. Вибірка називається ***повторною***, якщо перед вибором наступного елемента попередній повертається в генеральну сукупність. Вибірка називається ***безповторною***, якщо перед вибором наступного елемента попередній не повертається в генеральну сукупність.

Вибіркову сукупність називають ***репрезентативною***, якщо вона досить добре віддзеркалює основні характеристики генеральної сукупності. Щоб за даними вибірки мати можливість робити висновки про генеральну сукупність, вона має бути утворена з навмання взятих елементів, що певною мірою дозволяє знизити можливість помилок репрезентативності. Випадковість елементів у вибірці досягається завдяки принципу рівної можливості всіх елементів генеральної сукупності бути відібраними у вибірку.

Помилка репрезентативності — відхилення вибіркової сукупності за певними характеристиками від генеральної сукупності.

Чим більша величина відхилень, тим значніша помилка репрезентативності й тим нижча якість одержаних даних вибірки. Головне завдання на цьому етапі дослідження — урахувати помилку репрезентативності під час інтерпретації та узагальнення результатів дослідження, проведеного із застосуванням вибіркового методу.

Важливу роль у визначенні якості інформації, одержаної в результаті емпіричного дослідження, крім репрезентативності відіграють такі параметри, як надійність і валідність.

Надійність інформації — адекватність одержаних результатів дослідження випадкових явищ. Забезпечується вона врахуванням так званих випадкових помилок, які є неминучими внаслідок неоднорідності досліджуваних об'єктів. Чим однорідніші об'єкти обстеження і чим більший обсяг вибіркової сукупності, тим незначніша випадкова помилка вибірки і, відповідно, вища якість одержаної інформації. Випадкові помилки виникають також унаслідок низької якості інструментарію, непрофесійної роботи дослідника тощо.

Якість дослідження визначається величиною випадкової помилки (знаходять її за допомогою математичних формул), яку можна врахувати під час поширення висновків, зроблених на підставі вибіркового дослідження, на всю генеральну сукупність.

Валідність (***обґрунтованість***) ***інформації*** — відповідність результатів саме тим явищам і процесам, які передбачалося дослідити.

До зниження валідності можуть призвести не тільки помилки інструментарію, а й систематичні помилки вибірки.

Систематичні помилки виникають унаслідок неправильних вихідних статистичних даних про параметри контрольованих ознак генеральної сукупності, надто малого обсягу вибірки, хибного застосування способу відбору елементів для аналізу тощо.

Установити величину систематичних помилок за допомогою математичних формул неможливо, тому вони значно погіршують результати досліджень і взагалі можуть звести їх нанівець. Крім того, слід ураховувати, що надійність і валідність — самостійні параметри, які не залежать один від одного і характеризують якість дослідження з різних боків. Тому їх обов'язково треба визначати окремо.

Точність результатів вибіркового спостереження, нарешті, залежить від способу вибору елементів, міри коливання ознаки в сукупності та від кількості вибраних елементів.

Вирізняють такі види вибірок:

- власне випадкова вибірка, одержана випадковим вибором елементів без поділу їх на частини або групи;
- механічна вибірка, для якої елементи генеральної сукупності вибираються через деякий інтервал;
- типова вибірка, у котру навмання вибираються елементи з типових груп, на які за певною ознакою поділяється генеральна сукупність;
- серійна вибірка, в яку навмання потрапляють не елементи груп, а власне групи, які потім суцільно досліджуються.

Перед тим як визначити тип вибірки, зупиняються на методі (імовірнісний або цілеспрямований) вибору елементів вибірки.

Вибіркові спостереження за умови правильної їх організації і проведення дають достатньо вірогідні дані, придатні для використання у фінансово-господарському контролі. У контрольноаудиторському процесі до вибіркових спостережень вдаються з метою скорочення обсягу процедур, використовуваних у процесі дослідження об'єктів контролю.

Завдання, які вирішуються за допомогою вибіркового спостереження:

1) визначення середнього значення досліджуваної ознаки;
2) визначення питомої ваги (частки) досліджуваної ознаки в певній сукупності;
3) визначення середньої та граничної похибки вибірки;
4) знаходження меж для середньої і частки за повторного і безповторного відбору;
5) визначення потрібної чисельності вибірки;
6) поширення даних вибіркового спостереження на всю сукупність.

Відповідь на запитання про те, з якою ймовірністю можна стверджувати про збіг між генеральними й вибірковими узагальнювальними показниками, дає теорія вибіркового методу, що ґрунтується на законі великих чисел. За допомогою цього закону вирішують два взаємозв'язаних завдання:

— розраховують за заданої ймовірності межі можливих відхилень вибіркового показника від відповідного показника в генеральній сукупності;

— визначають імовірність перевищення встановленої межі можливими відхиленнями вибіркового показника від генерального.

Масові явища, які вивчає статистика, перебувають під впливом багатьох випадкових чинників. Тому використовуємо основний висновок граничних теорем теорії ймовірностей про те, що сукупна дія багатьох випадкових факторів приводить за деяких умов до результату майже незалежного від випадку. Оскільки вибіркове спостереження пов'язане з випадковими відхиленнями характеристик вибіркової й генеральної сукупностей, то ключове положення граничних теорем дає змогу стверджувати, що результати вибіркового спостереження вірогідні за достатньо великої кількості відібраних елементів. За цих умов вибіркові характеристики надійно відтворюють генеральні характеристики.

Випадкові похибки реєстрації для великої кількості спостережень не впливають суттєво на результат дослідження, оскільки вони взаємно погашаються, а тому від них можна абстрагуватись і надалі розглядати тільки похибки вибірки. Принцип строгої випадковості, який покладено в основу вибірки, забезпечує її об'єктивність, дає змогу встановити межі можливих похибок і дістати майже вірогідні дані для характеристики всієї сукупності явищ.

1.3. СТАТИСТИЧНІ РОЗПОДІЛИ ВИБІРКИ. ЕМПІРИЧНА ФУНКЦІЯ СТАТИСТИЧНОГО РОЗПОДІЛУ. ПОЛІГОН. ГІСТОГРАМА

Варіаційним рядом називається впорядкована статистична сукупність, елементи (варіанти) x_i якої розташовані в порядку зростання.

Якщо елемент x_i зустрічається n_i разів, тоді число n_i називається ***частотою*** елемента x_i, а відношення $w_i = \frac{n_i}{n}$ — ***відносною частотою*** елемента x_i.

Статистичним рядом (***розподілом***) називається послідовність пар (x_i, n_i), де x_i подано за зростанням значень.

У разі коли обсяг вибірки великий, статистичні дані записують у вигляді інтервального статистичного ряду.

Інтервальним статистичним рядом (***розподілом***) називається послідовність виділених інтервалів і частот, або відносних частот відповідних їм варіантів.

Для побудови інтервального статистичного ряду множину значень варіант розбивають на інтервали $[a_i; a_{i+1})$, тобто здійснюють їх групування. ***Кількість інтервалів k*** рекомендується розраховувати за формулою Стерджерса

$$k = 1 + 3{,}322 \lg n\,. \tag{1.1}$$

Довжину інтервалу, яку позначають літерою h (або Δ), найчастіше беруть однаковою. Підраховуючи кількість значень варіант, що потрапили в інтервал $[a_i; a_{i+1})$, отримують відповідні інтервалам частоти n_i для $i = \overline{1, k}$.

Нерівні інтервали застосовують тоді, коли кількісні зміни розміру ознаки мають неоднакове значення для нижчих і вищих груп. Так, якщо різниця лише в 2—4 роки для тих, хто одружується в молодому віці (до 30 років), відіграє велику роль, то для людей старшого віку така різниця буде набагато значніша — 10 років і більше.

Якщо значення ознаки змінюється рівномірно, то виділяють рівні інтервали груп. Їхню довжину визначають за формулою

$$h = \frac{x_{\max} - x_{\min}}{k}, \tag{1.2}$$

де $x_{\max} = \max\{x_1, x_2, ... x_n\}$; $x_{\min} = \min\{x_1, x_2, ... x_n\}$.

Як правило, статистичний ряд записується у вигляді таблиці, перший рядок якої містить варіанти x_i, а другий — відповідні їм частоти n_i (табл. 1.1):

Таблиця 1.1

Варіанти x_i	x_1	x_2	…	x_k
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

Інтервальний статистичний ряд у загальному вигляді можна подати таблицею (табл. 1.2):

Таблиця 1.2

Інтервали $[a_i; a_{i+1})$	$[a_1; a_2)$	$[a_2; a_3)$	…	$[a_k; a_{k+1}]$
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

Для статистичних рядів повинні виконуватися умови:

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

Для наочності використовують графічне зображення статистичних рядів у вигляді полігону частот (відносних частот) і, тільки в разі інтервального ряду, гістограми частот (відносних частот).

Полігоном частот (***відносних частот***) називається ламана лінія, що з'єднує прямолінійними відрізками точки з координатами (x_i; n_i) [або $(x_i; w_i)$) для $i = \overline{1,k}$ у разі дискретного статистичного ряду; (c_i; n_i) (або $(c_i; w_i)$) у разі інтервального ряду, де c_i — середина i-го інтервалу, $c_i = \frac{a_i + a_{i+1}}{2}$.

Гістограмою частот (***або просто гістограмою***), називають ступінчасту фігуру, яка складається з прямокутників, основами яких є частинні інтервали довжиною $h = x_{i+1} - x_i$, а висоти дорівнюють відповідному значенню частоти n_i, поділеному на h, тобто $\frac{n_i}{h}$ (за умови, що всі довжини інтервалів однакові).

Гістограмою відносних частот називається ступінчаста фігура, яка складається з прямокутників, основами яких є окремі інтервали довжиною $h = x_{i+1} - x_i$, а висоти дорівнюють відношенню відносної частоти w_i до h, тобто $\frac{w_i}{h}$.

За статистичним рядом можна знайти емпіричну функцію розподілу випадкової величини X.

Емпіричною функцією розподілу називається функція аргументу x, $x \in \mathrm{R}$, що визначає відносну частоту події $X < x$, що дорівнює відношенню числа n_x — кількості варіант менших за значення x — до обсягу вибірки n:

$$F^*(x) = W(X < x) = \frac{n_x}{n}. \tag{1.3}$$

Для статистичного ряду графік емпіричної функції розподілу має ступінчастий вигляд, для інтервального статистичного ряду графіком є кумулятивна крива (кумулята), яку одержують, з'єднуючи точки $(a_i; F^*(a_i))$, відрізками прямих, де a_i — кінці інтервалів, $i = \overline{1,k}$.

Властивості емпіричної функції:

1) $0 \le F^*(x) \le 1$;
2) $F^*(x)\big|_{x \le x_{\min}} = 0$, де $x_{\min}$ — найменше значення варіанти вибірки;
3) $F^*(x)\big|_{x > x_{\max}} = 1$, де $x_{\max}$ — найбільше значення варіанти вибірки;
4) $F^*(x)$ є неспадною функцією аргументу x, а саме $F^*(x_2) \ge F^*(x_1)$ для $x_2 \ge x_1$.

На відміну від емпіричної функції розподілу вибірки, інтегральну функцію розподілу генеральної сукупності називають теоретичною інтегральною функцією розподілу. З теореми Бернуллі випливає, що відносна частота події за ймовірністю прямує до ймовірності цієї події. Це означає, що емпірична функція вибірки за ймовірністю прямує до теоретичної функції розподілу генеральної сукупності. Тому емпірична функція розподілу вибірки є оцінкою теоретичної функції генеральної сукупності.

Для побудови кумуляти $F^*(x)$ для інтервального статистичного розподілу вибірки за основу береться припущення, що ознака на кожному частинному інтервалі має рівномірну щільність імовірностей. Тому кумулята матиме вигляд ламаної лінії, яка зростає на кожному частинному інтервалі і наближається до одиниці.

Приклад 1.1. Під час тестування службовців деякої компанії були одержані такі результати (у балах):

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

Побудувати дискретний статистичний ряд, полігон частот, емпіричну функцію розподілу та її графік.

Р о з в ' я з а н н я. Для побудови дискретного статистичного ряду записуємо в порядку зростання різні значення випадкової величини X і відповідні частоти (табл. 1.3).

Останній стовпчик таблиці використовується для перевірки правильності побудови статистичного ряду (усього в тестуванні взяли участь 25 осіб, тому сума частот має дорівнювати 25).

Таблиця 1.3

x_i	37	38	39	40	41	42	43	44	Сума
n_i	1	2	2	4	6	5	4	1	$n = \sum_{i=1}^{k} n_i = 25$

Полігон частот даного розподілу зображено на рис. 1.1.

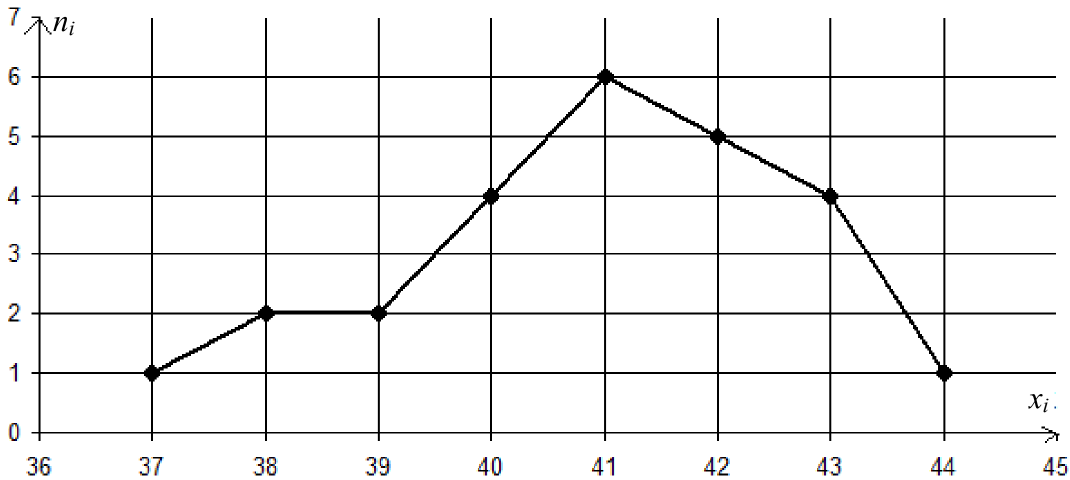


Рис. 1.1. Полігон частот

Для побудови емпіричної функції розподілу обчислимо суму частот варіант, менших за x (тобто $\sum\limits_{x_i < x} n_i$). Дістаємо:

якщо $x \le 37$, то $\sum\limits_{x_i<37} n_i = 0$, оскільки таких значень x_i немає, тоді $F^*(x) = 0$;

якщо $37 < x \le 38$, то $\sum\limits_{x_i<38} n_i = n_1 = 1$, а $F^*(x) = \dfrac{1}{25}$;

якщо $38 < x \le 39$, то $\sum\limits_{x_i<39} n_i = n_1 + n_2 = 1 + 2 = 3$, а $F^*(x) = \dfrac{3}{25}$;

якщо $39 < x \le 40$, то $\sum\limits_{x_i<40} n_i = n_1 + n_2 + n_3 = 1+2+2 = 5$, а $F^*(x) = \dfrac{5}{25}$;

якщо $40 < x \le 41$ то $\sum\limits_{x_i<41} n_i = n_1 + n_2 + n_3 + n_4 = 1+2+2+4 = 9$, а $F^*(x) = \dfrac{9}{25}$;

якщо $41 < x \le 42$, то $\sum\limits_{x_i<42} n_i = n_1 + n_2 + n_3 + n_4 + n_5 = 1+2+2+4+6 = 15$, а $F^*(x) = \dfrac{15}{25}$;

якщо $42 < x \le 43$, то $\sum\limits_{x_i<43} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 1+2+2+4+6+5 = 20$, а $F^*(x) = \dfrac{20}{25}$;

якщо $43 < x \le 44$, то $\sum\limits_{x_i<44} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 + n_7 = 1+2+2+4+6+5+4 = 24$, а $F^*(x) = \dfrac{24}{25}$;

якщо $x > 44$, то $\sum\limits_{x_i<x} n_i = 25$, адже всі значення x_i менші від будь-якого числа, що перевищує 44; тоді $F^*(x) = 1$.

Згідно з означенням емпірична функція має такий вигляд:

$$F^*(x) = \begin{cases} 0, & x \le 37, \\ \dfrac{1}{25}, & 37 < x \le 38, \\ \dfrac{3}{25}, & 38 < x \le 39, \\ \dfrac{5}{25}, & 39 < x \le 40, \\ \dfrac{9}{25}, & 40 < x \le 41, \\ \dfrac{15}{25}, & 41 < x \le 42, \\ \dfrac{20}{25}, & 42 < x \le 43, \\ \dfrac{24}{25}, & 43 < x \le 44, \\ 1, & x > 44 \end{cases} = \begin{cases} 0, & x \le 37, \\ 0{,}04, & 37 < x \le 38, \\ 0{,}12, & 38 < x \le 39, \\ 0{,}2, & 39 < x \le 40, \\ 0{,}36, & 40 < x \le 41, \\ 0{,}6, & 41 < x \le 42, \\ 0{,}8, & 42 < x \le 43, \\ 0{,}96, & 43 < x \le 44, \\ 1, & x > 44. \end{cases}$$

Графік одержаної емпіричної функції розподілу зображено на рис. 1.2.

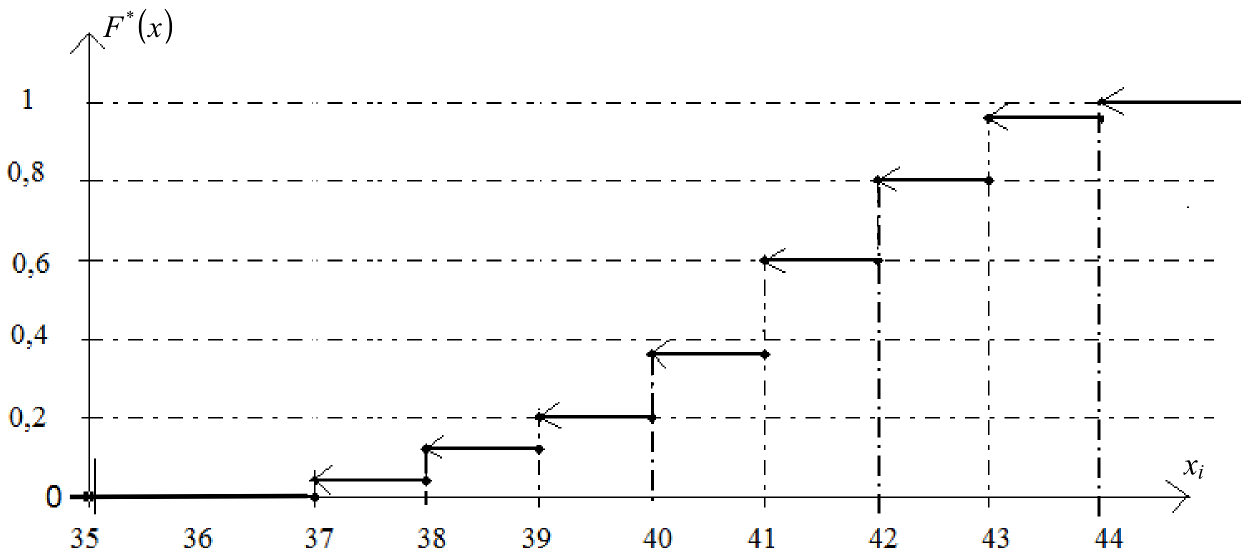


Рис. 1.2. Графік емпіричної функції розподілу

Приклад 1.2. За даними вибіркового дослідження було одержано розподіл родин за доходом на одного їх члена в умовних одиницях (табл. 1.4). Побудувати інтервальний статистичний ряд, гістограму частот, емпіричну функцію та її графік, полігон частот і полігон відносних частот.

Таблиця 1.4

28,92	27,54	22,36	29,09	32,19	26,04	17,06	26,83	24,55	33,22
17,53	30,07	36,27	24,24	26,03	31,05	13,94	14,56	21,40	23,04
13,09	38,84	25,57	22,87	6,11	27,79	25,68	16,30	17,93	24,37
28,92	27,54	22,36	29,06	32,19	26,04	17,06	26,83	24,55	33,22
17,53	30,07	36,27	24,24	26,03	31,05	13,94	14,56	21,40	23,04

Р о з в ’ я з а н н я. Табл. 1.4 містить 50 даних, тобто $n = 50$. Для побудови інтервального статистичного ряду знаходимо кількість інтервалів за формулою (1.1): $k = 1 + 3{,}322 \lg 50 \approx 6{,}477$, вибираємо $k = 7$; $x_{\min} = 6{,}11$, $x_{\max} = 38{,}84$; довжину кожного інтервалу знаходимо за формулою (1.2):

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{38{,}84 - 6{,}11}{7} \approx 4{,}68.$$

Отже, за початок першого інтервалу вибираємо $a_1 = x_{\min} = 6{,}11$. Тоді

$$a_2 = a_1 + h = 6{,}11 + 4{,}68 = 10{,}79.$$

Аналогічно,

$$a_3 = 15{,}47;\ a_4 = 20{,}15;\ a_5 = 24{,}83;\ a_6 = 29{,}51;\ a_7 = 34{,}19;\ a_8 = 38{,}87\,.$$

Підрахувавши кількість варіант, що потрапили в кожен інтервал, дістаємо інтервальний статистичний ряд (табл. 1.5).

Таблиця 1.5

$[a_i; a_{i+1})$	[6,11; 10,79)	[10,79; 15,47)	[15,47; 20,15)	[20,15; 24,83)	[24,83; 29,51)	[29,51; 34,19)	[34,19 38,87]
n_i	1	5	6	12	15	8	3

За даними таблиці 1.5 будуємо гістограму частот (рис. 1.3).

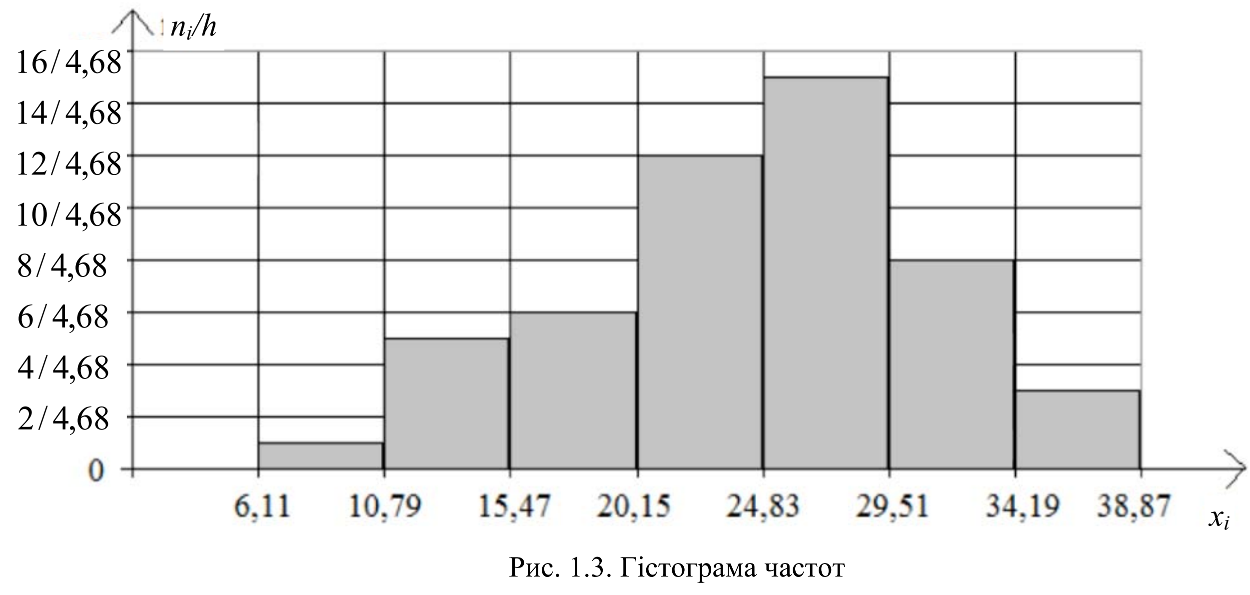


Рис. 1.3. Гістограма частот

Для побудови емпіричної функції розподілу обчислимо за формулою (1.3) суму відносних частот варіант менших за x [тобто $F^*(x)=\sum\limits_{x_i<x} w_i$]. За x беремо ліву границю кожного інтервалу. Отримаємо:

якщо , то $x \le a_1 = 6{,}11$, $F^*(x)=\sum\limits_{x_i<6,11} w_i = 0$, оскільки таких значень x_i немає;

якщо $x = a_2 = 10{,}79$, то $F^*(x)=\sum\limits_{x_i<10,79} w_i = w_1 = 0{,}02$;

якщо $x = a_3 = 15{,}47$, то $F^*(x)=\sum\limits_{x_i<15,47} w_i = w_1 + w_2 = 0{,}02 + 0{,}1 = 0{,}12$;

якщо $x = a_4 = 20{,}15$,то

$$F^*(x)=\sum\limits_{x_i<20,15} w_i = w_1 + w_2 + w_3 = 0{,}02 + 0{,}1 + 0{,}12 = 0{,}24;$$

якщо $x = a_5 = 24{,}83$, то $F^*(x)=\sum\limits_{x_i<24,83} w_i = w_1 + w_2 + w_3 + w_4 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 = 0{,}48$;

якщо $x = a_6 = 29{,}51$, то $F^*(x)=\sum\limits_{x_i<29,51} w_i = w_1 + w_2 + w_3 + w_4 + w_5 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 = 0{,}78$;

якщо $x = a_7 = 34{,}19$, то $F^*(x)=\sum\limits_{x_i<34,19} w_i = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 + 0{,}16 = 0{,}94$;

якщо $x = a_8 = 38{,}87$, то

$$F^*(x)=\sum\limits_{x_i<38,87} w_i = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7 =$$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 + 0{,}16 + 0{,}06 = 1.$

Останнє значення функції розподілу означає, що всі значення величини X менші за 38,87.

Згідно з означенням емпірична функція має такий вигляд:

x	$x \le 6{,}11$	10,79	15,47	20,15	24,83	29,51	34,19	$x \ge 38{,}87$
$F^*(x)$	0	0,02	0,12	0,24	0,48	0,78	0,94	1

Побудуємо графік емпіричної функції інтервального розподілу, з'єднуючи зазначені точки відрізками прямих (рис. 1.4).

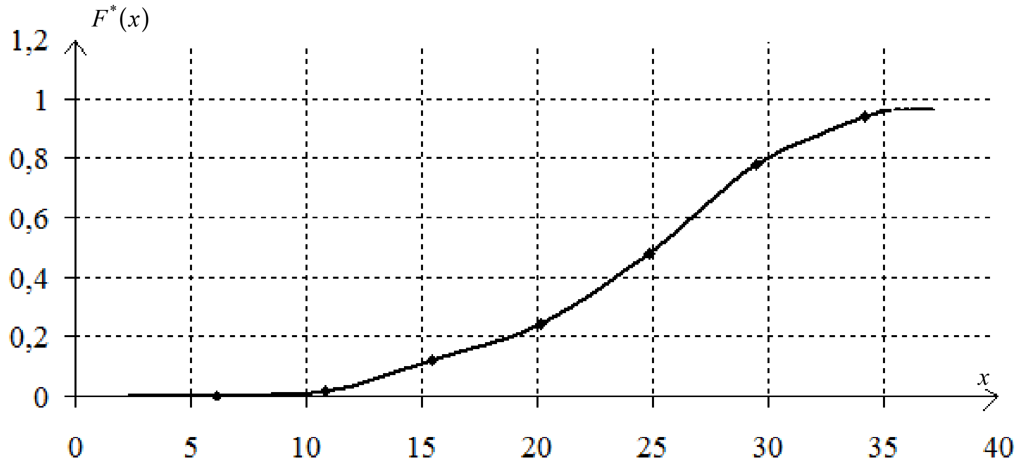


Рис. 1.4. Графік емпіричної функції розподілу

Для побудови полігону частот і полігону відносних частот обчислимо середини кожного інтервалу за формулою $c_i = \frac{a_i + a_{i+1}}{2}$. Дістанемо: $c_1 = \frac{6,11 + 10,79}{2} = 8,45$; $c_2 = 13,13$; $c_3 = 17,81$; $c_4 = 22,49$; $c_5 = 27,17$; $c_6 = 31,85$; $c_7 = 36,51$.

Розрахуємо відносні частоти:

$$w_1 = \frac{n_1}{n} = \frac{1}{50} = 0,02;\ w_2 = 0,1;\ w_3 = 0,12;\ w_4 = 0,24;\ w_5 = 0,3;\ w_6 = 0,16;\ w_7 = 0,06.$$

Результати оформимо у вигляді таблиці (табл. 1.6), останній стовпчик якої будемо використовувати для перевірки правильності розрахунків.

Таблиця 1.6

c_i	8,45	13,13	17,81	22,49	27,17	31,85	36,51	Перевірка
n_i	1	5	6	12	15	8	3	$\sum_{i=1}^{k} n_i = 50$
w_i	0,02	0,1	0,12	0,24	0,3	0,16	0,06	$\sum_{i=1}^{k} w_i = 1$

За даними табл. 1.6 будуємо полігон частот (рис. 1.5) і полігон відносних частот (рис. 1.6).

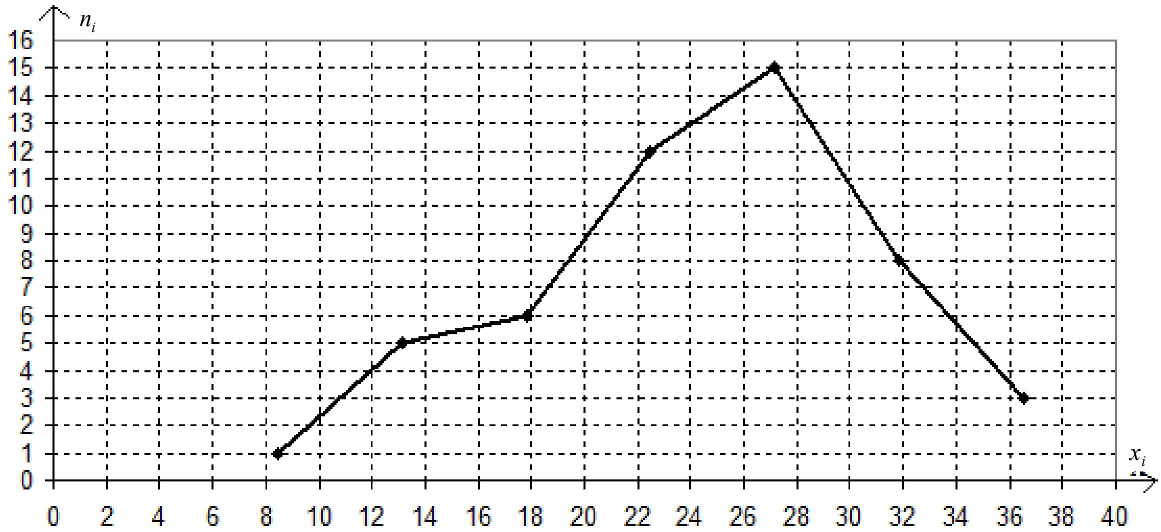


Рис. 1.5. Полігон частот

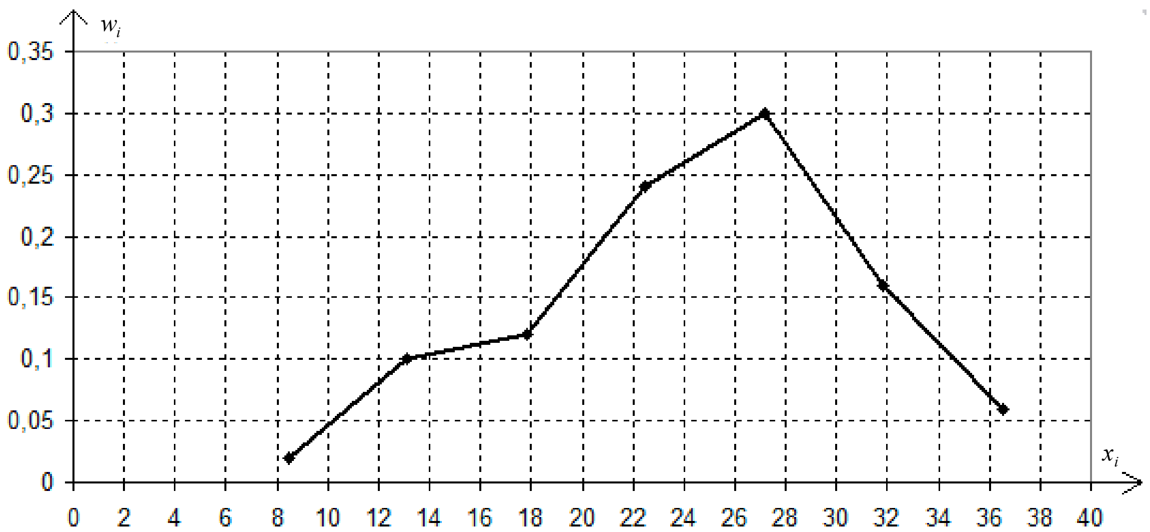


Рис. 1.6. Полігон відносних частот

§ 2. ЧИСЛОВІ ХАРАКТЕРИСТИКИ СТАТИСТИЧНОГО РОЗПОДІЛУ ВИБІРКИ

Кожна вибірка є частиною генеральної сукупності, яка має властивості (параметри), що певним чином її описують. Оскільки кожен статистичний розподіл вибірки може дати деяке уявлення про генеральну сукупність, то за допомогою вибіркових даних можна оцінити ці параметри, використовуючи відповідні вибіркові характеристики.

У попередньому параграфі розглядалися різні способи подання вибіркових даних, у тому числі графічний. Систематизація вибірки та побудова графіків, таких як полігон, гістограма, кумулята тощо, є достатньо важливою частиною дослідної роботи для підвищення наочності та інформативності результатів і передавання інформації. Вони є функціональними вибірковими характеристиками та дозволяють висунути гіпотезу про закон розподілу випадкової величини, що вивчається. Але для розв'язання низки практичних завдань немає потреби знати всі значення випадкової величини та відповідні їм імовірності, а зручніше користуватися деякими кількісними показниками, які давали б у стислому вигляді інформацію про випадкову величину. Такі показники будемо називати **числовими характеристиками**. Оскільки вибірковий розподіл не дає інформації про статистичні закономірності, які б описували числові характеристики генеральної сукупності, виникає потреба в знаходженні відповідних характеристик вибірки.

У цьому параграфі розглянемо числові характеристики статистичного (вибіркового) розподілу і пояснимо, які висновки можна зробити, вивчаючи їх. Питання про те, як оцінити відповідні параметри генеральної сукупності, використовуючи вибіркові (емпіричні) числові характеристики, розкриємо в наступних параграфах.

Характеристики вибіркового розподілу, які кількісно описують його структуру, поділяються:
— на характеристики положення;
— розсіювання;
— форми.

2.1. СЕРЕДНІ ВИБІРКОВІ ХАРАКТЕРИСТИКИ

До характеристик положення належать міри центральної тенденції статистичного розподілу вибірки, такі як вибіркова середня, мода, медіана, а також квантилі.

Як відомо, варіаційний ряд може бути представлений набором варіант x_i та їхніх частот n_i, $i=\overline{1,k}$. Однак оперувати таким набором досить складно. Для зручності досліджень треба вказати таку величину, яка, вірогідно відображаючи особливості даного ряду, була б підсумковою, зведеною. Таку величину називають **середньою.** Середня не може повністю замінити ряд, під час оперування нею втрачається частина інформації про вибірку. Але вона є таким собі «представником» усього ряду, оскільки навколо неї концентруються спостережувані значення ознаки. Тобто середня величина характеризує центральну тенденцію однорідної вибірки одним числом.

Зауваження. Умовами того, що середня величина є насправді узагальнювальною та може застосовуватись як така, що відображає типове для даного ряду, є якісна однорідність вибірки та достатньо великий її обсяг.

Застосування в дослідженнях середніх неоднорідних вибірок може призвести до спотвореного результату. Яскравим прикладом такого спотворення є відомий жарт про «середню температуру пацієнтів у лікарні», коли середньою в лікарні, включаючи морг та інфекційне відділення, є температура 36,6 °С. Такі дані є очевидно фіктивними і не віддзеркалюють реальної ситуації. Отже, обчисленню середніх вибіркових величин має передувати обґрунтоване виділення із сукупності, що вивчається, достатньо однорідних груп.

Середні величини поділяються на **степеневі середні** (серед яких — середнє арифметичне значення, середнє геометричне значення та ін.) та **структурні середні** (мода і медіана). Вибір того чи іншого виду середньої величини залежить від цілей економічного дослідження, економічної сутності досліджуваного показника тощо.

Нехай для вивчення генеральної сукупності щодо кількісної ознаки X маємо вибірку обсягу n: $x_1, x_2, \ldots, x_n$.

Вибірковою середньою (або середньою арифметичною) $\overline{x}_\text{в}$ будемо називати середнє арифметичне значення ознаки вибіркової сукупності.

Якщо всі значення $x_1, x_2, \ldots, x_n$ ознаки вибірки різні, то

$$\overline{x}_\text{в}=\frac{x_1+x_2+\ldots+x_n}{n}. \tag{2.1}$$

Якщо ж значення x_1, x_2, … , x_k мають відповідно частоти n_1, n_2, … , n_k, причому $\sum_{i=1}^{k} n_i = n$, то

$$\overline{x}_B = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i . \qquad (2.2)$$

У разі інтервального статистичного розподілу вибіркової сукупності вибіркову середню знаходять за формулою

$$\overline{x}_B = \frac{x_1^* n_1 + x_2^* n_2 + \ldots + x_k^* n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i , \qquad (2.3)$$

де x_1^*, x_2^*, … , x_k^* — середини відповідно першого, другого, … , k-го інтервалів; n_1, n_2, … , n_k — їхні частоти, причому $n = \sum_{i=1}^{k} n_i$.

Слід зазначити, що середня вибіркова, знайдена за інтервальним рядом, є наближеною, оскільки вона обчислюється за припущення, що розподіл окремих значень на кожному інтервалі є рівномірним, а це не завжди виконується. При цьому її точність залежить від того, наскільки розподіл значень усередині кожного інтервалу близький до такого, що його зважена середня збігається із серединою інтервалу. Крім того, точність залежить і від довжини інтервалу: чим вужчий інтервал, тим точніша середня вибіркова, знайдена за інтервальним рядом.

В економічних дослідженнях середню вибіркову часто застосовують для подання спостережуваних значень (приміром, середня ціна, середній обсяг продажу, середня заробітна плата, середня продуктивність праці тощо).

Приклад 2.1. Робітники деякої бригади одержують таку місячну заробітну плату (табл. 2.1):

Таблиця 2.1

Робітник	1	2	3	4	5	6	7	8	9	10
Зарплата, грн	1023	1164	1610	1718	2370	2850	3203	3890	3900	4150

Знайти середню місячну заробітну плату робітника даної бригади.

Р о з в ’ я з а н н я.

$$\overline{x}_B = \frac{1}{10}(1023 + 1164 + 1610 + 1718 + 2370 + 2850 + 3203 + 3890 + 3900 + 4150) =$$

$$= \frac{25878}{10} = 2587{,}8 \text{ грн.}$$

Отже, середньомісячна заробітна плата робітника бригади становить 2587,8 грн.

Зазначимо, що кожній вибірці обсягу n, яка відібрана з генеральної сукупності, відповідає своя $\overline{x}_в$. Отже, вибіркова середня є випадковою величиною, а значить, можна розглядати розподіл (теоретичний та емпіричний) вибіркових середніх і його числові характеристики.

Як буде показано в наступних параграфах, вибіркова середня є точковою оцінкою математичного сподівання випадкової величини *X*, що вивчається, яке надалі будемо називати **генеральною середньою** і позначати $\overline{x}_Г$.

Властивості вибіркової середньої. Наведемо деякі математичні властивості вибіркової середньої, що розкривають її суть.

1. Сума відхилень усіх спостережуваних значень ознаки від вибіркової середньої дорівнює нулю.

Справді, $\sum_{i=1}^{n}(x_i - \overline{x}_B) = \sum_{i=1}^{n} x_i - n \cdot \overline{x}_B = 0$.

2. Якщо всі варіанти збільшити (або зменшити) в m разів, то вибіркова середня також збільшиться (зменшиться) в таку саму кількість разів.

Наприклад, якщо грошові банківські вклади громадян скоригувати на рівень інфляції 1,5, то середній розмір вкладу відповідно збільшиться в 1,5 раза.

3. Якщо всі варіанти збільшити (зменшити) на однакову величину A, то й вибіркова середня збільшиться (зменшиться) на таку саму величину.

4. У разі пропорційного збільшення (зменшення) частот вибіркова середня не змінюється. Але її величина зміниться, якщо відбудуться структурні зрушення.

Скажімо за незмінної курсової вартості акцій окремих емітентів середня вартість акцій може зменшитись унаслідок збільшення частки «дешевих» акцій серед загальної їх кількості.

Властивості 2—4 пропонуємо читачеві довести самостійно.

5. Якщо вибіркова сукупність обсягу n розбита на s груп, що не перетинаються ($\sum_{i=1}^{s} n_i = n$, де n_i — кількість одиниць сукупності з i-ї групи), то загальна вибіркова середня $\bar{x}_в$ дорівнює середньому арифметичному групових середніх $\bar{x}_{вi}$ ($\bar{x}_{вi} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j$, де x_j, $j = \overline{1, n_i}$ — значення ознаки з i-ї групи), узятих з вагою n_i.

Справді, за означенням $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} x_j = \frac{1}{n}\sum_{i=1}^{s} n_i \cdot \bar{x}_{вi}$.

Приклад 2.2. Нехай на заводі три цехи, у першому з яких працюють 100 робітників, у другому — 500, у третьому — 400. Середньомісячна заробітна плата робітників першого цеху — 2562 грн, другого — 1858 грн і третього — 3104 грн. Знайти середню місячну заробітну плату в цілому на заводі.

Розв'язання. Скориставшись властивістю 5 вибіркової середньої, дістанемо

$$\bar{x}_в = \frac{1}{1000}(100 \cdot 2562 + 500 \cdot 1858 + 400 \cdot 3104) = 2426{,}8 \text{ грн.}$$

Отже, середньомісячна зарплата робітників заводу становить 2426,8 грн.

Розглянемо інші види степеневих середніх величин, хоч коло їх застосувань значно вужче, ніж у вибіркової середньої.

Якщо для заміни індивідуальних значень ознаки на середню величину треба залишити незмінним їх добуток (а не суму, як у разі вибіркової середньої), слід застосовувати **середню геометричну**:

$$\bar{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}, \tag{2.4}$$

де n — обсяг вибірки.

У разі коли значення ознаки x_1, x_2, … , x_k мають відповідно частоти n_1, n_2, … , n_k, причому $\sum_{i=1}^{k} n_i = n$, знаходять середню геометричну зважену:

$$\bar{x}_{геом} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}. \tag{2.5}$$

Середня геометрична найбільш широко застосовується для знаходження середніх темпів зростання. Якщо, наприклад, унаслідок інфляції ціна на певний товар зросла за перший рік у 2 рази, за другий — ще в 1,2 раза до рівня попереднього року, то за два роки ціна зросла в $2 \cdot 1{,}2 = 2{,}4$ раза. Тоді для знаходження середньорічного зростання ціни середнє арифметичне ($\bar{x}_в$) непридатне, адже тоді б середньорічне зростання ціни дорівнювало б $\frac{2+1{,}2}{2} = 1{,}6$ раза, і, відповідно, за два роки ціна зросла б у $1{,}6 \cdot 1{,}6 = 2{,}56$ раза, а не у 2,4. Правильне ж значення середньорічного зростання ціни дає середнє геометричне значення: $\bar{x}_{геом} = \sqrt{2{,}4} \approx 1{,}55$ (раза).

Зазначимо, що у разі коли дані про темпи зростання подані у вигляді динамічного ряду, формула середньої геометричної набуває вигляду:

$$\overline{t} = \sqrt[n-1]{\frac{y_n}{y_1}}\,, \qquad (2,6)$$

де $y_1,\ y_2,\ \ldots,\ y_n$ — значення ознаки динамічного ряду, $\overline{t}$ — середній темп зростання розглядуваної ознаки.

Геометрична середня також застосовується у випадках, коли окремі значення ознаки є далеко від усіх інших значень. Такі екстремальні значення на середню арифметичну впливають більше, ніж на середню геометричну, і тому вона як середня величина ліпше характеризує даний ряд.

Як уже зазначалось, вибіркова середня застосовується, у разі, коли варіанти вибіркової сукупності x_i та їхні частоти n_i відомі. А коли інформації про частоти для окремих значень варіант немає, і відомі лише добутки $x_i n_i$, застосовується **середня гармонійна зважена**. Щоб знайти середню величину в такому разі, позначимо $x_i n_i = v_i$. Тоді $n_i = \frac{v_1}{x_i}$ і середня арифметична, виражена через відомі дані x_i та v_i, набере вигляду:

$$\overline{x}_{\text{гарм}} = \frac{\sum_{i=1}^{k} v_i}{\sum_{i=1}^{k} \frac{v_i}{x_i}} = \frac{v_1 + v_2 + \ldots + v_k}{\frac{v_1}{x_1} + \frac{v_2}{x_2} + \ldots + \frac{v_k}{x_k}}. \qquad (2,7)$$

Така середня величина називається **середньою гармонійною**. Вона є перетвореною формою середньої арифметичної $\overline{x}_{\text{в}}$ і тотожна їй. Тобто замість знаходження середньої гармонійної можна знайти середню вибіркову, але для цього спочатку треба знайти частоти кожної з варіант, що закладені у формулі (2.7).

Отже, середня гармонійна застосовується, коли відомий чисельник логічної формули для знаходження середньої величини і невідомий знаменник. Найчастіше вона використовується для розрахунку загальної середньої із середніх групових, коли невідомі обсяги груп.

Приклад 2.3. У табл. 2.2 подано дані про наявні доходи населення за регіонами України у III кварталі 2010 р.[1]. Знайти середній наявний дохід у розрахунку на одну особу в цілому по Україні.

Таблиця 2.2

№	Область	Усього, млн грн, v_i	У розрахунку на одну особу, грн, x_i
1	Автономна Республіка Крим	8489	4321,0
2	Вінницька	7662	4650,1
3	Волинська	4317	4161,8
4	Дніпропетровська	17027	5085,4
5	Донецька	24211	5436,0
6	Житомирська	5593	4356,6
7	Закарпатська	4442	3565,3
8	Запорізька	9254	5119,2
9	Івано-Франківська	5686	4118,8
10	Київська	8312	4831,7
11	Кіровоградська	4334	4268,3
12	Луганська	10786	4681,6
13	Львівська	11031	4329,8
14	Миколаївська	5196	4376,3
15	Одеська	9561	4001,3
16	Полтавська	7254	4851,5
17	Рівненська	4622	4011,1

[1] Держкомстат України [Електронний ресурс]: експрес-випуск. — Січ. 2011. — Режим доступу : http://www.ukrstat.gov.ua

Закінчення табл. 2.2

№	Область	Усього, млн грн, v_i	У розрахунку на одну особу, грн, x_i
18	Сумська	5378	4602,5
19	Тернопільська	4198	3860,2
20	Харківська	13130	4752,1
21	Херсонська	4819	4414,2
22	Хмельницька	5821	4370,4
23	Черкаська	5322	4119,2
24	Чернівецька	3394	3752,8
25	Чернігівська	5246	4744,1
26	м. Київ	23898	8577,6
27	м. Севастополь	1657	4353,7
	Разом:	220640	

Р о з в ’ я з а н н я. Логічна формула для знаходження потрібної середньої така:

$$\text{Середній дохід} = \frac{\text{Загальний дохід, млн грн}}{\text{Кількість осіб, млн осіб}}.$$

Оскільки інформації про знаменник немає, скористаємося формулою середньої гармонійної (2.7), знайшовши перед цим суму:

$$\sum_{i-1}^{27}\frac{v_i}{x_i} = \frac{8489}{4321} + \frac{7662}{4650,1} + \frac{4317}{4161,8} + \frac{17027}{5085,4} + \frac{24211}{5436} + \frac{5593}{4356,6} + \frac{4442}{3565,3} + \frac{9254}{5119,2} +$$
$$+\frac{5686}{4118,8} + \frac{8312}{4831,7} + \frac{4334}{4268,3} + \frac{10786}{4681,6} + \frac{11031}{4329,8} + \frac{5196}{4376,3} + \frac{9561}{4001,3} + \frac{7254}{4851,5} +$$
$$+\frac{4622}{4011,1} + \frac{5378}{4602,5} + \frac{4198}{3860,2} + \frac{13130}{4752,1} + \frac{4819}{4414,2} + \frac{5821}{4370,4} + \frac{5322}{4119,2} + \frac{3394}{3752,8} +$$
$$+\frac{5246}{4744,1} + \frac{23898}{8577,6} + \frac{1657}{4353,7} \approx 45,893 \text{ млн осіб.}$$

Отже:

$$\bar{x}_{\text{гарм}} = \frac{\sum_{i=1}^{k} v_i}{\sum_{i=1}^{k}\frac{v_i}{x_i}} = \frac{220640}{45,893} \approx 4807,7 \text{ грн на особу.}$$

У разі коли частоти всіх варіант дорівнюють одиниці, застосовується **середня гармонійна проста**, яку знаходять за формулою

$$\bar{x}_{\text{гарм}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}, \tag{2,8}$$

де n — обсяг розглядуваної сукупності.

Властивість середньої гармонійної

Якщо сукупність розбита на s груп, що не перетинаються, з обсягами відповідно n_1, n_2, … , n_s, причому $\sum_{i=1}^{s} n_i = n$, а $\bar{x}_{1\text{гарм}}$, $\bar{x}_{2\text{гарм}}$, … , $\bar{x}_{s\text{гарм}}$ — гармонійні середні цих груп, то загальну середню гармонійну за всією сукупністю можна знайти за формулою

$$\bar{x}_{\text{гарм}} = \frac{n_1 + n_2 + \ldots + n_s}{\frac{n_1}{\bar{x}_{1\text{гарм}}} + \frac{n_2}{\bar{x}_{2\text{гарм}}} + \ldots + \frac{n_s}{\bar{x}_{s\text{гарм}}}}. \qquad (2,9)$$

Цю властивість пропонуємо читачеві довести самостійно.

Підсумуємо все викладене, що стосується степеневих середніх. Будь-яка середня величина має знаходитись так, щоб за заміни нею кожного значення ознаки не змінювався підсумковий, узагальнювальний, показник, пов'язаний з показником, що усереднюється. Наприклад, за заміни фактичних швидкостей на окремих відрізках шляху на середню швидкість не має змінитись відстань, пройдена за той самий час транспортним засобом; за заміни всіх заробітних плат працівників на середню не має змінитись фонд заробітної плати тощо. Тобто в кожному конкретному випадку існує лише один вид степеневої середньої, який відповідає сутності та властивостям соціально-економічного явища, що вивчається.

Зазначимо, що на тому самому вихідному матеріалі значення розглянутих видів середніх різні. У загальному випадку співвідношення між ними таке:

$$\bar{x}_{\text{гарм}} \le \bar{x}_{\text{геом}} \le \bar{x}_{\text{арифм}} = \bar{x}_{\text{в}}.$$

Розглянемо структурні середні величини — моду та медіану. Для обчислення степеневих середніх, зокрема найчастіше вживаної вибіркової середньої, використовуються всі відомі значення ознаки, тимчасом як мода та медіана визначаються структурою розподілу. Тому ці характеристики називаються структурними середніми.

Мода і медіана у більшості випадків характеризують значення варіанти, яка має певне положення у варіаційному ряді. Їх часто застосовують як середні числові характеристики, у разі коли знаходження вибіркової середньої неможливе або недоцільне.

Модою (M_o) точкового статистичного розподілу називають значення ознаки, що зустрічається найчастіше в даній сукупності. Її пошук не потребує обчислень — модою є варіанта з найбільшою частотою. Зазначимо, що мода не містить жодної інформації про те, наскільки широко розповсюджене дане значення ознаки. Тобто вона не відбиває міри модальності.

Моду широко застосовують в економічних дослідженнях. Її використовують, коли потрібно охарактеризувати величину ознаки, яка зустрічається найчастіше. Наприклад, знайти розмір взуття чи одягу, попит на який є максимальним, ціну на ринку, за якою було продано найбільше товару, тощо.

Приклад 2.4. За даними табл. 2.3 знайти модальний розмір жіночого взуття.

Таблиця 2.3

Розмір взуття	35	36	37	38	39	40	41
Кількість проданих пар	3	15	21	32	17	5	1

Р о з в ' я з а н н я. Модальним є 38 розмір, адже взуття саме цього розміру було продано найбільше — 32 пари.

Розглянемо інтервальний розподіл вибірки. Припустимо, що всі інтервали мають однакову довжину h. Довизначимо моду в цьому разі графічно. Побудуємо гістограму частот даного інтервального ряду і виберемо прямокутник, що має найбільшу висоту (назвемо його модальним). Природно припустити, що мода міститься на інтервалі (x_i, x_{i+1}), що є основою цього прямокутника (рис. 2.1). При цьому, якщо прямокутник гістограми, що передує модальному, має більшу висоту, ніж наступний за модальним, то мода міститься ближче до точки x_i, ніж до x_{i+1}. За означенням модою в цьому разі є абсциса точки перетину прямих AC та BD. Знайдемо її.

Нехай $M_o = x_i + \Delta x$. Через точку *B* проведемо пряму, паралельну *AC*. Нехай *E* — точка перетину побудованої прямої з прямою *CD*. З подібності трикутників *AOB* та *EBD* випливає:

$$\frac{|OF|}{|BC|} = \frac{|AB|}{|ED|} \text{ або } \frac{\Delta x}{x_{i+1} - x_i} = \frac{|AB|}{|DC| + |CE|}.$$

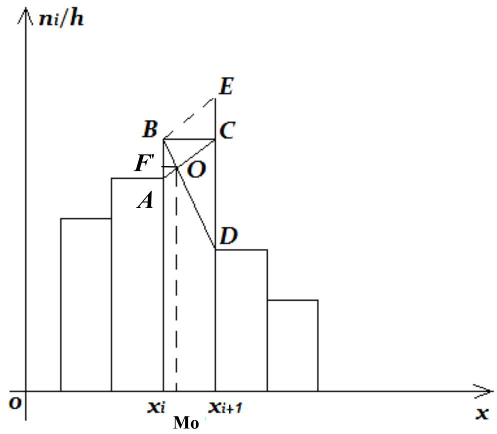


Рис. 2.1

Оскільки $|AB| = \frac{1}{h}(n_{\text{Mo}} - n_{\text{Mo-1}})$, $|DC| + |CE| = \frac{1}{h}[(n_{\text{Mo}} - n_{\text{Mo+1}}) + (n_{\text{Mo}} - n_{\text{Mo-1}})$, остаточно маємо

$$\text{Mo} = x_i + h\frac{n_{\text{Mo}} - n_{\text{Mo}-1}}{2n_{\text{Mo}} - n_{\text{Mo}-1} - n_{\text{Mo}+1}}, \qquad (2.10)$$

де x_i — початок модального інтервалу (тобто такого, що містить моду);

n_{Mo} — частота модального інтервалу;

$n_{\text{Mo}-1}$ — частота інтервалу, що передує модальному;

$n_{\text{Mo}+1}$ — частота інтервалу, наступного за модальним;

h — довжина модального інтервалу.

Отже, для знаходження моди за інтервального розподілу вибірки спочатку потрібно визначити модальний інтервал, тобто інтервал з найбільшою частотою ознаки. Далі, виконавши розрахунки за формулою (2.10), знайти значення M_{o}. Зазначимо, що коли інтервали мають різну довжину, модальним є інтервал з найбільшою щільністю, яку знаходять як відношення частоти даного інтервалу до його довжини $\frac{n_i}{h_i}$. Тоді аналогічно до виведення формули (2.10) можна знайти формулу для обчислення моди й у цьому разі. Вона матиме вигляд:

$$\text{Mo} = x_i + h\frac{y_{\text{Mo}} - y_{\text{Mo}-1}}{2y_{\text{Mo}} - y_{\text{Mo}-1} - y_{\text{Mo}+1}},$$

де y_{Mo}, $y_{\text{Mo}-1}$, $y_{\text{Mo}+1}$ — щільності відповідно модального, передуючого модальному та наступного за модальним інтервалів.

Приклад 2.5. Інтервальний розподіл річного доходу працюючих жінок у США[1] (1988) подано в табл. 2.4. Знайти моду.

Таблиця 2.4

Дохід, в дол.	до 5000	5000— 10 000	10000— 15 000	15000— 20 000	20000— 25 000	понад 25 000
Кількість працюючих жінок, тис. осіб	922	3813	7111	6601	4685	6676

Р о з в ’ я з а н н я. Модальний інтервал 10 000—15 000, адже його частота є максимальною — 7111 тис. осіб. За формулою (2.10) дістаємо:

$$\text{Mo} = 10000 + 5000 \cdot \frac{7111 - 3813}{2 \cdot 7111 - 3813 - 6601} = \$\,14\,330.$$

Отже, модальний річний дохід $ 14 330.

[1] Information Please Almanac // The New Universe of Information. 1990. — P. 64.

При знаходженні моди може виникнути ситуація, коли розподіл вибірки має кілька значень ознаки з максимальною частотою. Такий розподіл називається **мультимодальним**, а, якщо таких значень два — **бімодальним**. Серед структурних середніх таку властивість має лише мода, значення решти характеристик єдині. У разі мультимодального розподілу є підстави говорити про неоднорідність сукупності. Якщо ж усі значення ознаки зустрічаються однаково часто, то кажуть, що ряд не має моди.

Мода — це саме та середня характеристика, яка найчастіше застосовується для даних, що мають нечислову природу. Наприклад, за допомогою моди визначають тип покупця, що найчастіше зустрічається, або колір автомобілів, який має найбільшу популярність у покупців. Ця інформація може бути корисна для прогнозування продажу, планування виробництва тощо.

Медіаною (Me) називається таке можливе значення ознаки, яке міститься посередині варіаційного ряду, тобто ділить його на дві рівновеликі частини за кількістю спостережуваних значень. Якщо обсяг вибірки n — число непарне, то медіаною є член варіаційного ряду з номером $\frac{n+1}{2}$. У разі, коли n парне, медіана дорівнює середньому арифметичному $x_{\frac{n}{2}}$ та $x_{\frac{n}{2}+1}$ елементів ряду, тобто $\mathrm{Me}=\frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right)$. Наприклад, для варіаційного ряду 5, 6, 7, 10, 21 медіаною є варіанта 7; для ряду 1, 3, 4, 7, 8, 12 медіана дорівнює $\frac{1}{2}(4+7)=5{,}5$.

Приклад 2.6. Розглянемо дані прикладу 2.4 — розподіл розміру жіночого взуття (табл. 2.3). Знайдемо медіану даного ряду.

Розв'язання. Обсяг вибірки $n=94$ — парне число. Отже, медіаною є середнє арифметичне значень розмірів взуття, які у варіаційному ряді розташовуються під 47 та 48 номерами. Неважко переконатися, що обидва ці значення дорівнюють 38. Тож медіанний розмір взуття — 38.

Отже, у дискретному статистичному ряді медіаною є варіанта, для якої нагромаджена частота перевищує половину обсягу сукупності, за умови що нагромаджена частота попередньої варіанти менша за 0,5 n. Якщо ж нагромаджена частота деякої варіанти дорівнює 0,5 n, тоді медіану знаходять як середнє арифметичне цієї та наступної за нею варіанти дискретного ряду.

Знайдемо медіану за інтервального статистичного розподілу. Для цього спочатку визначимо інтервал, що містить медіану. Медіанним називається такий інтервал, нагромаджена (кумулятивна) частота якого дорівнює половині суми частот або перевищує цю суму, причому нагромаджена частота попереднього інтервалу менша від такої півсуми. Нехай x_i — початок медіанного інтервалу, довжина якого дорівнює h, частота — n_{Me}. Тоді $n_{\mathrm{Me}}. = S_{\mathrm{Me}} - S_{\mathrm{Me}-1}$, де S_{Me} та $S_{\mathrm{Me}-1}$ — нагромаджена частоти медіанного інтервалу та інтервалу, що йому передує, відповідно. Природно припустити, що всередині інтервалу варіанти розташовані рівномірно, тобто приріст частоти пропорційний приросту довжини інтервалу:

$$n_{\mathrm{Me}} : h = \left(\frac{1}{2}n - S_{\mathrm{Me}-1}\right) : (\mathrm{Me} - x_i).$$

Звідси

$$\mathrm{Me} = x_i + h\frac{0{,}5n - S_{\mathrm{Me}-1}}{n_{\mathrm{Me}}}. \qquad (2{,}11)$$

Отже, для розрахунку значення медіани інтервального розподілу вибірки можна використовувати формулу (2.11). Зазначимо, що медіанний інтервал може збігатися з модальним лише в окремих випадках; взагалі кажучи, це різні інтервали.

Приклад 2.7. Знайдемо медіанний річний дохід працюючих жінок за даними прикладу 2.5 (табл. 2.4).

Розв'язання. Для цього обчислимо півсуму частот:

$$0{,}5\ \mathrm{n} = 0{,}5(922 + 3813 + 7111 + 6601 + 4685 + 6676) = 14\,904.$$

Медіанним є інтервал 15 000—20 000, оскільки його нагромаджена частота дорівнює 18 447 (перевищує 0,5 n), тоді як нагромаджена частота попереднього інтервалу — 11 846 (менша за 0,5 n). За формулою (2.11) обчислимо медіану:

$$M_e = 15000 + 5000 \cdot \frac{14904 - 11846}{6601} = 17316{,}3 \, .$$

Отже, медіанний річний дохід дорівнює \$ 17316,3. Це означає, що рівно 50 % працюючих жінок мають дохід менший за \$ 17 316,3 і, відповідно, стільки саме — більший.

Графічно медіаною є абсциса точки на кумуляті, ордината якої дорівнює 0,5 (рис. 2.2). Формулу для знаходження медіани можна записати також, використовуючи значення емпіричної функції. Для цього поділимо на n знаменник та чисельник дробу формули (2.11).

Дістанемо

$$\mathrm{Me} = x_i + h \cdot \frac{0{,}5 - F^*(x_i)}{F^*(x_{i+1}) - F^*(x_i)}\,, \tag{2.12}$$

де x_i та x_{i+1} — відповідно початок і кінець медіанного інтервалу.

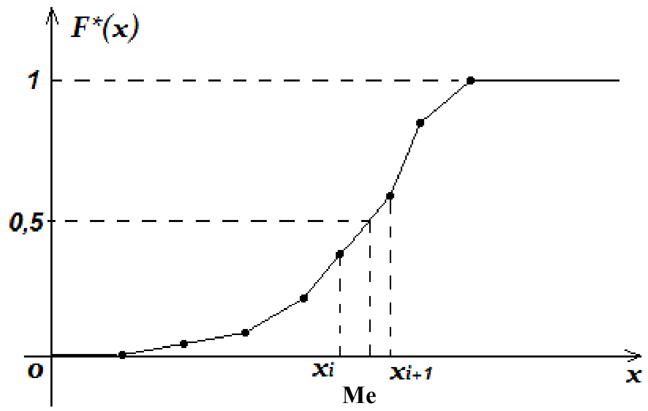


Рис. 2.2

Властивості медіани

1. Значення медіани не залежить від вибіркових значень ознаки, що містяться по обидва боки від неї. Завдяки цій властивості медіана в деяких випадках має певні переваги над вибірковою середньою. Якщо на значення останньої можуть досить суттєво впливати випадкові коливання крайніх варіант вибірки (за досить невеликого її обсягу), то значення медіани за таких коливань залишиться незмінним. Крім того, у разі інтервального розподілу з відкритими крайніми інтервалами медіана часто є ліпшою характеристикою центральної тенденції, ніж середнє арифметичне значення.

2. Сума абсолютних відхилень усіх значень ознаки від медіани є мінімальною проти суми відхилень від будь-якої іншої величини. Тобто функція $g(t) = \sum_{i=1}^{n} |x_i - t|$ досягає свого мінімуму при $t =$ Me. Ця властивість медіани застосовується в маркетинговій діяльності, у проектуванні розташування транспортних зупинок, АЗС, станцій техобслуговування тощо.

Для вивчення вибіркової сукупності в деяких випадках мода та медіана є більш прийнятними, ніж вибіркова середня. Адже їхні значення часто є певними варіантами ряду, тоді як вибіркова середня є величиною абстрактною. Тому, наприклад, під час планування обсягів виробництва та вивчення попиту на певний товар, як правило за середню величину береться модальне значення ознаки. Медіана частіше використовується для аналізу різного роду витрат та надходжень коштів. На державному рівні медіана може використовуватись для встановлення офіційного прожиткового мінімуму або рівня бідності. Так, у різних країнах за прожитковий мінімум беруть 40, 50 або 60 % медіанного доходу.

Закінчуючи огляд структурних середніх, зауважимо, що мода та медіана, на відміну від вибіркової середньої (див. властивість 5), не можуть бути виражені через відповідні значення модальних і медіанних значень груп, на які розбита розглядувана сукупність. Тобто арифметичні операції з Mo та Me неможливі, значення цих числових характеристик для всієї сукупності треба знаходити заново.

Розглянуті в цьому пункті середні величини $\bar{x}_в$, Mo і Me, взагалі кажучи, є різними числами, вони збігаються лише в разі симетричного розподілу частот статистичного ряду. Тому, аналізуючи співвідношення між цими величинами, можна оцінити асиметрію даного ряду. Крім того, такий аналіз дозволяє зробити певні висновки щодо форми закону розподілу випадкової величини, яка вивчається. Більш детально про це йтиметься в п. 2.3.

Крім розглянутих раніше середніх величин до характеристик положення належать квантилі. Як відомо, медіана є числовою характеристикою, яка має таку властивість: рівно 50 % значень ознаки розташовані ліворуч від медіани і 50 % — праворуч у варіаційному ряді. Природним узагальненням медіани є поняття квантиля.

Квантиль — це таке число, яке ділить варіаційний ряд у заданій пропорції. Тобто p-квантилем (або квантилем порядку p) статистичного розподілу є число x_p, нижче від якого у варіаційному ряді міститься p-та частина всіх спостережуваних значень. Залежно від кількості частин, на які вони розбивають сукупність, квантилі поділяють на такі види.

1. Квартилі Q_1, Q_2, Q_3 — ділять варіаційний ряд на 4 рівних частини за кількістю спостережуваних значень ознаки. Тобто, Q_1 — це 0,25-квантиль, його називають нижнім (або першим) квартилем; Q_2 — 0,5-квантиль (медіана, або другий квартиль); Q_3 — 0,75-квантиль (верхній, або третій, квартиль).

2. Децилі D_1, D_2, … , D_9 — ділять варіаційний ряд на 10 рівних частин. Так, наприклад, перший дециль D_1 відокремлює 10 % найменших значень ознаки у варіаційному ряді, тобто є 0,1-квантилем.

3. Перцентилі (або процентилі) P_1, P_2, … , P_{99} — ділять варіаційний ряд на 100 рівних частин. Тобто p-м перцентилем називають квантиль порядку $p/100$. Відповідно, медіана є 50-м перцентилем, перший і третій квартилі — 25-м і 75-м перцентилями.

Обчислюється p-квантиль за таким правилом. Якщо np не є цілим числом, то p-квантилем є член варіаційного ряду з номером int (np + 1) [int (a) — ціла частина числа a]. Якщо ж np — ціле, то p-квантилем є середнє арифметичне елементів ряду $x_{\text{пр}}$ та $x_{\text{пр}+1}$. У цьому разі p-квантиль може не збігатися з жодним членом даного варіаційного ряду, він є півсумою двох з них.

Приклад 2.8. Знайдемо нижній квартиль (Q_1) і 60-й перцентиль (P_{60}) для даних з прикладу 2.1 (табл. 2.1).

$Q_1 = P_{25}$, тому в цьому разі $np = 10 \cdot \dfrac{25}{100} = 2{,}5$. Отже, оскільки int (2,5 + 1) = 3, нижній квартиль збігається з третім членом варіаційного ряду $Q_1 = x_3 = 1610$ грн.

Це означає, що 25 % робітників даної бригади одержують заробітну плату меншу ніж 1610 грн.

Обчислимо P_{60} : $np = 10 \cdot \dfrac{60}{100} = 6$, тому

$$P_{60} = \frac{x_6 + x_7}{2} = \frac{2850 + 3203}{2} = 3026{,}5 \text{ грн.}$$

Розглянемо випадок, коли вибіркові дані подані інтервальним рядом. Формула для обчислення P_l є аналогом формули (2.11) для знаходження медіани:

$$P_l = x_i + h \cdot \frac{\left(l/100 \right) \cdot n - S_{P_l - 1}}{n_{P_l}}, \tag{2.13}$$

де x_i — початок інтервалу, в який потрапляє P_l; n_{P_l} — частота цього інтервалу; h — його довжина; S_{P_l-1} — нагромаджена частота попереднього інтервалу.

Приклад 2.9. Для даних з табл. 2.4 знайти нижній і верхній квартилі річного прибутку працюючих жінок.

Р о з в ' я з а н н я. Неважко переконатись, що Q_1 потрапляє в інтервал 10 000 — 15 000, адже нагромаджені частоти цього та попереднього інтервалів дорівнюють відповідно 11 846 та 4735, а $0,25n = 0,25 \cdot 29808 = 7452$. Підставимо одержані числові значення у формулу (2.13):

$$Q_1 = 10\,000 + 5000 \cdot \frac{0,25 \cdot 29\,808 - 4735}{7111} \approx \$\,11\,910\,.$$

Аналогічно, Q_3 лежить в інтервалі 20 000 — 25 000, адже нагромаджені частоти цього та попереднього інтервалів дорівнюють відповідно 23132 та 18 447, а $0,75n = 0,75 \cdot 29808 = 22356$. Тому, підставивши відповідні дані у формулу (2.13), дістанемо:

$$Q_3 = 20\,000 + 5000 \cdot \frac{0,75 \cdot 29\,808 - 18447}{4685} \approx \$\,24\,172\,.$$

Отже, річний дохід 25 % опитаних жінок менший ніж \$ 11910, у 75 % респондентів він менший ніж \$ 24172. Крім того, неважко помітити, що різниця між верхнім квартилем і медіаною, знайденою в прикладі 2.7, становить \$ 6856, а різниця між медіаною та нижнім квартилем менша, а саме \$ 6406. Цей факт свідчить про деяку несиметричність у середній частині розподілу. Більш детально характеризацію симетричності статистичного розподілу розглянемо в п. 2.3.

Зазначимо, що квантиль — числова характеристика, що застосовується для досить різноманітних типів даних. Слід підкреслити, що вона може визначатися лише з упорядкованої послідовності вибіркових значень ознаки, тобто з ранжированого варіаційного ряду, або з інтервального ряду, за допомогою апроксимаційної формули (2.13) .

Наведемо приклад використання квантилів у соціально-економічних дослідженнях. У вивченні диференціації доходів населення широко застосовується децильний коефіцієнт

$$K_{\partial} = \frac{D_9}{D_1}\,.$$

Цей показник характеризує рівень диференціації доходів, показуючи як співвідносяться мінімальне значення доходу 10 % найбільш забезпеченого та максимальне значення доходу 10 % найменш забезпеченого населення (у разах).

Крім децильного коефіцієнта часто знаходять фондовий коефіцієнт $К_ф$, який більш точно характеризує рівень диференціації доходів. Цей показник є відношенням середніх значень грошових доходів 10 % найбагатшого населення та 10 % найбіднішого.

2.2. ХАРАКТЕРИСТИКИ РОЗСІЮВАННЯ

Використання для аналізу статистичних даних лише середніх числових характеристик може призвести до помилок. Адже два статистичних вибіркових розподіли, які мають однакові вибіркові середні, можуть суттєво різнитись унаслідок розсіювання значень ознаки, що вивчається. Якщо всі спостережувані значення ознаки мало відрізняються одне від одного, то середня величина буде достатньо показовою характеристикою вибіркової сукупності. У протилежному ж разі вона не дає достатньої інформації про сукупність, тоді її застосування в практичних задачах буде обмеженим. Отже, велику роль в аналізі сукупності, що вивчається, відіграють характеристики розсіювання. Вони певним чином доповнюють середні величини як міри положення, характеризують однорідність статистичної сукупності, межі розсіювання та зв'язок між ознаками. Вивчення варіації допомагає зрозуміти суть того чи іншого явища; дослідження причин її виникнення та впливу на неї окремих факторів дає можливість робити правильні висновки та приймати науково обґрунтовані рішення.

До характеристик розсіювання належать: розмах варіації, середнє лінійне відхилення, дисперсія та середнє квадратичне (стандартне) відхилення, інтерквартильний розмах, коефіцієнт варіації. Зупинимось докладніше на кожній з них.

Найпростішою мірою розсіювання є **розмах варіації *R***, який показує різницю між найбільшим і найменшим значеннями ознаки у вибірці:

$$R = x_{\max} - x_{\min}\,. \tag{2.14}$$

Приклад 2.10. У табл. 2.5 наведено дані про середню заробітну плату в Україні помісячно за перше півріччя 2010 р. в розрахунку на одного штатного працівника[1]. Знайти розмах варіації середньої заробітної плати за даний період.

Таблиця 2.5

Місяць	Січень	Лютий	Березень	Квітень	Травень	Червень
Середня з/п, грн	1916	1955	2109	2107	2201	2373

Р о з в ' я з а н н я. За період із січня до червня 2010 р. включно середня заробітна плата зросла з 1916 грн до 2373 грн. Отже, розмах варіації $R = 2373 - 1916 = 457$ грн.

Перевагою застосування цієї характеристики розсіювання є простота її обчислення. Однак розмах варіації характеризується лише крайніми елементами варіаційного ряду і не відображає розсіювання інших спостережуваних значень ознаки. Тому його слід застосовувати лише для вивчення однорідних статистичних сукупностей.

Більш точною характеристикою є **середній розмах варіації**. Оскільки кожній вибірці відповідає єдине значення розмаху, то його величина змінюватиметься в процесі проведення однакових серій спостережень. Середнє арифметичне значення ряду таких розмахів називається середнім розмахом варіації. Таку характеристику розсіювання часто застосовують в економічних дослідженнях, наприклад, для контролю якості продукції.

Оскільки розмах не дає інформації про варіацію всіх значень ознаки, тому недоцільно обмежуватися під час її вивчення лише цим показником. Для аналізу розсіювання потрібно знайти числову характеристику, яка б відбивала всі коливання спостережуваних значень і була б узагальнювальною. Найпростішою з таких характеристик є середнє лінійне відхилення.

Кожне спостережуване значення ознаки відхиляється від вибіркової середньої $\bar{x}_в$ у той чи інший бік. Різниця $x_i - \bar{x}_в$ характеризує внесок *i*-ї варіанти у варіацію ознаки, що вивчається. Оскільки (за властивістю 1 вибіркової середньої) сума всіх відхилень значень ознаки від $\bar{x}_в$ дорівнює нулю, доцільно як характеристику варіації взяти суму абсолютних величин (модулів) таких відхилень.

Отже, **середнє лінійне відхилення**, у разі коли всі значення ознаки різні, знаходять за формулою

$$\bar{l} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}_в|, \tag{2.15}$$

де n — обсяг сукупності.

Для вибіркових даних, згрупованих у статистичний дискретний ряд,

$$\bar{l} = \frac{1}{n}\sum_{i=1}^{k}|x_i - \bar{x}_в| \cdot n_i, \tag{2.16}$$

де сума частот ряду $\sum_{i=}^{k} n_i = n$.

У разі інтервального ряду у формулі (2.16) за варіанти x_i слід узяти середини відповідних інтервалів.

Приклад 2.11. Для даних із прикладу 2.10 (табл. 2.5) знайдемо середнє лінійне відхилення.

Розв'язання. Спочатку обчислимо вибіркову середню:

$$\bar{x}_в = \frac{1}{6}(1916 + 1955 + 2109 + 2107 + 2201 + 2373) = 2110 \text{ грн.}$$

Отже, середня заробітна плата за перше півріччя 2010 р. становила 2110 грн. Знайдемо середнє лінійне відхилення:

$$\bar{l} = \frac{1}{6}(194 + 155 + 1 + 3 + 91 + 263) = 118 \text{ грн.}$$

Тобто в середньому щомісячна заробітна плата за даний період відхилялася від середньої зарплати за півріччя на 118 грн.

---

[1] Держкомстат України [Електронний ресурс] : 1998—2011. Режим доступу : http://www.ukrstat.gov.ua

Перевагою цієї характеристики розсіювання є простота її розрахунку та інтерпретації. Але попри це на практиці вона застосовується рідко, адже використання модуля для розрахунку цієї характеристики не дозволяє поставити її у відповідність одному з відомих імовірнісних законів розподілу. В економічних дослідженнях середнє лінійне відхилення використовується лише тоді, коли сума відхилень значень ознаки від середньої без урахування знаків має економічний зміст, наприклад, для аналізу ритмічності виробництва, обороту зовнішньої торгівлі тощо.

Найчастіше для характеристики варіації спостережуваних значень ознаки вибірки щодо вибіркової середньої використовуються вибіркова дисперсія та середнє квадратичне відхилення.

Вибірковою дисперсією ($D_в$) називається середнє арифметичне квадратів відхилень значень вибірки від вибіркової середньої $\bar{x}_в$.

Якщо всі значення ознаки вибірки $x_1, x_2, \ldots, x_n$ різні, то

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2. \tag{2.17}$$

У разі, коли значення ознаки $x_1, x_2, \ldots, x_k$ мають частоти відповідно $n_1, n_2, \ldots, n_k$, причому $\sum_{i=1}^{k} n_i = n$, то

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 n_i . \tag{2.18}$$

Для вибіркових даних, поданих інтервальним статистичним рядом, вибіркову дисперсію знаходять за формулою

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i^* - \bar{x}_в\right)^2 n_i , \tag{2.19}$$

де $x_1^*, x_2^*, \ldots, x_k^*$ — середини відповідних інтервалів; $n_1, n_2, \ldots, n_k$ — їхні частоти, причому $n = \sum_{i=1}^{k} n_i$.

Як уже зазначалось, для розрахунку вибіркової середньої та дисперсії інтервального розподілу вибірки спостережувані значення ознаки замінюються серединами відповідних інтервалів. Природно, це може призвести до певних неточностей. В. Ф. Шеппард установив, що в середньому різниця між вибірковими середніми, знайденими за даними, згрупованими в інтервальний ряд, та початковими даними, дорівнює нулю. Для дисперсії ж така різниця в середньому дорівнює $1/12$ квадрата довжини інтервалу (у разі, коли всі інтервали мають однакову довжину). Тобто щоб скоригувати дисперсію, знайдену за інтервальним рядом, від неї слід відняти поправку Шеппарда $\frac{1}{12}h^2$, де h — довжина інтервалу. Варто наголосити, що дана поправка має застосовуватись лише для неперервного розподілу, близького до нормального, побудованого за достатньо великою кількістю даних ($n > 500$). В інших випадках до використання поправки Шеппарда в обчисленнях вибіркової дисперсії слід ставитися з обережністю, оскільки замість уточнення результату можна дістати протилежне.

Зауважимо, що вибіркова дисперсія, характеризуючи міру розсіювання випадкової величини, сама є випадковою величиною і має свої числові характеристики: середню та дисперсію. У §3 і §4 буде показано, що вибіркова дисперсія за виконання деяких умов є оцінкою **генеральної дисперсії**, тобто дисперсії ознаки в генеральній сукупності, яку далі будемо позначати $D_г$.

Оскільки одиниці, в яких вимірюється дисперсія, є квадратами одиниць спостережуваної випадкової величини (наприклад, грн^2, м^2, г^2), то за показник розсіювання доцільно брати **вибіркове середнє квадратичне відхилення** ($\sigma_в$), що дорівнює квадратному кореню з вибіркової дисперсії, тобто

$$\sigma_в = \sqrt{D_в} . \tag{2.20}$$

Середнє квадратичне (стандартне) відхилення є узвичаєною мірою варіації ознаки, для його обчислення враховуються всі значення відхилень варіант від вибіркової середньої. Розмірність середньоквадратичного відхилення збігається з розмірністю спостережуваної ознаки, тому воно є зручним з погляду інтерпретації.

Приклад 2.12. Інспектор відібрав 30 зразків пакетів вівсянки швидкого приготування для контролю ваги. Результати зважування (г):

28,7; 29,8; 30,6; 31,6; 29,5; 30,1; 30; 31,9; 31,7; 29,4; 29,3; 30,1; 30,4; 29,4; 29,8; 30,9; 30; 30,2; 29,8; 30,7; 30,5; 29,9; 30,6; 30,4; 29,8; 30,8; 30,1; 30,4; 30,7; 30.

Знайти вибіркову дисперсію та середньоквадратичне відхилення спостережуваної ознаки, згрупувавши дані в інтервальний ряд (з кроком $h = 0,5$).

Р о з в ' я з а н н я. Побудуємо інтервальний ряд, і проміжні дані для обчислення вибіркової середньої та дисперсії подамо в табл. 2.6.

Таблиця 2.6

Вага одного пакета, г	Кількість пакетів, шт. (n_i)	x_i^*	$x_i^* \cdot n_i$	$x_i^* - \bar{x}_в$	$(x_i^* - \bar{x}_в)^2$	$(x_i^* - \bar{x}_в)^2 \cdot n_i$
до 29	1	28,75	28,75	–1,48	2,1904	2,1904
29 — 29,5	3	29,25	87,75	–0,98	0,9604	2,8812
29,5 — 30	6	29,75	178,5	–0,48	0,2304	1,3824
30 — 30,5	10	30,25	302,5	0,02	0,0004	0,004
30,5 — 31	7	30,75	215,25	0,52	0,2704	1,8928
31 — 31,5	2	31,25	62,5	1,02	1,0404	2,0808
31,5 і більше	1	31,75	31,75	1,52	2,3104	2,3104
Разом:	30	—	907	—	—	12,742

Знайдемо середню вибіркову, дисперсію та середньоквадратичне відхилення вибірки за інтервальним рядом:

$$\bar{x}_в = \frac{1}{30}\sum_{i=1}^{7} x_i^* n_i = \frac{907}{30} = 30,23 \text{ г},$$

$$D_в = \frac{1}{30}\sum_{i=1}^{7}\left(x_i^* - \bar{x}_в\right)\cdot n_i = \frac{12,742}{30} = 0,425 \text{ г}^2 ,$$

$$\sigma_в = \sqrt{D_в} = \sqrt{0,425 = 0,652} \text{ г}.$$

У розв'язанні практичних задач також часто використовується величина

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot n_i ,$$

або

$$s^2 = \frac{1}{n-1} D_в ,$$

яка називається **виправленою вибірковою дисперсією** (див. п. 3.2).

Відповідно, **виправленим середнім квадратичним відхиленням** називається величина

$$s = \sqrt{s^2} .$$

Тож стандартне відхилення, як і середнє лінійне відхилення, показує, на скільки в середньому відхиляються значення ознаки від середньої величини у вибірці. Але слід мати на увазі, що середнє лінійне відхилення буде мінімальним, якщо воно розраховане відносно медіани (Ме), тобто функція

$$\bar{l}(t) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - t\right|$$

досягає мінімуму при t = Ме (див. властивості медіани), тоді як середньоквадратичне відхилення є мінімальним у разі розрахунку його від середньої вибіркової $\bar{x}_в$ (див. наведені далі властивості дисперсії). Установлено, що середнє лінійне відхилення $\bar{l}$ завжди менше, ніж стандартне, зокрема, для симетричних чи близьких до симетричного розподілів $\sigma_в \approx 1625\ \bar{l}$. За нормального закону розподілу, який є симетричним відносно середньої величини, $\sigma_в : \bar{l} \approx 1{,}2$. Збільшення значення відношення ($\sigma_в : \bar{l}$) свідчить про наявність у сукупності варіант, відхилення яких від середньої мають екстремальні значення. Тобто дане відношення може служити певним індикатором «засміченості» сукупності неоднорідними елементами.

Перевагою використання середньоквадратичного відхилення як міри розсіювання є те, що воно входить до більшості теорем теорії ймовірностей, на яких побудовано основні методи математичної статистики.

Продемонструємо на даних з прикладу 2.12, які висновки можна зробити про концентрацію вибіркових значень ознаки навколо вибіркової середньої за величиною $\sigma_в$, використовуючи теорему Чебишова.

Отже, середня маса відібраних для контролю зразків пакетів із вівсяними пластівцями $\bar{x}_в = 30{,}23$ г, середньоквадратичне відхилення $\sigma_в$ = 0,652 г (див. приклад 2.12). Тоді

$$\bar{x}_в - \sigma_в = 29{,}578,\ \bar{x}_в + \sigma_в = 30{,}882;$$
$$\bar{x}_в - 2\sigma_в = 28{,}926,\ \bar{x}_в + 2\sigma_в = 31{,}534;$$
$$\bar{x}_в - 3\sigma_в = 28{,}274,\ \bar{x}_в + 3\sigma_в = 32{,}186.$$

Проаналізувавши вибіркові дані, можна побачити, що 21 значення з 30 відібраних (70 %) міститься на інтервалі ($\bar{x}_в - \sigma_в; \bar{x}_в + \sigma_в$); 26 значень з 30 (87 %) — на інтервалі ($\bar{x}_в - 2\sigma_в; \bar{x}_в + 2\sigma_в$); 30 із 30 значень (100 %) — на інтервалі ($\bar{x}_в - 3\sigma_в; \bar{x}_в + 3\sigma_в$).

Що ж можна сказати загалом про будь-яку вибіркову сукупність? Відповідно до теореми Чебишова можемо стверджувати, що незалежно від закону розподілу випадкової величини, що вивчається, як мінімум 75 % значень ознаки потрапляють на інтервал ($\bar{x}_в - 2\sigma_в; \bar{x}_в + 2\sigma_в$), 89 % значень — на інтервал ($\bar{x}_в - 3\sigma_в; \bar{x}_в + 3\sigma_в$). Загалом принаймні $(1 - 1/k^2)\cdot 100\,\%$ вибіркових даних містяться між $\bar{x}_в - k\sigma_в$ та $\bar{x}_в - k\sigma_в$. Отже, якщо стандартне відхилення $\sigma_в$ невелике, то $\bar{x}_в$ добре представляє даний ряд, і в такому разі кажуть, що він є достатньо однорідним. І чим більша величина $\sigma_в$, тим ширше розсіяні значення ознаки відносно вибіркової середньої.

У разі, якщо закон розподілу спостережуваної ознаки близький до нормального, можна стверджувати, що на інтервалі ($\bar{x}_в - \sigma_в;\ \bar{x}_в + \sigma_в$) містяться близько 68 % спостережуваних значень, на інтервалі ($\bar{x}_в - 2\sigma_в;\ \bar{x}_в + 2\sigma_в$); — близько 95 %, і інтервал ($\bar{x}_в - 3\sigma_в;\ \bar{x}_в + 3\sigma_в$); охоплює майже всі 100 % значень.

Приклад 2.13. Для випадково відібраних 200 автомобільних страхових виплат деякої страхової компанії за певний період маємо: $\bar{x}_в = 5923$ грн, $\sigma_в = 1280$ грн. Який висновок можна зробити про розташування вибіркових значень ознаки навколо вибіркової середньої? Що можна стверджувати про значення усіх 200 спостережуваних виплат за припущення, що дана ознака розподілена за нормальним законом?

Р о з в ’ я з а н н я. Незалежно від закону розподілу розміру страхових виплат, спираючись на нерівність Чебишова, можна стверджувати, що 89 % спостережуваних значень (178 із 200) лежать у межах від $\bar{x}_в - 3\sigma_в = 2083$ грн до $\bar{x}_в + 3\sigma_в = 9763$ грн. У разі, ж коли закон розподілу даної ознаки нормальний, майже всі 200 відібраних значень ознаки лежать у знайдених межах — від 2083 грн до 9763 грн.

Вибіркова дисперсія як міра варіації сама по собі є незручною через свою «квадратну» розмірність. Але вона широко застосовується в економічних дослідженнях, адже розчленування її на складові дозволяє оцінити вплив різних факторів на величину варіації ознаки. Наведемо основні **властивості вибіркової дисперсії**.

1. Якщо всі варіанти збільшити (зменшити) в m раз, то дисперсія збільшиться (зменшиться) в m^2 разів.

2. За збільшення (зменшення) усіх варіант на ту саму величину дисперсія залишиться незмінною.

3. За збільшення (зменшенні) усіх частот в однакову кількість раз значення дисперсії не зміниться.

(Властивості 1—3 пропонуємо читачеві довести самостійно).

4. Дисперсія відносно вибіркової середньої дорівнює різниці дисперсії відносно довільної сталої і квадрата різниці середньої вибіркової та цієї сталої, тобто

$$D_{\text{в}} = D_{\text{вC}} - (\overline{x}_{\text{в}} - C)^2$$

$$\text{де } D_{\text{вC}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - C)^2 .$$

Доведемо це. Справді,

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_{\text{в}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - C) - (\overline{x}_{\text{в}} - C)\right]^2 =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - C)^2 - 2(x_{\text{в}} - C)\sum_{i=1}^{n}(x_i - C) + n(\overline{x}_{\text{в}} - C)^2\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - C)^2 - 2\left(\overline{x}_{\text{в}} - C\right)\left(\frac{1}{n}\sum_{i=1}^{n}x_i - C\right) + \left(\overline{x}_{\text{в}} - C\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - C)^2 - 2\left(\overline{x}_{\text{в}} - C\right)^2 + \left(\overline{x}_{\text{в}} - C\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - C)^2 - (\overline{x}_{\text{в}} - C)^2.$$

Отже,

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - C\right)^2 - \left(\overline{x}_{\text{в}} - C\right)^2 . \tag{2.21}$$

Урахувавши позначення, уведені раніше, рівність (2.21) можна записати у вигляді

$$D_{\text{в}} = D_{\text{вC}} - \left(\overline{x}_{\text{в}} - C\right)^2 .$$

Звідси випливає, що $D_{\text{вC}} = D_{\text{в}} + \left(\overline{x}_{\text{в}} - C\right)^2$. Це означає, що для будь-якої сталої $D_{\text{вC}} \geq D_{\text{в}}$. Отже, вибіркова дисперсія (дисперсія відносно $\overline{x}_{\text{в}}$) має властивість мінімальності: вона менша, ніж дисперсія, обчислена відносно будь-якої іншої сталої $C \in R$.

5. Вибіркова дисперсія дорівнює різниці середнього арифметичного квадратів значень ознаки та квадрата середньої вибіркової, тобто

$$D_{\text{в}} = \overline{x_{\text{в}}^2} - \left(\overline{x}_{\text{в}}\right)^2 , \tag{2.22}$$

де

а) $\overline{x_{\text{в}}^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2$, якщо всі значення ознаки $x_1,\ x_2,\ \ldots,\ x_n$ різні;

б) $\overline{x_{\text{в}}^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 n_i$, якщо значення ознаки $x_1,\ x_2,\ \ldots,\ x_k$ мають частоти відповідно $n_1,\ n_2,\ \ldots,\ n_k$, причому $\sum_{i=1}^{k}n_i = n$.

Справді, узявши у (2.21) $C = 0$, дістанемо рівність (2.22). Тобто дістали ще одну формулу для обчислення вибіркової дисперсії.

Зауважимо, що рівність дисперсій, знайдених за формулами (2.17) [(або (2.18)] та (2.22), виконується лише в разі, коли значення середньої вибіркової, що використовувалось у розрахунках, було точним. Якщо ж значення середньої було заокруглене, слід у розрахунку дисперсії надати перевагу формулі (2.17). Річ у тім, що розрахунок за формулою (2.17) призведе до похибки того самого порядку, що й похибка, допущена за заокруглення середньої. У разі ж застосування формули (2.22) похибка дисперсії буде на порядки більша, ніж похибка середньої вибіркової.

6. Правило додавання дисперсій. Варіація ознаки залежить від різних факторів, які можна поділити на дві групи — систематичні та випадкові. Для практичних і наукових потреб необхідно оцінити роль кожної групи факторів у формуванні варіації. При цьому загальну варіацію досліджува-

ної ознаки потрібно розкласти на дві компоненти: систематичну та випадкову. Це можна зробити на основі аналітичного групування, при цьому групувальна ознака розглядається як систематичний фактор. У такому разі для вивчення варіації використовуються три види дисперсії: загальна ($D_{\text{в}}$), внутрішньогрупова ($D_{\text{вн}}$) та міжгрупова ($D_{\text{м}}$). Зазначимо, що правило додавання дисперсій, викладене далі для вибірки, є справедливим і для генеральної сукупності.

Нехай усі значення ознаки розглядуваної сукупності розбиті на s груп, що не перетинаються. Якщо кожну з таких груп розглядати як окрему сукупність, можна знайти групову середню (див. властивість 5 вибіркової середньої) та групову дисперсію відносно такої середньої:

$$\overline{D_{\text{в}j}} = \frac{1}{N_j}\sum_{i=1}^{k_j}(x_i - \overline{x_{\text{в}j}})^2 n_i\,, \tag{2.23}$$

де j — номер групи, $\overline{x_{\text{в}j}}$ — її вибіркова середня, k_j — кількість варіант у групі j, n_i — частота значення x_i, $N_j = \sum_{i=1}^{k_j} n_i$ — обсяг групи з номером j.

Групова дисперсія відображає випадкову варіацію, тобто ту частину варіації, яка не залежить від фактора, на основі якого було проведено групування. Оскільки таку дисперсію знаходять для кожної групи окремо, то для того щоб увести поняття **внутрішньогрупової дисперсії** в цілому за сукупністю, знаходять її середнє значення, зважене за обсягами груп:

$$D_{\text{вн}} = \frac{1}{n}\sum_{j=1}^{s} D_{\text{в}j} N_j\,, \tag{2.24}$$

де $n = \sum_{j=1}^{s} N_j$ — обсяг сукупності.

Міжгруповою дисперсією за означенням називається дисперсія групових середніх відносно загальної вибіркової середньої, тобто

$$D_{\text{м}} = \frac{1}{n}\sum_{j=1}^{s}(\overline{x_{\text{в}j}} - \overline{x_{\text{в}}})^2 N_j\,. \tag{2.25}$$

Міжгрупова дисперсія характеризує систематичну варіацію, яка обумовлюється впливом групувальної ознаки.

Загальна ж дисперсія є мірою варіації ознаки всієї сукупності під впливом усіх факторів, що її обумовлюють. Вона розраховується як вибіркова дисперсія даної сукупності за однією з формул — (2.17), (2.18) чи (2.19).

Тепер можна сформулювати правило додавання дисперсій. Отже, загальна дисперсія дорівнює сумі внутрішньогрупової та міжгрупової дисперсій, тобто

$$D_{\text{в}} = D_{\text{вн}} + D_{\text{м}}\,. \tag{2.26}$$

Доведемо це. Зазначимо, що, використовуючи властивість 5, формули (2.23) та (2.25) для обчислення групової та міжгрупової дисперсій можна записати у вигляді:

$$D_{\text{в}j} = \frac{1}{N_j}\sum_{i=1}^{k_j} x_i^2 n_i - \overline{x_{\text{в}j}}^2\,, \tag{2.27}$$

$$D_{\text{м}} = \frac{1}{n}\sum_{j=1}^{s} \overline{x_{\text{в}j}}^2 N_j - \overline{x_{\text{в}}}^2\,. \tag{2.28}$$

Із (2.27) для кожної групи маємо:

$$\sum_{i=1}^{k_j} x_i^2 n_i = N_j D_{\text{в}j} + N_j \overline{x_{\text{в}j}}^2\,.$$

Додавши почленно s таких рівностей, складених для кожної з груп, дістанемо:

$$\sum_{i=1}^{k} x_i^2 n_i = \sum_{j=1}^{s} D_{\text{в}j} N_j + \sum_{j=1}^{s} \overline{x_{\text{в}j}}^2 N_j\,,$$

де k — кількість варіант у розглядуваній сукупності.

Поділимо обидві частини одержаної рівності на n та віднімемо від кожної з них $\overline{x_{\text{в}}}^2$:

$$\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - \overline{x_{\text{в}}}^2 = \frac{1}{n}\sum_{j=1}^{s} D_{\text{в}j} N_j + \frac{1}{n}\sum_{j=1}^{s} \overline{x_{\text{в}j}}^2 N_j - \overline{x_{\text{в}}}^2,$$

або, з використанням формул (2.22), (2.24), (2.28),

$$D_{\text{в}} = D_{\text{вн}} + D_{\text{м}}.$$

Тож, користуючись правилом додавання дисперсій, можна розкласти загальну варіацію результативної ознаки на систематичну та випадкову. При цьому мірою систематичної варіації є міжгрупова дисперсія $D_{\text{м}}$, а випадкової — внутрішньогрупова дисперсія $D_{\text{вн}}$. Крім того, скориставшись формулою (2.26), за двома відомими дисперсіями можна завжди знайти третю, що значно спрощує розрахунки.

Приклад 2.14. Менеджер повинен дослідити, наскільки варіація обсягів продажу протягом певного часу залежала від продавця. Було проведено спостереження щоденного продажу магазину одягу, де працюють три продавці. Одержані результати (x_i, $1 \le i \le 24$, тис. грн) подано в табл. 2.7.

Таблиця 2.7

Продавець 1	1,7	1,9	1,6	2,0	1,3	2,1	2,4	1,6
Продавець 2	2,0	1,8	2,3	2,5	3,0	2,1	2,4	1,9
Продавець 3	1,5	1,1	1,7	1,2	0,9	1,4	1,6	1,4

Р о з в ' я з а н н я. Отже, групувальною ознакою в даному разі є кваліфікація продавця, результативною — щоденні обсяги продажу. Знайдемо три види дисперсій: загальну, внутрішньогрупову та міжгрупову.

Обчислимо середній обсяг продажу та групову дисперсію [див. формулу (2.23)] для першого продавця:

$$\overline{x_{\text{в}1}} = \frac{1}{8}(1{,}7 + 1{,}9 + 1{,}6 + 2{,}0 + 1{,}3 + 2{,}1 + 2{,}4 + 1{,}6) = 1{,}825 \text{ тис. грн},$$

$$D_{\text{в}1} = \frac{1}{8}(1{,}7^2 + 1{,}9^2 + 1{,}6^2 + 2{,}0^2 + 1{,}3^2 + 2{,}1^2 + 2{,}4^2 + 1{,}6^2) - 1{,}825^2 = 0{,}104.$$

Для другого і третього відповідно:

$$\overline{x_{\text{в}2}} = \frac{1}{8}(2{,}0 + 1{,}8 + 2{,}3 + 2{,}5 + 3{,}0 + 2{,}1 + 2{,}4 + 1{,}9) = 2{,}25 \text{ тис. грн},$$

$$D_{\text{в}2} = \frac{1}{8}(2{,}0^2 + 1{,}8^2 + 2{,}3^2 + 2{,}5^2 + 3{,}0^2 + 2{,}1^2 + 2{,}4^2 + 1{,}9^2) - 2{,}25^2 = 0{,}1325;$$

$$\overline{x_{\text{в}3}} = \frac{1}{8}(1{,}5 + 1{,}1 + 1{,}7 + 1{,}2 + 0{,}9 + 1{,}4 + 1{,}6 + 1{,}4) = 1{,}35 \text{ тис. грн},$$

$$D_{\text{в}3} = \frac{1}{8}(1{,}5^2 + 1{,}1^2 + 1{,}7^2 + 1{,}2^2 + 0{,}9^2 + 1{,}4^2 + 1{,}6^2 + 1{,}4^2) - 1{,}35^2 = 0{,}0625.$$

Групові дисперсії показують варіації обсягів продажу кожним продавцем, обумовлені всіма можливими факторами, крім різниці якості їхньої роботи.

Внутрішньогрупова дисперсія:

$$D_{\text{вн}} = \frac{8 \cdot 0{,}104 + 8 \cdot 0{,}1325 + 8 \cdot 0{,}0625}{24} = 0{,}1.$$

Для того щоб знайти міжгрупову дисперсію, обчислимо загальну середню вибіркову (див. властивість 5 вибіркової середньої):

$$\overline{x_{\text{в}}} = \frac{1}{3}\left(1{,}825 + 2{,}25 + 1{,}35\right) = 1{,}8083 \text{ тис. грн}.$$

Отже,

$$D_{\text{м}} = \frac{1}{3}\left(1{,}825^2 + 2{,}25^2 + 1{,}35^2\right) - 1{,}8083^2 = 0{,}135 .$$

Міжгрупова дисперсія показує варіацію групових середніх, обумовлену відмінностями в роботі продавців.

Знайдемо загальну дисперсію вибірки:

$$D_{\text{в}} = \frac{1}{24}\sum_{i=1}^{24} x_i^2 - \bar{x}_{\text{в}}^2 = \frac{1}{24}(1{,}7^2 + 1{,}9^2 + 1{,}6^2 + 2{,}0^2 + 1{,}3^2 + 2{,}1^2 + 2{,}4^2 + 1{,}6^2 +$$

$$+2{,}0^2 + 1{,}8^2 + 2{,}3^2 + 2{,}5^2 + 3{,}0^2 + 2{,}1^2 + 2{,}4^2 + 1{,}9^2 + 1{,}5^2 + 1{,}1^2 + 1{,}7^2 + 1{,}2^2 + 0{,}9^2 +$$

$$+1{,}4^2 + 1{,}6^2 + 1{,}4^2) - 1{,}8083^2 = \frac{84{,}12}{24} - 1{,}8083^2 = 0{,}235 .$$

Загальна дисперсія показує сумарний вплив усіх можливих факторів на варіацію обсягів щоденного продажу.

Як бачимо, правило додавання дисперсій виконується:

$$D_{\text{в}} = D_{\text{вн}} + D_{\text{м}} = 0{,}1 + 0{,}135 = 0{,}235 .$$

Очевидно, що чим більше відношення міжгрупової дисперсії до загальної, тим сильніший вплив на варіацію результативної ознаки групувального фактора. Тому в статистиці широко використовується **емпіричний коефіцієнт детермінації**, за допомогою якого обчислюють міру впливу досліджуваного факторного показника на результативний діленням міжгрупової дисперсії на загальну:

$$\eta^2 = \frac{D_{\text{м}}}{\mathrm{D}_{\text{в}}} . \tag{2.29}$$

У наведеному прикладі цей коефіцієнт становить $\eta^2 = \frac{0{,}135}{0{,}235} = 0{,}574$. Це означає, що на 57,4 % варіація обсягів продажів залежала від кваліфікації продавця, і на 42,6 % — від інших факторів.

Тісноту ж зв'язку між групувальною та результативною ознаками характеризує **емпіричне кореляційне відношення**:

$$\eta = \sqrt{\frac{D_{\text{м}}}{\mathrm{D}_{\text{в}}}} .$$

Емпіричне кореляційне відношення, як і емпіричний коефіцієнт детермінації, може набирати значень від 0 до 1. Причому чим ближче воно до 1, тим сильніший зв'язок між ознаками, що вивчаються.

Однак усе викладене стосувалося кількісних ознак. Але поряд із ними часто на практиці доводиться мати справу з ознаками, які притаманні одним одиницям, що розглядається, і не притаманні іншим (наприклад, виріб якісний або бракований, стать чоловіча або жіноча). Такі ознаки називають **альтернативними**. Наявність такої ознаки в елемента сукупності позначимо 1, її відсутність — 0. Нехай p — частка одиниць сукупності з даною ознакою, $q = 1 - p$ — частка одиниць без неї. Знайдемо **вибіркову середню**, **вибіркову дисперсію** та **вибіркове середнє квадратичне відхилення альтернативної ознаки**:

$$\overline{x_{\text{в}}} = 1 \cdot p + 0 \cdot q = p ; \tag{2.30}$$

$$D_{\text{в}} = (1-q)^2 p + (0-p)^2 q = q^2 p + qp^2 = pq(q+p) = pq , \tag{2.31}$$

$$\sigma_{\text{в}} = \sqrt{pq} . \tag{2.32}$$

Отже, дисперсія альтернативної ознаки дорівнює добутку частки одиниць з даною ознакою і частки одиниць без неї. Максимальним значенням дисперсії альтернативної ознаки є 0,25; воно досягається при $p=0,5$.

Приклад 2.15. Нехай з відібраних 100 лампочок 2 виявилися бракованими. Знайти середньоквадратичне відхилення частки браку.

Розв'язання. У даному разі $p=0,98$, $q=0,02$. Тоді за формулою (2.32)

$$\sigma=\sqrt{pq}=\sqrt{0,98\cdot 0,02}=\sqrt{0,0196}=0,14\,.$$

Своєрідною мірою варіації ознаки є також **інтерквартильний розмах**:

$$\Delta Q=Q_3-Q_1\,. \tag{2.33}$$

Зазначимо, що точки, які відповідають значенням Q_1, Q_2, Q_3, ділять у загальному випадку відрізок $[x_{\min};x_{\max}]$ на чотири не рівних за довжиною частини. Між Q_1 та Q_2 міститься половина всіх спостережуваних значень ознаки. Тож, чим більш щільно зосереджені значення ознаки біля своєї середньої (у даному разі медіани), тим меншим буде ΔQ.

Інтерквартильний розмах, як і розмах варіації R, є нечутливим до випадкових коливань крайніх значень ознаки. Цю міру варіації, а також медіану як міру центральної тенденції використовують, коли середню вибіркову та середнє квадратичне відхилення знаходити недоцільно з погляду природи розглядуваної ознаки або неможливо. Медіану та інтерквартильний розмах як числові характеристики ознаки широко застосовують, наприклад, у медицині для розподілів, відмінних від нормального (більшість розподілів медичних і біологічних параметрів є саме такими).

Іноді як міру розсіювання знаходять **інтердецильний розмах**:

$$\Delta D=D_9-D_1\,. \tag{2.34}$$

Інтердецильний розмах характеризує розсіювання 80 % даних. Він також є більш показовим, ніж розмах варіації R, адже майже не залежить від екстремальних значень ознаки.

Приклад 2.16. Знайти ΔQ для даних з прикладу 2.12, одержаних у результаті зважування 30 пакетів вівсянки для контролю ваги.

Р о з в ' я з а н н я. За одержаними даними (маса одного пакета, г) побудуємо варіаційний ряд:

28,7; 29,3; 29,4; 29,4; 29,5; 29,8; 29,8; 29,8; 29,8; 29,9; 30; 30; 30; 30,1; 30,1; 30,1; 30,2; 30,4; 30,4; 30,4; 30,5; 30,6; 30,6; 30,7; 30,7; 30,8; 30,9; 31,6; 31,7; 31,9.

Знайдемо перший і третій квартилі.

$Q_1=P_{25}$, тому $np=30\cdot\dfrac{25}{100}=7,5$. Отже, нижній квартиль збігається з восьмим членом варіаційного ряду $Q_1=x_8=29,8$ г.

$Q_3=P_{75}$, тому $np=30\cdot\dfrac{75}{100}=22,5$. Тобто верхній квартиль $Q_3=x_{23}=30,6$ г.

Звідси інтерквартильний розмах:

$$\Delta Q=30,6-29,8=0,8 \text{ г.}$$

Інтерпретація одержаного значення така: якщо відкинути 25 % найбільших і 25 % найменших значень, то розмах 50 % значень розглядуваної ознаки (маса одиниці продукції, г), що залишилися, дорівнює 0,8 г.

Отже, підсумовуючи, можна сказати, що чим менша величина, що характеризує варіацію (середнє лінійне відхилення, середньоквадратичне відхилення тощо), тим більш однорідною є розглядувана сукупність і тим більш типовим представником цієї сукупності буде середня величина. Однак у статистиці часто виникає потреба в порівнянні варіацій кількох ознак, наприклад, вартості та прибутку, стажу роботи і розміру заробітної плати, кваліфікації працівника та продуктивності праці тощо. Оскільки середньоквадратичне відхилення (чи середнє лінійне відхилення) вимірюється в тих самих одиницях, що й спостережувана ознака, то воно є незручною для порівняння характеристикою. Наприклад, знайшовши вибіркові середні квадратичні відхилення стажу роботи та заробітної плати працівників, неможливо оцінити, яка з ознак має більшу варіацію, адже перша вимірюється в роках, друга — у гривнях.

Для порівняння варіації різних ознак, а також однієї ознаки, але в різних сукупностях з різними середніми, використовується відносна міра розсіювання — **коефіцієнт варіації**

$$V = \frac{\sigma_{\text{в}}}{\bar{x}_{\text{в}}} \cdot 100\,\%\,. \tag{2.35}$$

Зазначимо, що коефіцієнт варіації використовується не тільки для порівняння варіації ознак, а й для характеризації однорідності сукупності. Сукупність уважається однорідною, якщо її коефіцієнт варіації не перевищує 33 %.

Приклад 2.17. За даними 20 фермерських господарств регіону з однаковими посівними площами врожайність зернових і зернобобових культур $(x_i$, $1 \le i \le 20$, ц з 1 га) та картоплі $(y_i$, $1 \le i \le 20$, ц з 1 га) подана в табл. 2.8. Знайти коефіцієнти варіації для обох вибіркових сукупностей, порівняти одержані результати.

Таблиця 2.8

№ госп-ва	1	2	3	4	5	6	7	8	9	10
x_i	25,1	24,8	22	20,2	28,1	26,5	29,8	27,7	21,4	30,5
y_i	136,2	126,6	154,2	144	125,2	148,8	148,2	133,1	141,2	154,7

№ госп-ва	11	12	13	14	15	16	17	18	19	20
x_i	25,1	29,8	21	20,2	28,1	26,5	25,8	30,7	22,4	31,5
y_i	162,8	147,1	108,9	138	151,2	160,4	149,6	141	150,4	156,1

Р о з в ’ я з а н н я. Знайдемо числові характеристики (вибіркову середню та вибіркову дисперсію) для кожної сукупності:

$$\overline{x}_{\text{в}} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{1}{20}(25{,}1 + 24{,}8 + 22 + 20{,}2 + 28{,}1 + 26{,}5 + 29{,}8 + 27{,}7 + 21{,}4 +$$

$$+30{,}5 + 25{,}1 + 29{,}8 + 21 + 20{,}2 + 28{,}1 + 26{,}5 + 25{,}8 + 30{,}7 + 22{,}4 + 31{,}5) = 25{,}86 \text{ ц/га},$$

$$D_{\text{в}x} = \frac{1}{20}\sum_{i=1}^{20} x_i^2 - \bar{x}_{в}^2 = \frac{1}{20}(25{,}1^2 + 24{,}8^2 + 22^2 + 20{,}2^2 + 28{,}1^2 + 26{,}5^2 + 29{,}8^2 + 27{,}7^2 + 21{,}4^2 +$$

$$30{,}5^2 + 25{,}1^2 + 29{,}8^2 + 21^2 + 20{,}2^2 + 28{,}1^2 + 26{,}5^2 + 25{,}8^2 + 30{,}7^2 + ,$$

$$+22{,}4^2 + 31{,}5^2) - 25{,}86^2 = 12{,}94 \ (\text{ц/га})^2;$$

$$\sigma_{\text{в}x} = \sqrt{12{,}94} \approx 3{,}6 \ \text{ ц/га}.$$

$$\overline{y}_{\text{в}} = \frac{1}{20}\sum_{i=1}^{20} y_i = \frac{1}{20}(136{,}2 + 126{,}2 + 154{,}2 + 144 + 125{,}2 + 148{,}8 + 148{,}2 + 133{,}1 +$$

$$+141{,}2 + 154{,}7 + 162{,}8 + 147{,}1 + 108{,}9 + 138 + 151{,}2 + 160{,}4 + 149{,}6 + 141 +$$

$$+150{,}4 + 156{,}1) = 143{,}885 \ \text{ ц/га},$$

$$D_{\text{в}} = \frac{1}{20}\sum_{i=1}^{20} y_i^2 - \overline{y}_{\text{в}}^{\,2} = \frac{1}{20}(136{,}2^2 + 126{,}2^2 + 154{,}2^2 + 144^2 + 125{,}2^2 + 148{,}8^2 +$$

$$+148{,}2^2 + 133{,}1^2 + 141{,}2^2 + 154{,}7^2 + 162{,}8^2 + 147{,}1^2 + 108{,}9^2 + 138^2 + 151{,}2^2 +$$

$$+160{,}4^2 + 149{,}6^2 + 141^2 + 150{,}4^2 + 156{,}1^2) - 143{,}885^2 = 164{,}11 (\text{ц/га})^2,$$

$$\sigma_{\text{в}y} = \sqrt{164{,}11} \approx 12{,}81 \ \text{ ц/га}.$$

Отже, коефіцієнти варіації для обох сукупностей:

$$V_x = \frac{3{,}6}{25{,}86} \cdot 100\,\% \approx 14\,\%\,;$$

$$V_y = \frac{12{,}81}{143{,}885} \cdot 100\,\% \approx 9\,\%.$$

Це означає, що обидві сукупності достатньо однорідні, причому за врожайністю картоплі сукупність господарств даного регіону є більш однорідною, ніж за врожайністю зернових і зернобобових культур.

Розглянутий приклад показує, що застосування абсолютних характеристик розсіювання ($D_в$, $\sigma_в$) у порівнянні варіації тієї самої ознаки різних сукупностей може призвести до іншого результату, ніж застосування відносних характеристик (наприклад, V).

Так, у розглянутому прикладі середнє квадратичне відхилення врожайності картоплі більше, ніж урожайності зернових, що, здавалося б, могло б свідчити про більшу варіативність даної ознаки. Але й середнє значення другої ознаки значно перевищує середнє першої. У результаті коефіцієнт варіації врожайності картоплі менший, і, відповідно, ця ознака є більш стійкою, ніж урожайність зернових.

Крім розглянутого коефіцієнта варіації існують і інші відносні показники варіації, які визначаються як відношення різних абсолютних мір розсіювання до середньої величини. Серед них:

1) відносний розмах варіації (коефіцієнт осциляції)

$$V_R = \frac{R}{\bar{x}_B} \cdot 100\,\%;$$

2) відносне лінійне відхилення

$$V_{\bar{l}} = \frac{\bar{l}}{\bar{x}_в} \cdot 100\,\% \text{ або } V_{\bar{l}} = \frac{\bar{l}}{Me} \cdot 100\,\%;$$

3) відносний показник квартильної варіації

$$V_Q = \frac{\Delta Q}{2Me} \cdot 100\,\%.$$

2.3. ЕМПІРИЧНІ МОМЕНТИ. ХАРАКТЕРИСТИКИ ФОРМИ: АСИМЕТРІЯ, ЕКСЦЕС

Розглянуті в п. 2.1 та 2.2 основні характеристики центральної тенденції та розсіювання, очевидно, є окремим випадками деякої системи числових характеристик сукупності. Таку систему можна задати, наприклад, за допомогою початкових і центральних моментів статистичного розподілу.

Початковий емпіричний момент порядку *k* визначається так:

$$\tilde{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{2.36}$$

де $k = 1, 2, 3, \ldots$.

У разі коли значення ознаки x_1, x_2, …, x_m мають частоти відповідно n_1, n_2, …, n_m, причому $\sum_{i=1}^{m} n_i = n$,

$$\tilde{\nu}_k = \frac{1}{n}\sum_{i=1}^{m} x_i^k n_i, \tag{2.37}$$

де $k = 1, 2, 3, \ldots$.

Якщо дані згруповано в інтервальний ряд, початкові моменти знаходять за формулою, аналогічною до формули (2.37), лише значення варіант замінюються значеннями середин інтервалів, їхні частоти — частотами відповідних інтервалів.

Очевидно, що вибіркова середня $\bar{x}_в$ є початковим моментом першого порядку.

Центральний емпіричний момент $\tilde{\mu}_k$ порядку *k* дорівнює

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_в)^k, \tag{2.38}$$

де $k = 1, 2, 3, \ldots$.

Якщо дані подані дискретним статистичним рядом, в якому кожній варіанті x_i відповідає частота n_i, причому $\sum_{i=1}^{m} n_i = n$, то

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_{\text{в}}}\right)^k n_i\,, \tag{2.39}$$

де $k = 1, 2, 3, \ldots$.

Отже, вибіркова дисперсія $D_{\text{в}}$ — це центральний емпіричний момент другого порядку. Величина та знак центрального моменту третього порядку залежать від переваги додатних відхилень у кубі над від'ємними, або навпаки. Ця властивість дозволяє використовувати $\tilde{\mu}_3$ для оцінки асиметрії статистичного розподілу, про що йтиметься далі.

На практиці у статистичних дослідженнях застосовуються моменти до четвертого порядку включно. Зауважимо, що їхні значення змінюються від вибірки до вибірки, і тому вони є випадковими величинами. Точність емпіричних моментів зменшується зі зростанням їх порядку. Так, зокрема, дисперсія початкових моментів порядку k залежить від моментів порядку $2k$, тому вона стає досить великою для моментів високих порядків навіть за великих обсягів вибірки.

Початкові та центральні емпіричні моменти пов'язані такими співвідношеннями:

$$\tilde{\mu}_2 = \tilde{v}_2 - \tilde{v}_1^2\,; \tag{2.40}$$

$$\tilde{\mu}_3 = \tilde{v}_3 - 3\tilde{v}_1^2\tilde{v}_2 + 2\tilde{v}_1^3\,; \tag{2.41}$$

$$\tilde{\mu}_4 = \tilde{v}_4 - 4\tilde{v}_1\tilde{v}_3 + 6\tilde{v}_1^2\tilde{v}_2 - 3\tilde{v}_1^4\,. \tag{2.42}$$

Крім раніше розглянутих числових показників, які характеризують середнє значення вибірки та розсіювання всіх значень ознаки навколо нього, для повної характеристики сукупності, що вивчається, необхідні показники форми статистичного розподілу.

Як відомо (див. п. 1.3), для одержання приблизного уявлення про форму розподілу будуються графіки — полігон і гістограма. Але для того щоб порівнювати різні статистичні розподіли з відомими ймовірнісними (наприклад, з нормальним розподілом), а також один з одним, потрібно визначити числові характеристики форми. Розглянемо основні з них.

На практиці у статистичних дослідженнях доводиться зустрічатися з великою кількістю різних розподілів. Як зазначалось у п. 2.1, однорідні вибіркові сукупності є унімодальними, їхні розподіли, відповідно, одновершинні. Для **симетричних розподілів** очевидною є рівність середньої вибіркової, моди та медіани:

$$\overline{x_{\text{в}}} = \text{Me} = \text{Mo}\,.$$

Важливим прикладом симетричного розподілу є нормальний. Але на практиці нерідко зустрічаються й несиметричні розподіли. Як же оцінити міру асиметрії розподілу?

У п. 2.1 під час розгляду таких числових характеристик, як квартилі, було показано, що на основі аналізу співвідношення різниць $Q_2 - Q_1$ та $Q_3 - Q_2$ можна зробити певні висновки про наявність асиметрії розглядуваного розподілу. Отже, одною з відносних мір асиметрії є величина:

$$As_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(Q_3 + Q_1 - 2\text{Me})}{Q_3 - Q_1}\,. \tag{2.43}$$

Значення показника асиметрії, визначене таким способом, лежить у межах від –1 до 1. Зазначимо, що іноді коефіцієнт асиметрії знаходять з формули, аналогічної до (2.43), де замість Q_1 і Q_3 взято D_1 та D_9, тобто

$$As_D = \frac{(D_9 + D_1 - 2\text{Me})}{D_9 - D_1}\,.$$

Для того щоб визначити інший показник асиметрії, розглянемо асиметричний унімодальний розподіл (рис. 2.3).

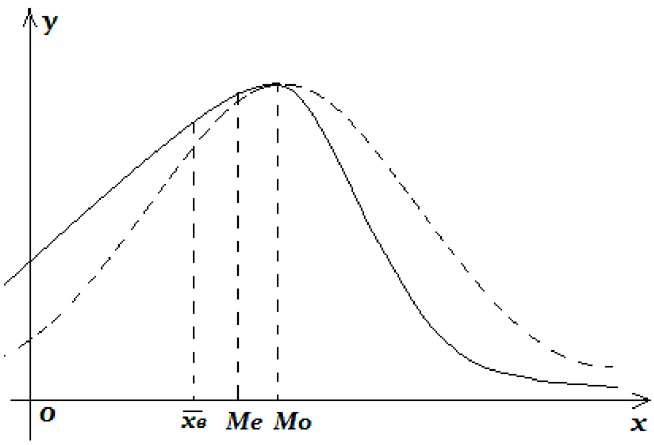


Рис. 2.3

Порівняємо його з базовим симетричним, в якого середня вибіркова, мода та медіана рівні між собою (на рисунку він зображений пунктиром). Якщо в розглядуваного розподілу лівий кінець «піднятий», а правий «опущений», то мода залишиться такою самою, як і в симетричного, а середня вибіркова зміститься ліворуч, адже на її величину впливають усі значення ознаки. Медіана також зміститься ліворуч, але оскільки її значення залежить більшою мірою від частот варіант, ніж від їхніх значень, зміщення буде відносно невеликим порівняно зі зміщенням вибіркової середньої. Тому взаємне розташування в цьому разі моди, медіани та вибіркової середньої таке: $\overline{x_в} < \mathrm{Me} < \mathrm{Mo}$ (див. рис. 2.3). Зазначимо, що таку асиметрію заведено вважати лівобічною. У разі правобічної асиметрії, міркуючи аналогічним чином, можна дійти висновку, що взаємне розташування даних середніх величин таке: $\mathrm{Mo} < \mathrm{Me} < x_в$.

Отже, аналізуючи взаємне розташування Mo, Me та $\overline{x_в}$ одномодального розподілу, можна зробити висновок про його форму. Зокрема, якщо значення даних середніх рівні (або близькі) між собою, то розподіл симетричний (чи близький до симетричного) .

Англійський математик і статистик К. Пірсон на основі вирівнювання різних типів кривих установив, що для помірно асиметричних розподілів справедливим є таке наближене співвідношення між середньою вибірковою, медіаною та модою:

$$\left|\mathrm{Mo} - \overline{x_в}\right| = 3\left|\mathrm{Me} - \overline{x_в}\right|.$$

Крім того, Пірсон запропонував визначення коефіцієнта асиметрії на основі різниці між середньою вибірковою та модою:

$$As_п = \frac{\overline{x_в} - \mathrm{Mo}}{\sigma_в}. \qquad (2.44)$$

Тож із викладеного випливає, що в разі правобічної асиметрії $As_п > 0$, лівобічної — $As_п < 0$.

Вважається, що при $As_п < 0{,}25$ асиметрія низька, при $As_п > 0{,}5$ — висока. Зазначимо, що розглянутий коефіцієнт асиметрії — безрозмірна величина, оскільки різниця між модою та середнім вибірковим ділиться на $\sigma_в$. Це дозволяє порівнювати міру асиметрії різних статистичних розподілів.

Інший коефіцієнт, запропонований шведським математиком Ліндбергом, знаходять за формулою

$$As_Л = P - 50, \qquad (2.45)$$

де P — частка значень ознаки, що перевищують середню вибіркову (у відсотках).

Найбільш точним і найчастіше вживаним є коефіцієнт асиметрії, що обчислюється за допомогою центрального моменту третього порядку:

$$As = \frac{\tilde{\mu}_3}{\sigma^3}. \qquad (2.46)$$

Визначений у такий спосіб показник асиметрії, як і коефіцієнт асиметрії Пірсона ($As_п$), теж є додатним за правобічної асиметрії та від'ємним — за лівобічної. Крім того, він має суттєву перевагу над іншими коефіцієнтами, визначеними раніше. Адже застосування цього показника дає змогу не тільки визначити міру асиметрії, а й установити, чи симетричний розподіл генеральної сукупності. Оцінка міри значущості цього коефіцієнта асиметрії дається за допомогою його стандартного відхилення, що залежить від обсягу вибірки й обчислюється за формулою

$$\sigma_{As} = \sqrt{\frac{6(n-1)}{(n+1)(n+13)}}\,. \tag{2.47}$$

Отже, якщо виконується нерівність

$$\frac{As}{\sigma_{As}} > 3\,,$$

то асиметрія є суттєвою, і розподіл ознаки в генеральній сукупності несиметричний. У протилежному разі асиметрія несуттєва і її наявність у вибіркового розподілу може бути спричинена різними випадковими обставинами.

Слід зазначити, що всі згадані коефіцієнти асиметрії характеризують певним чином форму розподілу й дорівнюють нулю в разі симетричних розподілів. Застосування коефіцієнтів As_Q та $As_П$ дає змогу оцінити міру асиметрії в середній частині ряду розподілу, тимчасовим як показник, побудований на основі центрального моменту третього порядку, залежить більшою мірою від крайніх членів ряду.

Приклад 2.18. Дослідити вибірковий статистичний розподіл урожайності картоплі з прикладу 2.17 (табл. 2.8) на симетричність. Проаналізувати одержані результати і зробити висновок про значущість коефіцієнта асиметрії для генеральної сукупності.

Р о з в ' я з а н н я. Знайдемо інтервальний розподіл вибірки (табл. 2.9):

Таблиця 2.9

Урожайність x, ц/га	108,9—116,6	116,6—124,3	124,3—132	132—139,7	139,7—147,4	147,4—155,1	155,1—162,8
Частота, n_i	1	—	2	3	4	7	3

Маємо (див. приклад 2.17):

$$\overline{x_в} = 143{,}885 \text{ ц/га}; \ \sigma_в = 12{,}81 \text{ ц/га}.$$

Знайдемо структурні середні моду та медіану. Модальним є інтервал 147,4—155,1, отже:

$$\text{Mo} = 147{,}4 + 7{,}7 \cdot \frac{7-4}{2\cdot 7 - 4 - 3} = 150{,}7 \text{ ц/га}.$$

Медіанним є інтервал 139,7—147,4, адже його кумулятивна частота дорівнює 0,5 n. Тобто медіана збігається з правим кінцем цього інтервалу, тобто

$$\text{Me} = 147{,}4 \text{ ц/га}.$$

Звідси маємо таке взаємне розташування середньої вибіркової, моди та медіани:

$$x_в < \text{Me} < \text{Mo}\,,$$

що свідчить про наявність лівобічної асиметрії.

Обчислимо коефіцієнти асиметрії, розглянуті раніше.

Показник асиметрії, що його знаходять за допомогою квартилів (пропонуємо читачеві самостійно переконатись у тому, що $Q_1 = 136{,}2$ ц/га, $Q_3 = 151{,}2$ ц/га):

$$As_Q = \frac{151{,}2 + 136{,}2 - 2\cdot 147{,}4}{151{,}2 - 136{,}2} = -0{,}49\,.$$

Коефіцієнт асиметрії за Пірсоном:

$$As_П = \frac{143{,}885 - 150{,}7}{12{,}81} = -0{,}53\,.$$

Для обчислення коефіцієнта асиметрії, що задається формулою (2.46), знайдемо емпіричний центральний момент третього порядку (проміжні результати — у табл. 2.10):

$$\tilde{\mu}_3 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x_в})^3 = -\frac{39919{,}5}{20} = -1995{,}975 .$$

Таблиця 2.10

i	x_i	$x_i - \overline{x_B}$	$(x_i - \overline{x_B})^3$	$(x_i - \overline{x_B})^4$
1	108,9	–34,985	–42819,9	1 498 054
2	125,2	–18,685	–6523,48	121891,2
3	126,6	–17,285	–5164,26	89264,24
4	133,1	–10,785	–1254,47	13529,46
5	136,2	–7,685	–453,87	3487,992
6	138	–5,885	–203,817	1199,46
7	141	–2,885	–24,0125	69,27607
8	141,2	–2,685	–19,3568	51,97293
9	144	0,115	0,001521	0,000175
10	147,1	3,215	33,23096	106,8375
11	148,2	4,315	80,34196	346,6755
12	148,8	4,915	118,7328	583,5715
13	149,6	5,715	186,6589	1066,756
14	150,4	6,515	276,5306	1801,597
15	151,2	7,315	391,42	2863,237
16	154,2	10,315	1097,508	11320,8
17	154,7	10,815	1264,968	13680,63
18	156,1	12,215	1822,554	22262,5
19	160,4	16,515	4504,387	74389,96
20	162,8	18,915	6767,356	128004,5
	Разом:	0	–39919,5	1983975

Отже,

$$As = \frac{-1995{,}975}{12{,}81^3} = -0{,}95 .$$

Як бачимо, значення всіх коефіцієнтів асиметрії вказують на сильну лівобічну асиметрію розподілу. Це видно й з гістограми (рис. 2.4).

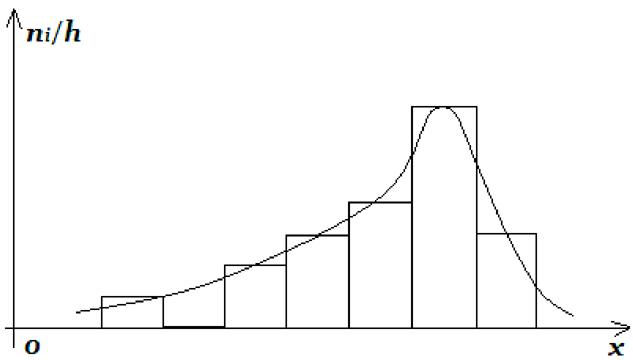


Рис. 2.4

При цьому коефіцієнти, що характеризують асиметрію в центральній частині розподілу (As_Q та $As_П$), менші за абсолютною величиною, ніж коефіцієнт асиметрії, значення якого залежить більшою мірою від екстремальних значень (As). Це означає, що асиметрія в центральній частині менш значна, ніж на кінцях ряду розподілу.

Знайдемо середньоквадратичне відхилення коефіцієнта асиметрії за формулою (2.47):

$$\sigma_{As} = \sqrt{\frac{6 \cdot 19}{21 \cdot 23}} = 0{,}486 .$$

Отже, $\frac{|As|}{\sigma_{As}} = \frac{0{,}95}{0{,}486} = 1{,}95 < 3$. Тому немає підстав уважати, що розподіл генеральної сукупності несиметричний. Можна припускати, що асиметрія вибіркового розподілу може бути викликана впливом випадкових факторів або невеликим обсягом вибірки.

Для симетричних (або близьких до симетричних) одномодальних розподілів можна знайти показник, що характеризує гостровершинність даного розподілу порівняно з нормальним, що має такі самі характеристики (середню та стандартне відхилення).

У разі коли вершина розглядуваного розподілу міститься вище, ніж вершина нормального розподілу, кажуть про додатний ексцес (тобто «надлишок»). У такому разі крива розподілу має більш гостру вершину. Це означає, що в розглядуваній сукупності навколо середньої величини сформоване «ядро», що слабо варіює за даною ознакою. Якщо ж вершина міститься нижче від вершини нормального розподілу і, відповідно, крива більш полога, ексцес вважається від'ємним. За суттєвого від'ємного ексцесу розподіл не має такого «ядра».

Найпростішою мірою ексцесу, в основу розрахунку якої покладено відношення між інтерквартильним та інтердецильним розмахами, є величина

$$Es_Q = \frac{Q_3 - Q_1}{2(D_9 - D_1)} = \frac{\Delta Q}{2\Delta D} . \tag{2.48}$$

Для нормального розподілу $Ex_Q = 0{,}263$. Оскільки зменшення різниці між першим і третім квартилями вказує на те, що так зване ядро більш щільно сформоване навколо середньої величини, то при $Ex_Q < 0{,}263$ ексцес додатній, а при $Ex_Q > 0{,}263$ — від'ємний.

Дж. Ліндберг запропонував такий коефіцієнт ексцесу:

$$Es_л = P - 38{,}29 , \tag{2.49}$$

де P — частка (%) тих значень ознаки, які лежать на відстані меншій, ніж половина середньоквадратичного відхилення, від вибіркової середньої. У нормально розподіленій сукупності ця частка становить 38,29 % від усіх значень ознаки.

Найточніший показник ексцесу знаходять за допомогою центрального моменту четвертого порядку:

$$Es = \frac{\tilde{\mu}_4}{\sigma^4} - 3 . \tag{2.50}$$

Для нормального розподілу величина визначеного в такий спосіб коефіцієнту ексцесу дорівнює нулю. Розподіли з від'ємним та додатним значеннями Es показані на рис. 2.5.

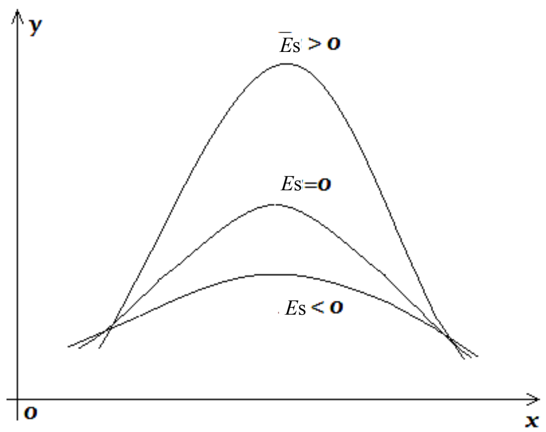


Рис. 2.5

Остання характеристика ексцесу зручна в застосуванні ще й тому, що відоме її середньоквадратичне відхилення, що залежить від обсягу вибірки:

$$\sigma_{Es} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}\,. \tag{2.51}$$

Це дає змогу оцінити, наскільки суттєва величина ексцесу для розподілу генеральної сукупності. Так, якщо

$$\frac{|Es|}{\sigma_{Es}} < 3\,,$$

то немає підстав уважати, що розподілу генеральної сукупності властивий ексцес.

Зазначимо, що на практиці, оцінюючи значення відношення абсолютної величини коефіцієнта асиметрії або ексцесу до відповідного стандартного відхилення цього показника, не слід забувати про те, що висновки, сформульовані раніше, можна робити лише за досить великих обсягів вибірок. У протилежному ж разі вони не завжди коректні.

Приклад 2.19. Знайдемо для даних про врожайність картоплі у 20 господарствах регіону з прикладу 2.17 (табл. 2.8) коефіцієнт ексцесу.

Р о з в ' я з а н н я. Для обчислення показника ексцесу за формулою (2.48) знайдемо інтердецильний розмах:

$$D_1 = P_{10} = 125{,}2\,,\quad D_9 = P_{90} = 156{,}1\,,$$
$$\Delta D = 30{,}9\,.$$

Отже,

$$Es_Q = \frac{\Delta Q}{2\Delta D} = \frac{15}{2\cdot 30{,}9} = 0{,}243\,.$$

Одержане значення коефіцієнта ексцесу менше за аналогічне значення для нормального розподілу, що свідчить про невеликий додатний ексцес, тобто більшу концентрацію центральних значень ряду навколо середньої.

Коефіцієнт ексцесу за Ліндбергом у даному разі становить:

$$Es_{\text{л}} = 40 - 38{,}29 = 1{,}71\,\%\,,$$

що також вказує на наявність додатного ексцесу.

Обчислимо найчастіше вживаний коефіцієнт ексцесу, що його знаходять за формулою (2.50). Для цього визначимо емпіричний центральний момент четвертого порядку (проміжні розрахунки див. у табл. 2.10):

$$\tilde{\mu}_4 = \frac{1}{20}\sum_{i=1}^{20}\left(x_i - \overline{x}_{\text{в}}\right)^4 = \frac{1983975}{20} = 99198{,}75\,.$$

Звідси

$$Es = \frac{\tilde{\mu}_4}{D_{\text{в}}^2} - 3 = \frac{99198{,}75}{164{,}11^2} - 3 = 0{,}68\,.$$

Отже, можна зробити висновок про незначну гостровершинність емпіричного розподілу порівняно з нормальним.

Знайдемо середньоквадратичне відхилення ексцесу за формулою (2.51):

$$\sigma_{Es} = \sqrt{\frac{24\cdot 20\cdot 18\cdot 17}{19^2\cdot 23\cdot 25}} = 0{,}84\,.$$

Отже, відношення $\frac{|\mathrm{E}s|}{\sigma_{Es}} = \frac{0{,}68}{0{,}84} = 0{,}81 < 3$. Це свідчить про те, що немає підстав уважати, що ексцес властивий розподілу генеральної сукупності.

Слід зазначити, що показники асиметрії та ексцесу не тільки характеризують форму емпіричного розподілу даної сукупності, а й служать певним орієнтиром для дальшого вивчення соціально-економічних явищ. Наприклад, значний від'ємний ексцес може свідчити про якісну неоднорідність розглядуваної сукупності. Крім того, значення коефіцієнтів асиметрії та ексцесу дають змогу робити певні висновки про близькість розглядуваного розподілу до нормального.

Закінчуючи розгляд показників асиметрії та ексцесу розподілу, наведемо приклад їх застосування в економічних дослідженнях. У вивченні альтернатив типу «ризик—дохідність» широко використовуються не тільки середні характеристики та показники варіації, а й коефіцієнти асиметрії та ексцесу. Зміст цих коефіцієнтів полягає в такому. Зменшення від'ємного значення показника асиметрії характеризує позитивну тенденцію, коли збільшується ймовірність появи значень дохідності, що перевищують середнє її значення (середнє арифметичне). І навпаки, збільшення додатного показника асиметрії означає, що збільшується ймовірність одержання менших за середню величину доходів. Тобто, асиметрія є певним критерієм, що характеризує міру ризику.

Охарактеризуємо суть ризику, яка відображається коефіцієнтом ексцесу за різних його значень. Зменшення від'ємного значення показника ексцесу приводить до того, що графік функції щільності ймовірностей випадкової величини доходу стає більш пологим і розтягнутим. Остання властивість «розтягнутості» означає розширення області значень очікуваної дохідності в обидва боки від середньої величини. Посилення такої властивості економісти трактують як зростання ймовірності отримати малий дохід і зменшення ймовірності одержати середній. Водночас збільшення додатного значення коефіцієнта ексцесу показує, що графік функції щільності ймовірностей витягується, стає більш гостровершинним, формуючи при цьому ядро значень, що слабко варіює в центральній частині розподілу. Посилення цієї властивості трактують як збільшення ймовірності одержати значення дохідності близьке до середнього і зменшення ймовірності одержати невелике значення дохідності.

2.4. УМОВНІ ВАРІАНТИ

Під час обчислення числових характеристик вибірки, заданої інтервальним статистичним рядом з ***однаковими довжинами інтервалів***, зручно перейти до умовних варіант за формулою

$$u_l = \frac{x_i^* - C}{\Delta x}, \ i = 1, 2, \ldots, k; \tag{2.52}$$

де x_i^* — середини інтервалів, на які розбито розподіл; $\Delta x = h$ — довжини інтервалів; k — кількість інтервалів, C — умовний нуль, за який можна взяти будь-яке значення x_i^*.

Обчислення числових характеристик набувають максимальної простоти, якщо за уявний нуль взяти значення x_i^*, яке міститься якнайближче до середини інтервального ряду.

Покажемо, що умовні варіанти u_i є цілими числами. Справді, якщо довжини інтервалів Δx однакові, то варіанти x_i^* утворюють арифметичну прогресію, різниця якої Δx. Тому за формулою загального члена арифметичної прогресії $x_i^* = x_1^* + (i-1)\cdot\Delta x, \ i = 1, 2, \ldots, k.$ Виберемо за умовний нуль $C = x_m^*$, тоді

$$u_i = \frac{x_i^* - x_m^*}{\Delta x} = \frac{x_1^* + (i-1)\Delta x - x_1^* - (m-1)\Delta x}{\Delta x} = \frac{(i-m)\Delta x}{\Delta x} = i - m.$$

Номери інтервалів i та m — цілі числа, тому і їхня різниця також ціле число. Варіанті x_m^*, взятій за умовний нуль, відповідатиме умовна варіанта $u_m = 0.$ Інші умовні варіанти завжди мають такий вигляд:

$$\ldots, -2, -1, 0, 1\ 2, \ldots$$

Уводяться умовні статистичні моменти різних порядків: початкові

$$v_{ru} = \frac{1}{n}\sum_{i=1}^{k} u_i^r n_i \tag{2.53}$$

та центральні

$$\mu_{ru} = \frac{1}{n}\sum_{i=1}^{k}\left(u_i - \overline{u}\right)^r n_i \tag{2.54}$$

тут $\overline{u} = \frac{1}{n}\sum_{i=1}^{k} u_i n_i$ — умовне середнє значення.

Між початковими та центральними умовними моментами існують такі самі залежності, як і залежності (2.40)—(2.42) між початковими та центральними емпіричними моментами.

Виразимо вибіркову середню через умовні моменти:

$$\overline{x}_{в} = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i - C + C = \frac{1}{n}\sum_{i=1}^{k} x_i^* n_i - \frac{1}{n}\sum_{i=1}^{k} C n_i + C = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - C) n_i + C =$$

$$= \Delta x \cdot \frac{1}{n}\sum_{i=1}^{k}\frac{x_i^* - C}{\Delta x} \cdot n_i + C = \Delta x \cdot \frac{1}{n}\sum_{i=1}^{k} u_i n_i + C = \Delta x \cdot \overline{u} + C.$$

Отже, маємо

$$\overline{x}_{в} = \overline{u} \cdot \Delta x + C. \tag{2.55}$$

Виразимо центральні емпіричні моменти через умовні:

$$\tilde{\mu}_r = \frac{1}{n}\sum_{i=1}^{k}(x_i^* - \overline{x}_{в})^r n_i = \frac{1}{n}\sum_{i=1}^{k}(u_i \Delta x + C - \overline{u}\Delta x - C)^r n_i =$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left((u_i - \overline{u})\Delta x\right)^r n_i = (\Delta x)^r \cdot \frac{1}{n}\sum_{i=1}^{k}(u_i - \overline{u})^r n_i = (\Delta x)^r \cdot \mu_{ru}.$$

Тож дістаємо

$$\tilde{\mu}_r = \mu_{ru} \cdot (\Delta x)^r. \tag{2.56}$$

Уведемо позначення:

$$\overline{u} = \nu_{1u};\ \overline{u^2} = \nu_{2u};\ \overline{u^3} = \nu_{3u};\ \overline{u^4} = \nu_{4u}.$$

Тоді з (2.56), ураховуючи співвідношення між початковими та центральними моментами, дістанемо формули для обчислення вибіркової дисперсії, середнього квадратичного відхилення, емпіричних коефіцієнтів асиметрії та ексцесу:

$$D_B = \left(\overline{u^2} - (\overline{u})^2\right) \cdot (\Delta x)^2; \tag{2.57}$$

$$\sigma_{в} = \sqrt{\overline{u^2} - (\overline{u})^2} \cdot \Delta x; \tag{2.58}$$

$$As = \frac{\overline{u^3} - 3\overline{u^2} \cdot \overline{u} + 2(\overline{u})^3}{\sqrt{\left(\overline{u^2} - (\overline{u})^2\right)^3}}; \tag{2.59}$$

$$Es = \frac{\overline{u^4} - 4\overline{u^3} \cdot \overline{u} + 6\overline{u^2} \cdot (\overline{u})^2 - 3(\overline{u})^4}{\left(\overline{u^2} - (\overline{u})^2\right)^2} - 3. \tag{2.60}$$

Приклад 2.20. У табл. 2.11 подано дані про чисельність населення у сільській місцевості в регіонах України станом на 1 лютого 2011 р.[1]. Знайти числові характеристики розподілу.

Таблиця 2.11

Область	Кількість жителів	Область	Кількістьжителів
АР Крим	729479	Миколаївська	382195
Вінницька	826152	Одеська	793725
Волинська	499929	Полтавська	578702
Дніпропетровська	550774	Рівненська	601458
Донецька	419575	Сумська	378005
Житомирська	537233	Тернопільська	608965
Закарпатська	783716	Харківська	549541
Запорізька	415659	Херсонська	422577
Івано-Франківська	782325	Хмельницька	598584
Київська	663763	Черкаська	564075
Кіровоградська	384129	Чернівецька	522977
Луганська	304540	Чернігівська	408047
Львівська	997873	Севастополь	23469

Р о з в ' я з а н н я. Визначимо серед статистичних даних мінімальне $x_{\min} = 23469$ та максимальне $x_{\max} = 997869$ значення і розіб'ємо діапазон $[x_{\min}; x_{\max}]$ на шість однакових інтервалів шириною

$$\Delta x = \frac{x_{\max} - x_{\min}}{6} = \frac{997869 - 23469}{6} = 162400.$$

Підрахуємо частоти кожного інтервалу і складемо інтервальний статистичний ряд

Інтервали	23469–185869	185869–348269	348269–510669	510669–673069	673069–835469	835469–997869
Частоти	1	1	8	10	5	1

Для визначення числових характеристик перейдемо до дискретного статистичного ряду, замінивши кожний інтервал його серединою:

x_i^*	104669	267069	429469	591869	754269	916669
n_i	1	1	8	10	5	1

Виберемо умовний нуль $C = 591869$ і перейдемо до умовних варіант

$$u_i = \frac{x_i^* - C}{\Delta x} = \frac{x_i^* - 591869}{162400}.$$

Розрахунки подамо у вигляді табл. 2 12.

Таблиця 2.12

	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i u_i^3$	$n_i u_i^4$	$n_i(1+u_i)^4$
	1	–3	–3	9	–27	81	16
	1	–2	–2	4	–8	16	1
	8	–1	–8	8	–8	8	0
	10	0	0	0	0	0	10
	5	1	5	5	5	5	80
	1	2	2	4	8	16	81
Разом	26	–3	–6	30	–30	126	188

[1] Держкомстат України :[Електронний ресурс] : експрес-вип. — 2011. — 16 берез. — Режим доступу : http://www.ukrstat.gov.ua.

Останній стовпчик таблиці призначений для контролю обчислень, який проводиться на основі тотожності

$$\sum_{i=1}^{k} n_i(1+u_i)^4 = \sum_{i=1}^{k} n_i + 4\sum_{i=1}^{k} n_i u_i + 6\sum_{i=1}^{k} n_i u_i^2 + 4\sum_{i=1}^{k} n_i u_i^3 + \sum_{i=1}^{k} n_i u_i^4.$$

У нашому випадку маємо

$$188 = 26 + 4(-6) + 6\cdot 30 + 4(-30) + 126 = 188,$$

тож розрахунки проведено правильно.

Обчислимо початкові умовні моменти:

$$\nu_{1u} = -\frac{3}{13}; \quad \nu_{2u} = \frac{15}{13}; \quad \nu_{3u} = -\frac{15}{13}; \quad \nu_{4u} = \frac{63}{13}.$$

За формулою (2.55) знайдемо середнє значення

$$\bar{x}_{\text{в}} = -\frac{3}{13}\cdot 162400 + 591869 \approx 524392{,}0769.$$

Знайдемо дисперсію за формулою (2.57):

$$D = \left(\frac{15}{13} - \frac{9}{169}\right)\cdot 162400^2 = \frac{186}{169}\cdot 162400^2 = 29026741775{,}15.$$

$$\sigma = \frac{\sqrt{186}}{13}\cdot 162400 \approx 170372{,}36.$$

Коефіцієнт варіації $V = \frac{170372{,}36}{524392{,}0769}\cdot 100$ % = 32,4 % це свідчить про те, що населення в сільській місцевості в регіонах України розподілено досить нерівномірно.

Обчислимо коефіцієнти асиметрії та ексцесу за формулами (2.59) : (2.60)

$$As = \frac{-\frac{15}{13} - 3\cdot\frac{15}{13}\cdot\left(-\frac{3}{13}\right) + 2\cdot\left(-\frac{3}{13}\right)^3}{\left(\frac{\sqrt{186}}{13}\right)^3} = -\frac{834}{\sqrt{186^3}} \approx -0{,}3618;$$

$$Es = \frac{\frac{63}{13} - 4\cdot\left(-\frac{15}{13}\right)\cdot\left(-\frac{3}{13}\right) + 6\cdot\frac{15}{13}\cdot\frac{9}{169} - 3\cdot\frac{81}{13^4}}{\left(\frac{186}{169}\right)^2} - 3 = \frac{118278}{34596} - 3 \approx 0{,}4188.$$

Зауваження. Обчислення числових характеристик за схемою, запропонованою в табл. 2.12, що в літературі зі статистики має назву методу добутків обчислення числових характеристик.

Розв'язання типових задач

Задача 2.1. У табл. 2.13 представлено дані про темпи зростання реальної заробітної плати[1] в Україні у 1995 — 2010 рр. Яким є середній темп росту заробітної плати за рік?

Таблиця 2.13

Рік	1995	1996	1997	1998	1999	2000	2001	2002	2003
Реальна заробітна плата, % до попереднього року	110,6	96,6	96,6	96,2	91,1	99,1	119,3	118,2	115,2

2004	2005	2006	2007	2008	2009	2010
123,8	120,3	118,3	112,5	106,3	90,8	110,2

[1] Держкомстат України [Електронний ресурс] : 1998—2011. — Режим доступу, http://www.ukrstat.gov.ua

Р о з в ’ я з а н н я. Для знаходження середнього темпу зростання зарплати вибіркова середня (середня арифметична) непридатна. Адже за заміни індивідуальних значень ознаки (у даному разі темпів зростання у відсотках до попереднього року) на середню величину має залишитися незмінним добуток цих значень. Тому слід застосувати середню геометричну:

$$\overline{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \sqrt[16]{110{,}6 \cdot 96{,}6^2 \cdot 96{,}2 \cdot 91{,}1 \cdot 99{,}1 \cdot 119{,}3 \cdot 118{,}2 \cdot 115{,}2 \cdot 123{,}8 \cdot 120{,}3 \cdot 118{,}3 \cdot 112{,}5 \cdot 106{,}3 \cdot 90{,}8 \cdot 110{,}2} \approx 107{,}2616.$$

Отже, середньорічний темп зростання зарплати за даний період становить 107,2616 %.

Задача 2.2. За даними про динаміку середньомісячних сукупних витрат у розрахунку на одне домогосподарство[1] (табл. 2.14) знайти середньорічний темп зростання витрат.

Таблиця 2.13

Рік	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Сукупні витрати, грн	426,5	541,3	607,0	658,3	736,8	903,5	1229,4	1442,8	1722,0	2590,4	2754,1

Р о з в ’ я з а н н я. У даному разі для обчислення середньорічного темпу зростання сукупних щомісячних витрат у розрахунку на одне домогосподарство скористаємося формулою (2.6) для знаходження середньої геометричної:

$$\overline{t} = \sqrt[n-1]{\frac{y_n}{y_1}} = \sqrt[10]{\frac{2754{,}1}{426{,}5}} \approx 1{,}205$$

Отже, середньорічний темп зростання витрат — 120,5 %.

Задача 2.3. Для перевезення сировини зі складу на підприємство було зроблено два рейси. Під час перевезення першої партії вантажу вантажівка їхала зі швидкістю 55 км/год, другої — 65 км/год. Назад вона щоразу поверталася зі швидкістю 70 км/год. Знайти середню швидкість вантажівки за чотири поїздки.

Р о з в ’ я з а н н я. Логічна формула для знаходження середньої швидкості така:

$$\mathrm{v}_{сер} = \frac{\text{шлях, км}}{\text{час, год}}.$$

Оскільки відстань перевезень залишалася незмінною (її умовно можна взяти рівною 1), скористаємося формулою середньої гармонійної простої:

$$\mathrm{v}_{сер} = x_{гарм} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{4}{\frac{1}{55}+\frac{1}{70}+\frac{1}{65}+\frac{1}{70}} = 64{,}4 \text{ км/год}.$$

Отже, середня швидкість вантажівки складає 64,4 км/год

Задача 2.4. У табл. 2.15 подано дані щодо чотирьох груп вишів одного регіону. Знайти середні значення наведених показників та обґрунтувати вибір виду середньої величини у кожному випадку.

Таблиця 2.15

Рівень акредитації	Загальна кількість викладачів	Середня кількість викладачів у одному виші	Кандидати та доктори наук, %
I	1276	58	—
II	1940	97	40
III	2220	185	55
IV	3192	399	75

[1] Держкомстат України [Електронний ресурс] : 1998—2011. — Режим доступу, http://www.ukrstat.gov.ua

Р о з в ’ я з а н н я.

1. Оскільки немає даних про кількість вишів різних рівнів акредитації в даному регіоні, для обчислення середньої кількості викладачів у всіх навчальних закладах скористаємось середньою гармонійною:

$$\bar{x}_1 = \frac{1276 + 1940 + 2220 + 3192}{\frac{1276}{58} + \frac{1940}{97} + \frac{2220}{185} + \frac{3192}{399}} = \frac{8628}{62} = 139 .$$

2. Середній відсоток викладачів з науковим ступенем обчислимо за формулою середньої арифметичної зваженої, адже логічною формулою для її знаходження є така:

$$\bar{x}_2 = \frac{\text{Кількість докторів і кандидатів}}{\text{Загальна кількість викладачів, осіб}} .$$

Отже,

$$\bar{x}_2 = \frac{0 \cdot 1276 + 0,4 \cdot 1940 + 0,55 \cdot 2220 + 0,75 \cdot 3192}{1276 + 1940 + 2220 + 3192} = \frac{4391}{8628} = 0,51 \cdot$$

Тобто, середній відсоток кандидатів і докторів наук у всіх вишах даного регіону — 51 %.

Задача 2.5. Дані про кількість угод, укладених за посередництва агентства нерухомості «Комфорт», помісячно за останні два роки подані в табл. 2.16. Проаналізувати, наскільки сильно впливали сезонні коливання попиту на нерухомість на варіацію кількості угод, що укладалися даним агентством протягом зазначеного періоду.

Таблиця 2.16

1-й рік

Місяць	1	2	3	4	5	6	7	8	9	10	11	12
К-сть угод	17	14	18	11	10	8	12	14	21	25	20	16

2-й рік

Місяць	1	2	3	4	5	6	7	8	9	10	11	12
К-сть угод	14	12	13	10	8	9	5	8	10	13	18	9

Р о з в ’ я з а н н я. Згрупуємо дані за сезонною ознакою (табл. 2.17): 1 гр. — зимово-весняний період (1, 2, 3 і 4-й місяці), 2 гр. — літній (5, 6, 7, 8-й місяці), 3 гр. — осінньо-зимовий (9, 10, 11 і 12-й місяці).

Таблиця 2.17

Група даних (за сезонною ознакою)	1	2	3
К-сть угод на місяць	17, 14, 18, 11, 14, 12, 13, 10	10, 8, 12, 14, 8, 9, 5, 8	21, 25, 20, 16, 10, 13, 18, 9

Спостережувана ознака — кількість угод, укладених протягом місяця, варіює під впливом систематичного фактора — сезонних коливань попиту (міжгрупова варіація), а також інших випадкових факторів (внутрішньогрупова варіація). Задача полягає у вимірюванні цих варіацій за допомогою міжгрупової, внутрішньогрупової та загальної дисперсій.

Для розрахунку групових дисперсій знайдемо середню кількість угод за кожною групою та загальну середню кількість угод:

за першою групою $\bar{x}_{в1} = \frac{1}{8}(17 + 14 + 18 + 11 + 14 + 12 + 13 + 10) = 13,625$;

за другою групою $\bar{x}_{в2} = \frac{1}{8}(10 + 8 + 12 + 14 + 8 + 9 + 5 + 8) = 9,25$;

за третьою групою $\bar{x}_{в3} = \frac{1}{8}(21 + 25 + 20 + 16 + 10 + 13 + 18 + 9) = 16,5$;

загальна $\bar{x}_в = \frac{1}{3}(13,625 + 9,25 + 16,5) = 13,125$.

Обчислимо групові дисперсії:

$$D_{в1} = \frac{1}{8}(17^2 + 14^2 + 18^2 + 11^2 + 14^2 + 12^2 + 13^2 + 10^2) - 13,625^2 = 6,734;$$

$$D_{в2} = \frac{1}{8}(10^2 + 8^2 + 12^2 + 14^2 + 8^2 + 9^2 + 5^2 + 8^2) - 9,25^2 = 6,6875,$$

$$D_{в3} = \frac{1}{8}(21^2 + 25^2 + 20^2 + 16^2 + 10^2 + 13^2 + 18^2 + 9^2) - 16,5^2 = 27,25.$$

Групові дисперсії характеризують варіацію кількості угод за кожною з груп, що спричинена всіма можливими факторами, крім сезонності.

Знайдемо внутрішньогрупову дисперсію, тобто середню з групових:

$$D_{вн} = \frac{1}{3}(6,734 + 6,6875 + 27,25) = 13,557.$$

Обчислимо за формулою (2.25) міжгрупову дисперсію:

$$D_M = \frac{1}{3}\left[(13,625 - 13,125)^2 + (9,25 - 13,125)^2 + (16,5 - 13,125)^2\right] = 8,885.$$

Ця дисперсія характеризує варіацію групових середніх відносно загальної середньої, тобто вона показує величину впливу сезонних коливань попиту на варіацію кількості укладених угод.

Згідно з формулою (2.26) загальна вибіркова дисперсія є сумою двох знайдених дисперсій:

$$D_в = D_{вн} + D_м = 13,557 + 8,885 = 22,442.$$

Знайдемо емпіричний коефіцієнт детермінації, що показує частку міжгрупової дисперсії в загальній дисперсії:

$$\eta^2 = \frac{8,885}{22,442} = 0,396.$$

Це означає, що майже на 40 % (39,6 %) варіація кількості угод, укладених за посередництва даного агентства нерухомості протягом двох років, залежала від сезонних коливань попиту на нерухомість.

Задача 2.6. У табл. 2.18 подано дані про ціну на яблука (грн/кг) у 10 крамницях і на 6 ринках міста у вересні 2010 р.

А. Знайти середні ціни на ринку та в крамниці.

Б. Знайти середню ціну на яблука в усіх торгових точках міста.

В. Знайти показники варіації ціни на яблука в місті за зазначений період.

Таблиця 2.18

№ з/п		1	2	3	4	5	6	7	8	9	10
Ринки	Ціна за 1 кг, грн	6	7,2	6,5	6,8	6,4	7				
	Кількість проданого товару, т	5	3	2,7	1,8	4	1,6				
Крамниці	Ціна за 1 кг, грн	6,2	5	5,7	6	7	6,6	7,2	5	5,7	6
	Кількість проданого товару, т	3	2	2,3	1,9	1,3	3,5	2	4	3	1,8

Р о з в ' я з а н н я.

А. Середня ціна на яблука на ринках міста та в крамницях відповідно дорівнює:

$$\bar{x}_{B1} = \frac{6 \cdot 5 + 7,2 \cdot 3 + 6,5 \cdot 2,7 + 6,8 \cdot 1,8 + 6,4 \cdot 4 + 7 \cdot 1,6}{5 + 3 + 2,7 + 1,8 + 4 + 1,6} = 6,52 \text{ грн/кг},$$

$$\bar{x}_{в2} = \frac{6,2 \cdot 3 + 5 \cdot 2 + 5,7 \cdot 2,3 + 6 \cdot 1,9 + 7 \cdot 1,3 + 6,6 \cdot 3,5 + 7,2 \cdot 2 + 5 \cdot 4 + 5,7 \cdot 3 + 6 \cdot 1,8}{3 + 2 + 2,3 + 1,9 + 1,3 + 3,5 + 2 + 4 + 3 + 1,8} =$$

$= 5,95$ грн/кг.

Б. Знайдемо середню ціну на яблука в усіх торгових точках міста, скориставшись 5-ю властивістю вибіркової середньої:

$$\bar{x}_в = \frac{6,52 \cdot 18,1 + 5,95 \cdot 24,8}{42,9} = 6,19 \text{ грн/кг.}$$

В. Спочатку знайдемо абсолютні показники варіації: середнє квадратичне відхилення, середнє лінійне відхилення, розмах варіації та інтерквартильний розмах. Для цього зручно побудувати таблицю, в якій будуть знайдені відхилення всіх варіант від середньої та їхні квадрати, зважені на частоти, які в даному прикладі є кількостями проданого товару (у тоннах):

№	x_i	n_i	$\lvert x_i - \bar{x}_в \rvert \cdot n_i$	$(x_i - \bar{x}_в)^2 \cdot n_i$
1	5	6	7,14	8,4966
2	5,7	5,3	2,597	1,27253
3	6	8,7	1,653	0,31407
4	6,2	3	0,03	0,0003
5	6,4	4	0,84	0,1764
6	6,5	2,7	0,837	0,25947
7	6,6	3,5	1,435	0,58835
8	6,8	1,8	1,098	0,66978
9	7	2,9	2,349	1,90269
10	7,2	5	5,05	5,1005
$\sum$		42,9	23,029	18,78069

Отже, розмах варіації:

$$R = x_{\max} - x_{\min} = 7,2 - 5 = 2,2 \text{ грн/кг;}$$

середнє лінійне відхилення:

$$\bar{l} = \frac{23,029}{42,9} = 0,537 \text{ грн/кг;}$$

середнє квадратичне відхилення:

$$\sigma_в = \sqrt{D_в} = \sqrt{\frac{18,78069}{42,9}} = 0,6616 \text{ грн/кг .}$$

Як бачимо, середньоквадратичне відхилення більше за середнє лінійне відхилення, а точніше,

$$\sigma_в : \bar{l} \approx 1,232 .$$

Це свідчить про те, що розподіл розглядуваної ознаки близький до симетричного.
Знайдемо квартилі:

$$Q_1 = 5,7 \text{ грн/кг, } Q_2 = \text{Me} = 6,2 \text{ грн/кг, } Q_3 = 6,6 \text{ грн/кг.}$$

Тоді інтерквартильний розмах дорівнює:

$$\Delta Q = Q_3 - Q_1 = 6,6 - 5,7 = 0,9 \text{ грн/кг.}$$

Обчислимо відносні показники варіації:
коефіцієнт варіації:

$$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\,\% = \frac{0{,}6616}{6{,}19} \cdot 100\,\% \approx 10{,}7\,\%\,;$$

коефіцієнт осциляції:

$$V_R = \frac{R}{\bar{x}_в} \cdot 100\,\% = \frac{2{,}2}{6{,}19} \cdot 100\,\% \approx 35{,}5\,\%\,;$$

відносне лінійне відхилення:

$$V_{\bar{l}} = \frac{\bar{l}}{x_в} \cdot 100\,\% = \frac{0{,}537}{6{,}19} \cdot 100\,\% = \frac{0{,}537}{6{,}19} \cdot 100\,\% \approx 8{,}7\,\%\,;$$

відносний показник квартильної варіації:

$$V_Q = \frac{\Delta Q}{2\text{Me}} \cdot 100\,\% = \frac{0{,}9}{2 \cdot 6{,}19} 100\,\% \approx 7{,}3\,\%\,.$$

Знайдені відносні показники варіації показують частку розмаху, середнього лінійного відхилення, середньоквадратичного відхилення та середнього квартильного розмаху в середній величині. Вони дозволяють порівнювати варіацію різних ознак або однієї ознаки, але в різних сукупностях. Наприклад, знаючи відносні показники варіації цін на інші фрукти чи овочі, можна було б порівнювати, наскільки варіативною в місті є ціна на яблука проти цін на інші види продукції.

Крім того, відносні показники дають можливість оцінити міру однорідності розглядуваної сукупності. У нашому випадку сукупність цін на яблука в місті є достатньо однорідною, оскільки $V \approx 10{,}7\,\% < 33\,\%$.

Задача 2.7. У табл. 2.19 подано дані про розподіл населення України за рівнем середньодушових загальних доходів у першому півріччі 2010 р.[1]

А. Знайти середню величину щомісячного середньодушового доходу в Україні за зазначений період і показник варіації доходів. Наскільки нерівномірно розсіяні доходи навколо свого середнього значення?

Б. Знайти частку населення із середньодушовими загальними доходами в місяць нижчими від прожиткового мінімуму, який у той період становив 832 грн.

Таблиця 2.19

Середньодушовий дохід у місяць за всіма домогосподарствами, грн	до 300	300,1—480	480,1—660	660,1—840	840,1—1020
Населення, %	0,5	3,6	8,9	15,7	17,3

1020,1—1200	1200,1—1380	1380,1—1560	1560,1—1740	1740,1—1920	понад 1920
15,5	11,4	7,2	5,9	3,8	10,2

Р о з в ’ я з а н н я.

А. Оскільки середнє арифметичне значення доходу досить чутливе до екстремальних значень розглядуваної ознаки, то воно менш показове, ніж, наприклад, медіана. Слід зазначити, що в статистиці більшості країн для характеристики загального рівня доходів використовується саме медіанний рівень, тобто такий, нижчим і вищим від якого є середньодушовий дохід 50 % населення.

Отже, оскільки фактично маємо інтервальний ряд з відносними частотами, скористаємось формулою (2.12) для знаходження медіани:

$$Me = 1020{,}1 + \frac{0{,}5 - 0{,}46}{0{,}615 - 0{,}46} \cdot 180 = 1066{,}6 \text{ грн.}$$

Тож рівно половина населення мала в той період загальний середньодушовий дохід менший за 1066,6 грн, і половина — більший.

[1] Держкомстат України [Електронний ресурс] : 1998—2011. — Режим доступу : http://www.ukrstat.gov.ua.

Знайдемо показник варіації доходів. Оскільки як середню величину у даному разі було взято медіану, то для характеризації розсіювання доходів навколо середньої величини доцільно знайти інтерквартильний розмах. Для цього обчислимо нижній і верхній квартилі:

$$Q_1 = 660,1 + \frac{0,25 - 0,13}{0,287 - 0,13} \cdot 180 = 797,7 \text{ грн;}$$

$$Q_3 = 1380,1 + \frac{0,75 - 0,729}{0,801 - 0,729} \cdot 180 = 1432,6 \text{ грн.}$$

Отже,

$$\Delta Q = Q_3 - Q_1 = 1432,6 - 797,7 = 634,9 \text{ грн.}$$

Для того щоб оцінити однорідність сукупності середньодушових доходів, знайдемо відносний показник розсіювання, а саме відносний показник квартильної варіації:

$$V_Q = \frac{\Delta Q}{2Me} \cdot 100\ \% = \frac{634,9}{2 \cdot 1066,6} \cdot 100\ \% \approx 29,8\ \% .$$

Знайдене значення свідчить про достатню однорідність розглядуваної сукупності. Але не слід забувати, що знайдений у такий спосіб відносний показник варіації характеризує розсіювання лише середніх 50 % відсотків населення за ознакою середньодушового доходу, і в його обчисленні не враховувались екстремальні значення цієї ознаки.

Б. Для того щоб знайти частку населення із середньодушовими загальними доходами нижчими від прожиткового мінімуму, який дорівнював 832 грн, потрібно розв'язати обернену задачу: маючи значення p-квантиля, треба знайти величину p.

Отже, маємо рівняння

$$832 = 660,1 + \frac{100p - 13}{15,7} \cdot 180 \cdot$$

Неважко пересвідчитись, що єдиним його розв'язком є $p = 0,28$. Тож 28 % населення у першому півріччі 2010 р. мали місячний дохід нижчий від прожиткового мінімуму.

Задача 2.8. Маємо інтервальний ряд розподілу 50 найбільших комерційних банків України за величиною активів станом на 01.05.2011 р.[1] (табл. 2.20). Знайти середні величини, показники варіації та форми даного розподілу. На основі знайдених числових характеристик зробити висновок про близькість розглядуваного статистичного розподілу до нормального. Знайти емпіричний коефіцієнт детермінації.

Таблиця 2.20

Активи, млрд грн	2,5—3,5	3,5—5,5	5,5—10,5	10,5—20,5	20,5—40,5	40,5—80,5
Кількість банків	7	10	15	9	6	3

Р о з в ' я з а н н я.

Для того щоб знайти числові характеристики інтервального варіаційного ряду, побудуємо таку таблицю:

	Середини інтервалів, x_i^*	Частоти, n_i	$x_i^* \cdot n_i$	$x_i^{*2} \cdot n_i$	$x_i^{*3} \cdot n_i$	$x_i^{*4} \cdot n_i$
	3	7	21	63	189	567
	4,5	10	45	202,5	911,25	4100,625
	8	15	120	960	7680	61440
	15,5	9	139,5	2162,25	33514,875	519480,5625
	30,5	6	183	5581,5	170235,75	5192190,375
	60,5	3	181,5	10980,75	664335,375	40192290,2
$\sum$		50	690	19950	876866,25	45970068,75

[1] Для побудови інтервального варіаційного ряду було використано матеріали сайта http://bank-ua.com. Для уникнення спотворення результатів дослідження перший за величиною активів банк не було враховано як одиницю розглядуваної сукупності, оскільки розмір його активів сильно відрізняється від інших значень.

Отже, середнє арифметичне значення

$$\bar{x}_{\text{в}}=\frac{\sum_{i=1}^{6}x_i^* n_i}{n}=\frac{690}{50}=13,8 \text{ млрд грн.}$$

Знайдемо структурні середні.

Медіанним є інтервал 5,5—10,5, оскільки його нагромаджена частота перевищує 0,5 n. За формулою (2.11) маємо

$$Me=5,5+\frac{25-17}{15}\cdot 5=8,167 \text{ млрд грн.}$$

Для обчислення моди потрібно знайти модальний інтервал, у даному разі той, що має найбільшу щільність $\frac{n_i}{h_i}$. Неважко переконатись, що мода міститься на інтервалі 2,5—3,5. Тому, замінивши у формулі (2.10) частоти на щільності, дістаємо

$$Mo=2,5+\frac{7-0}{2\cdot 7-0-5}\cdot 1=3,278 \text{ млрд грн.}$$

Взаємне розташування одержаних середніх величин таке:

$$Mo< Me< \bar{x}_{\text{в}},$$

що свідчить про правобічну асиметрію розподілу. Для того, щоб кількісно охарактеризувати асиметрію, далі буде знайдено її числовий показник.

Щоб зрозуміти, наскільки широко розсіяні значення ознаки навколо середньої величини і чи є однорідною дана сукупність, знайдемо показники варіації.

Дисперсія та середньоквадратичне відхилення:

$$D_{\text{в}}=\overline{x_{\text{в}}^2}-(\overline{x_{\text{в}}})^2=\frac{19950}{50}-13,8^2=208,56;$$

$$\sigma_{\text{в}}=\sqrt{D_{\text{в}}}=\sqrt{208,56}=14,44 \text{ млрд грн;}$$

коефіцієнт варіації:

$$V=\frac{\sigma_{\text{в}}}{\bar{x}_{\text{в}}}\cdot 100\,\%=\frac{14,44}{13,8}\cdot 100\,\%=104,65\,\%.$$

Оскільки значення коефіцієнта варіації значно більше за 33 %, розглядувана сукупність є достатньо неоднорідною. Справді, у її складі є і відносно невеликі, і великі банки, причому обсяг активів останніх перевищує в десятки разів активи перших.

Знайдемо показники форми розподілу — асиметрію та ексцес, що дасть можливість зробити висновок про близькість емпіричного розподілу до нормального. Для цього визначимо центральні моменти третього та четвертого порядку, використовуючи відповідні початкові моменти:

$$\tilde{v}_1=13,8;$$

$$\tilde{v}_2=\frac{19950}{50}=399;$$

$$\tilde{v}_3=\frac{876866,25}{50}=17537,325;$$

$$\tilde{v}_4=\frac{45970068,75}{50}=919401,375,$$

а значить,

$$\tilde{\mu}_3 = \tilde{v}_3 - 3\tilde{v}_1\tilde{v}_2 + 2\tilde{v}_1^3 = 17537{,}325 - 3\cdot 13{,}8\cdot 399 + 2\cdot 13{,}8^3 = 6274{,}869;$$

$$\tilde{\mu}_4 = \tilde{v}_4 - 4\tilde{v}_1\tilde{v}_3 + 6\tilde{v}_1^2\tilde{v}_2 - 3\tilde{v}_1^4 = 919401{,}375 - 4\cdot 13{,}8\cdot 17537{,}325 +$$

$$+6\cdot 13{,}8^2\cdot 399 - 3\cdot 13{,}8^4 = -298452{,}2.$$

Коефіцієнти асиметрії та ексцесу відповідно дорівнюють:

$$As = \frac{\tilde{\mu}_3}{\sigma_B{}^3} = \frac{6274{,}869}{14{,}44^3} = 2{,}084;$$

$$Es = \frac{\tilde{\mu}_4}{\sigma_B{}^4} - 3 = \frac{-298452{,}2}{14{,}44^4} - 3 = -9{,}86.$$

Значення цих коефіцієнтів свідчать про сильну правобічну асиметрію розподілу та його плосковершинність порівняно з нормальним законом. Тому, можемо зробити висновок, що емпіричний розподіл не є близьким до нормального.

Знайдемо емпіричний коефіцієнт детермінації $K_д = \frac{D_9}{D_1}$. Для цього обчислимо перший і дев'ятий децилі:

$$D_1 = 2{,}5 + \frac{5-0}{7}\cdot 1 = 3{,}214 \text{ млрд грн};$$

$$D_9 = 20{,}5 + \frac{45-41}{6}\cdot 20 = 33{,}833 \text{ млрд грн}.$$

Отже,

$$K_д = \frac{D_9}{D_1} = \frac{33{,}833}{3{,}214} \approx 10{,}5 \text{ раза}.$$

Знайдений коефіцієнт показує, як співвідносяться мінімальний обсяг активів 10 % найбільших банків і максимальний обсяг активів 10 % найменших.

§ 3. ТОЧКОВІ ОЦІНКИ ПАРАМЕТРІВ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

3.1. ОСНОВНІ ПОНЯТТЯ І ЗАГАЛЬНІ ВИМОГИ

Нехай досліджується випадкова величина *X*. У багатьох задачах, що виникають на практиці, вид закону розподілу досліджуваної випадкової величини відомий, і потрібно знайти лише параметри, від яких він залежить. Наприклад, відомо, що похибки в розмірах деталей, що їх виробляє верстат-автомат, або похибки вимірювання різних величин розподілені за нормальним законом. У цих випадках задачею обробки статистичних даних є визначення двох його параметрів — математичного сподівання *a* та середнього квадратичного відхилення σ. У багатьох фізичних, економічних і біологічних процесах теоретичним законом розподілу є експоненціальний закон, і потрібно визначити один його параметр λ. У деяких випадках закон розподілу загалом несуттєвий, і потрібно знати тільки його числові характеристики.

З генеральної сукупності зроблено вибірку обсягу n результати якої $x_1, x_2, \ldots, x_n$. Значення $x_1, x_2, \ldots, x_n$, які спостерігались у конкретній вибірці, можна тлумачити як реалізації n-вимірної випадкової величини $(X_1, X_2, \ldots, X_n)$, компоненти якої X_i можна розглядати як n екземплярів випадкової величини X, тобто n незалежних випадкових величин, кожна з яких розподілена за тим самим законом, що й випадкова величина X.

Кожну характеристику, одержану на основі даних вибірки, також слід розглядати як значення деякої випадкової величини, що в різних вибірках може набувати різних значень. Наприклад, середню вибіркову $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i$ можна розглядати як середнє значення випадкової величини $\bar{X}_в = \frac{1}{n}\sum_{i=1}^{n} X_i$ Надалі в теоретичних міркуваннях, там де не виникає плутанини, ми будемо говорити про $\bar{x}_в$ саме як про випадкову величину $\bar{X}_в$, не запроваджуючи для цього спеціального позначення. Аналогічні зауваження стосуються й усіх інших вибіркових характеристик, тобто в теоретичних міркуваннях вони розглядатимуться як випадкові величини, а не як конкретні значення цих випадкових величин, одержані з конкретної вибірки.

Будь-яка функція від результатів вибірки $\eta(x_1, x_2, \ldots, x_n)$ називається *статистикою*. Статистиками є, наприклад, вибіркова середня, вибіркова дисперсія, емпірична функція розподілу, емпіричні моменти розподілу різних порядків.

Нехай закон розподілу випадкової величини X містить k параметрів $\theta_1, \theta_2, ..., \theta_k$. У загальному випадку задача оцінки параметрів розподілу зводиться до знаходження таких статистик $\hat{\theta}_1(x_1, x_2, \ldots, x_n)$, $\hat{\theta}_2(x_1, x_2, \ldots, x_n)$, …, $\hat{\theta}_k(x_1, x_2, \ldots, x_n)$, які можна використати для наближеного визначення невідомих параметрів розподілу. Ці функції $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ називаються *статистичними оцінками параметрів генеральної сукупності*.

Статистичні оцінки параметрів генеральної сукупності $\hat{\theta}_i(X_1, X_2, \ldots, X_n)$ ($i = 1, 2, \ldots, k$), як функції випадкових величин, самі є випадковими величинами. Їхні закони розподілу залежать, по-перше, від закону розподілу $F(x, \theta_1, \theta_2, ..., \theta_k)$ випадкової величини X (тобто від самих параметрів θ_i ($i = 1, 2, \ldots, k$), по-друге, від кількості спостережень n.

До статистичних оцінок параметрів висувається низка вимог, а саме змістовність, незміщеність, ефективність.

Означення. *Оцінка* $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ *невідомого параметра* θ *називається змістовною (ґрунтовною, обґрунтованою, слушною, консистентною), якщо в разі збільшення обсягу вибірки n вона збігається за ймовірністю до оцінюваного параметра, тобто для будь якого* $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n(X_1, X_2, \ldots, X_n) - \theta\right| < \varepsilon\right) = 1 = 1. \tag{3.1}$$

Це означає, що за збільшення обсягу вибірки оцінка чимраз більше наближається до параметра $\hat{\theta}$.

Якщо параметр $\Theta(\theta_1, \theta_2, ..., \theta_k)$ векторний, то для змістовності векторної оцінки $\Theta(\theta_1, \theta_2, ..., \theta_k)$ потрібна змістовність усіх її компонент.

Властивість змістовності необхідна для будь-якої статистичної оцінки, щоб вона мала практичний сенс, адже в протилежному разі збільшення обсягу інформації не поліпшуватиме оцінки.

З іншого боку, властивість змістовності — це *асимптотична* за n властивість (гранична при $n \to \infty$), яка може виявлятись лише за таких великих обсягів вибірки, яких на практиці важко досягти. Вона не накладає жодних обмежень на поведінку оцінки за скінченних n. Якщо існує одна змістовна оцінка $\hat{\theta}_n$ то можна побудувати нескінченну кількість інших змістовних оцінок. Наприклад, за будь-яких скінченних a та b оцінки $\frac{n+a}{n+b}\hat{\theta}_n$ також будуть змістовні, тому що $\frac{n+a}{n+b} \to 1$ при $n \to \infty$ Але, скажімо, при $n = 20$, $a = 5$, $b = 30$ нова оцінка становить лише половину оцінки $\hat{\theta}_n$.

Наведемо інший приклад. Нехай маємо змістовну оцінку $\hat{\theta}_n$. Її можна замінити іншою —

$$\hat{\theta}_n^1 = \begin{cases} 0, \text{якщо } n \le 10^{10} \\ \hat{\theta}_n \text{ якщо } n > 10^{10} \end{cases},$$

яка також змістовна, але від неї немає жодної користі в будь-якій практичній задачі, де обсяг вибірки великий, але не як завгодно великий.

Наведені міркування свідчать про те, що властивості змістовності недостатньо для характеристики оцінки. Тому її треба доповнити іншими властивостями.

Другою вимогою, яка висувається до статистичної оцінки, є її *незміщеність*, тобто відсутність у оцінки *систематичної похибки*.

Означення. *Статистична оцінка називається незміщеною, якщо її математичне сподівання дорівнює оцінюваному параметру за будь-якого обсягу вибірки, тобто*

$$M\left(\hat{\theta}_n(X_1, X_2, \ldots, X_n)\right) = \theta. \tag{3.2}$$

Якщо параметр $\Theta\ (\theta_1, \theta_2, ..., \theta_k)$ векторний, то для незміщеності векторної оцінки $\Theta\ (\theta_1, \theta_2, ..., \theta_k)$ потрібна незміщеність усіх її компонент.

Центр розсіювання незміщеної оцінки збігається з точним значенням оцінюваного параметра.

Якщо $M\left(\hat{\theta}_n\ (X_1, X_2, \ldots, X_n)\right) \neq \theta$ то оцінка називається *зміщеною*: при $M\left(\hat{\theta}\right) < \theta$ оцінка називається *від'ємно зміщеною*, при $M\left(\hat{\theta}\right) > \theta$ оцінка називається *додатно зміщеною*. При використанні від'ємно або додатно зміщених оцінок ми будемо робити систематичні похибки в менший або більший бік.

Якщо ми виявили зміщеність оцінки, то, як буде показано далі, її можна легко усунути.

Не кожна змістовна оцінка є незміщеною. Її математичне сподівання за кожного скінченного *n* може не дорівнювати параметру θ, а тільки наближатись до нього за необмеженого зростання обсягу вибірки *n*.

Якщо $M\left(\hat{\theta}_n\ (X_1, X_2, \ldots, X_n)\right) = \theta$ при $n \to \infty$ то оцінка $\hat{\theta}_n$ називається *асимптотично* незміщеною.

Вимога незміщеності оцінки особливо важлива в разі *малого обсягу вибірки*. У розпорядженні дослідника часто є дані щодо великої кількості спостережень з малими обсягами вибірки. Для того щоб у результаті усереднення одержати точнішу оцінку, треба бути впевненим, що немає систематичної похибки оцінки, одержаної за кожною вибіркою. Це буде тільки тоді, коли ці оцінки незміщені.

Як уже зазначалось, не кожна змістовна оцінка буде незміщеною. Також можна навести приклади незміщених оцінок, які не є змістовними. Отже, ці властивості не випливають одна з одної. Утім справедлива така теорема.

Теорема 3.1. *Якщо оцінка $\hat{\theta}_n$ параметра θ є незміщеною, тобто $M\left(\hat{\theta}_n\right) = \theta$ і за необмеженого зростання n дисперсія оцінки прямує до нуля, тобто $\lim\limits_{n\to\infty} D\left(\hat{\theta}_n\right) = 0$, то оцінка $\hat{\theta}_n$ є змістовною.*

► Доведення. Запишемо нерівність Чебишова для випадкової величини $\hat{\theta}_n$:

$$P\left(\left|\hat{\theta}_n - M\left(\hat{\theta}_n\right)\right| < \varepsilon\right) \geq 1 - \frac{D\left(\hat{\theta}_n\right)}{\varepsilon^2}.$$

За умовою теореми $M\left(\hat{\theta}_n\right) = \theta$ і $\lim\limits_{n\to\infty} D\left(\hat{\theta}_n\right) = 0$, тому

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1,$$

а це означає, що оцінка $\hat{\theta}_n$ є змістовною. ◄

Приклад 3.1. Нехай X_i — результат *i*-го спостереження під час дослідження випадкової величини *X*. Довести, що $\hat{\theta} = X_i$ є незміщеною, але незмістовною оцінкою генеральної середньої.

Р о з в ' я з а н н я. Математичне сподівання оцінки

$$M\left(\hat{\theta}\right) = M(X_i) = M(X) = \bar{x}_\text{г},$$

тому вона є незміщеною. Покажемо, що ця оцінка незмістовна. По-перше, X_i за фіксованого *i* не залежить від обсягу вибірки і тому її розподіл не змінюється зі збільшенням *n.* А по-друге, завжди знайдеться $\varepsilon > 0$ для випадкової величини $X \neq \text{const}$, таке що

$$P(|X_i - \bar{x}_\text{г}| < \varepsilon) = \int\limits_{\bar{x}_\text{г}-\varepsilon}^{\bar{x}_\text{г}+\varepsilon} f(x)dx \neq 1,$$

тут $f(x)$ — щільність розподілу випадкової величини *X*.

Отже, оцінка $\hat{\theta} = X_i$ не є змістовною.

Для параметра генеральної сукупності можуть існувати кілька змістовних незміщених оцінок. Наприклад, якщо ознака генеральної сукупності розподілена за нормальним законом, то можна довести, що вибіркова середня і медіана вибірки є змістовними незміщеними оцінками середнього значення ознаки генеральної сукупності. Тому потрібні додаткові критерії, які б дозволяли серед змістовних незміщених оцінок вибрати найліпшу. Один з таких критеріїв виникає в результаті дисперсії оцінок.

Якщо дисперсія оцінки $D(\hat{\theta})$ велика, то можливі значення $\hat{\theta}$ сильно розсіяні відносно їхньої середньої величини, і знайдене за результатами вибірки конкретне значення оцінки $\hat{\theta}_1(x_1, x_2, \ldots, x_n)$

може виявитись досить далеким від оцінюваного параметра $\hat{\theta}$. Чим менша дисперсія оцінки, тим більша частина її розподілу сконцентрована біля середнього значення, а отже, і біля самого параметра $\hat{\theta}$. Тож змістовна незміщена оцінка з меншою дисперсією в середньому буде менше відхилятись від істинного значення параметра, ніж оцінка з більшою дисперсією. Тому логічно вважати її ліпшою.

Наприклад, дисперсії вибіркової середньої і медіани вибірки, зробленої з нормально розподіленої сукупності, дорівнюють

$\mathrm{D}(\bar{x}_{\text{в}}) = \dfrac{D_{\text{г}}}{n}$ (для будь якого n), $D(\widetilde{Me}) = \dfrac{\pi}{2}\cdot\dfrac{D_{\text{г}}}{n}$ (для великих n).

Множник $\dfrac{\pi}{2} = 1{,}57 > 1$, тому за великих значень n середня вибіркова є ефективнішою оцінкою, ніж медіана (такий самий висновок справедливий і для малих значень n).

О з н а ч е н н я. *Незміщена оцінка $\hat{\theta}_n$ параметра θ називається ефективною, якщо вона має найменшу дисперсію серед усіх можливих незміщених оцінок параметра θ, одержаних за вибірками однакового обсягу n.*

3.2. ТОЧКОВІ ОЦІНКИ МАТЕМАТИЧНОГО СПОДІВАННЯ ТА ДИСПЕРСІЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

Нехай досліджується випадкова величина X_i з математичним сподіванням $M(X) = \bar{x}_{\text{г}} = \mu$ та дисперсією $D(X) = D\text{г} = \sigma^2$. Обидві числові характеристики невідомі, і потрібно знайти їхні точкові оцінки.

Розглянемо повторну вибірку $X_1, X_2, \ldots, X_n$ обсягу n з генеральної сукупності. Тут X_i — випадкова величина, яка характеризує значення ознаки X у i-го елемента вибірки ($i = 1,2, \ldots, n$). Якщо вибірка повторна, то випадкові величини X_i незалежні і розподілені за тим самим законом, що й досліджувана випадкова величина X. Тому

$$M(X_1) = M(X_2) = \ldots = M(X_n) = \mu, \tag{3.3}$$

$$D(X_1) = D(X_2) = \ldots = D(X_n) = \sigma^2. \tag{3.4}$$

Теорема 3.2. *Вибіркова середня $\bar{x}_{\text{в}}$ повторної вибірки є незміщеною змістовною оцінкою математичного сподівання генеральної сукупності (генеральної середньої $\bar{x}_{\text{г}}$).*

► Доведення. Доведемо незміщеність оцінки. Для цього знайдемо математичне сподівання середньої вибіркової. Ураховуючи (3.3), маємо

$$M(\bar{x}_{\text{в}}) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

Отже, вибіркова середня $(\bar{x}_{\text{в}})$ є незміщеною оцінкою генеральної середньої $\bar{x}_{\text{г}}$.

Доведемо змістовність оцінки. Знайдемо дисперсію вибіркової середньої:

$$D(\bar{x}_{\text{в}}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}.$$

Дисперсія оцінки $D(\bar{x}_{\text{в}}) \to 0$ при $n \to \infty$, тому за теоремою 3.1 середня вибіркова $\bar{x}_{\text{в}}$ є змістовною оцінкою генеральної середньої $\bar{x}_{\text{г}}$. ◄

Тож вибіркова середня $\bar{x}_{\text{в}}$ повторної вибірки є незміщеною змістовною оцінкою генеральної середньої $\bar{x}_{\text{г}}$. Дисперсія та середнє квадратичне відхилення цієї оцінки дорівнюють:

$$D(\bar{x}_{\text{в}}) = \frac{D_{\text{г}}}{n} = \frac{\sigma^2}{n}; \tag{3.5}$$

$$\sigma(\bar{x}_{\text{в}}) = \sqrt{\frac{D_{\text{г}}}{n}} = \frac{\sigma}{\sqrt{n}}. \tag{3.6}$$

Розглянемо тепер безповторну вибірку обсягу n з генеральної сукупності обсягу N. Тоді випадкові величини $X_1, X_2, \ldots, X_n$ є залежними, але й у цьому разі справедлива теорема, яку наведемо без доведення.

Теорема 3.3. *Вибіркова середня $\bar{x}_в$ безповторної вибірки є незміщеною змістовною оцінкою математичного сподівання генеральної сукупності* (*генеральної середньої $\bar{x}_г$*). *Дисперсія та середнє квадратичне відхилення цієї оцінки дорівнюють:*

$$D(\bar{x}_в) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}; \tag{3.7}$$

$$\sigma(\bar{x}_в) = \sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}}. \tag{3.8}$$

Як бачимо, середнє квадратичне відхилення оцінки в повторній і безповторній вибірках відрізняються лише множником

$$C = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1-\frac{n}{N}}{1-\frac{1}{N}}} \approx \sqrt{1-\frac{n}{N}}, \tag{3.9}$$

який менший від одиниці, тобто похибка оцінки безповторної вибірки менша від похибки оцінки повторної вибірки. Але ця відмінність здебільшого не має великого практичного значення, тому що зазвичай відношення $\frac{n}{N}$ обсягу n вибірки до обсягу N генеральної сукупності мале, і множник C близький до одиниці.

Так, за 10 %-вої вибірки $\frac{n}{N} = 0{,}1$ і $\sqrt{0{,}9} \approx 0{,}95 < C < 1,$ тому $\sigma(\bar{x}_в)$ у безповторній вибірці становить від 0,95 до 1 свого значення у повторній вибірці.

Взагалі точність оцінки для $\bar{x}_г$ за $\bar{x}_в$, як випливає з (3.8), мало залежить від обсягу N генеральної сукупності і набагато більше залежить від обсягу вибірки n.

За зростання обсягу вибірки n точність оцінки зростає приблизно в $\sqrt{n}$ разів, адже саме у стільки раз зменшується середнє квадратичне відхилення $\sigma(\bar{x}_в)$ Така залежність $\sigma(\bar{x}_в)$ від n пояснює той факт, що обсяги вибірок, які використовуються на практиці, звичайно невеликі. Адже незначне дальше їх збільшення мало впливає на точність оцінки, а суттєве збільшення точності вимагає незіставно великих обсягів вибірки.

Приклад 3.2. У цех надійшла партія з 2000 деталей. Для контролю їхньої якості було відібрано 120 деталей. Середній діаметр деталі у вибірці виявився $\bar{x}_в = 32$ мм. Треба порівняти середні квадратичні відхилення $\sigma(\bar{x}_в)$ для повторної та безповторної вибірок, якщо σ = 0,25 мм.

Р о з в ’ я з а н н я. Для повторної вибірки за формулою (3.6) маємо

$$\sigma(\bar{x}_в) \approx \frac{0{,}25}{\sqrt{120}} = 0{,}228 \text{ мм}.$$

Для безповторної вибірки за формулою (3.8) дістанемо

$$\sigma(\bar{x}_в) \approx \sqrt{\frac{2000-120}{2000-1}} \cdot \frac{0{,}25}{\sqrt{120}} = 0{,}221 \text{ мм}.$$

У даному прикладі різниця між цими значеннями становить 3 %.

Можна показати, що в разі нормально розподіленої генеральної сукупності оцінка $\bar{x}_в$ буде ефективною оцінкою $\bar{x}_г$.

Знайдемо оцінку дисперсії генеральної сукупності. На перший погляд такою оцінкою буде вибіркова дисперсія.

Розглянемо спочатку повторну вибірку. Подамо вираз для вибіркової дисперсії за формулою (2.21), в якій візьмемо $C = \mu$:

$$D_в = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{x}_в - \mu)^2.$$

Знайдемо математичне сподівання вибіркової дисперсії

$$M(D_в) = M\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) - M(\bar{x}_в - \mu)^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}M(X_i - \mu)^2 - D(\bar{x}_в) = \left| \begin{array}{l} \text{Оскільки } M(X_i - \mu)^2 = D_г \text{ і за формулою (3,5)} \\ M(\bar{x}_в - \mu)^2 = D(\bar{x}_в) = \dfrac{D_г}{n} = \end{array} \right|$$

$$= \frac{1}{n} \cdot nD_r - \frac{D_г}{n} = \frac{n-1}{n}D_г .$$

Отже, остаточно

$$M(D_в) = \frac{n-1}{n}D_г . \tag{3.10}$$

З рівності (3.10) випливає, що вибіркова дисперсія є зміщеною точковою статистичною оцінкою дисперсії генеральної сукупності $\bar{x}_г$. Множник $\frac{n-1}{n} < 1$ тому в разі використання вибіркової дисперсії як оцінки дисперсії генеральної сукупності ми матимемо систематичну похибку в менший бік. Для того щоб ліквідувати цю похибку, уводиться поняття ***виправленої дисперсії*** —

$$s^2 = \frac{n}{n-1}D_в \tag{3.11}$$

Теорема 3.4. *Виправлена дисперсія s^2 повторної вибірки є незміщеною змістовною оцінкою дисперсії генеральної сукупності* $D_г$.

► Доведення. Доведемо незміщеність оцінки. Знайдемо математичне сподівання виправленої дисперсії

$$M(s^2) = M\left(\frac{n}{n-1}D_в\right) = \frac{n}{n-1}M(D_в) = \frac{n}{n-1} \cdot \frac{n-1}{n}D_г = D_г .$$

Звідси випливає, що виправлена дисперсія s^2 повторної вибірки є незміщеною оцінкою дисперсії генеральної сукупності $D_г$.

Змістовність оцінки приймемо без доведення. ◄

Для безповторної вибірки, коли випадкові величини $X_1, X_2, \ldots, X_n$ залежні, можна довести, що

$$M(D_в) = \frac{n-1}{n} \cdot \frac{N}{N-1}D_г , \tag{3.12}$$

де N — обсяг генеральної сукупності.

Множник $\frac{N}{N-1}$ частіше за все дуже близький до одиниці, тому для безповторної вибірки $M(D_в) = \frac{n-1}{n}D_г$ і справедливі всі висновки, зроблені раніше для повторної вибірки.

Отже, як для повторної, так і для безповторної вибірки незміщеною змістовною статистичною оцінкою дисперсії генеральної сукупності є виправлена дисперсія, яка обчислюється за формулами:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2 , \tag{3.13}$$

або

$$s^2 = \frac{n}{n-1}\left(\overline{x^2} - (\bar{x}_в)^2\right). \tag{3.14}$$

Величина

$$s = \sqrt{\frac{n}{n-1}D_в} \tag{3.15}$$

називається виправленим середнім квадратичним відхиленням.

Виправлене середнє квадратичне відхилення є зміщеною статистичною оцінкою середнього квадратичного відхилення σ генеральної сукупності.

Зазначимо, що за великих значень n різниця між вибірковою дисперсією $D_в$ і виправленою дисперсією s^2 дуже мала, і вони практично однакові, тому виправлену дисперсію s^2 використовують за малих вибірок ($n \le 30$).

Зауваження. Якщо математичне сподівання $M(X) = \mu$ випадкової величини X відоме, то **незміщеною оцінкою дисперсії генеральної сукупності є статистика**

$$D_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 . \tag{3.16}$$

Справді,

$$M(D_0) = M\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(M(X_i - \mu)^2\right) = \frac{1}{n}\cdot nD_г = D_г .$$

Можна довести, що оцінка D_0 є змістовною.

Приклад 3.3. Результати вибірки подано у вигляді статистичного ряду

x_i	2,71	3,83	4,95	6,07	7,19	8,31	9,43
n_i	2	3	4	8	5	4	1

Знайти змістовні незміщені оцінки генеральної середньої та дисперсії генеральної сукупності.

Р о з в ' я з а н н я. Оскільки змістовною незміщеною оцінкою генеральної середньої $\bar{x}_г$ є вибіркова середня $\bar{x}_в$ то обчислюємо

$$\bar{x}_в = \frac{1}{27}(2,71\cdot 2 + 3,83\cdot 3 + 4,95\cdot 4 + 6,07\cdot 8 + 7,19\cdot 5 + 8,31\cdot 4 + 9,43\cdot 1) = 6,07.$$

Для визначення змістовної незміщеної оцінки дисперсії генеральної сукупності обчислимо вибіркову дисперсію

$$D_в = \overline{x^2} - (\bar{x}_в)^2 = \frac{1}{27}(2,71^2\cdot 2 + 3,83^2\cdot 3 + 4,95^2\cdot 4 + 6,07^2\cdot 8 +$$
$$+7,19^2\cdot 5 + 8,31^2\cdot 4 + 9,43^2\cdot 1) - 6,07^2 = 38,8183 - 36,8449 = 1,9734.$$

Тоді за формулою (3.11) змістовною незміщеною оцінкою дисперсії генеральної сукупності буде

$$s^2 = \frac{n}{n-1}D_в = \frac{27}{27-1}\cdot 1,9734 = 2,0493.$$

Можна довести, що вибіркові початкові та центральні моменти є змістовними оцінками відповідних моментів генеральної сукупності в разі, якщо останні існують. Але ці оцінки, крім вибіркової середньої, яка є початковим моментом першого порядку, зміщені.

Насамкінець наведемо дві теореми.

Теорема 3.5. *Відносна частота* $\frac{m}{n}$ *появи події A у n незалежних випробуваннях є незміщеною, змістовною та ефективною оцінкою ймовірності P(A)=p появи події A в кожному випробуванні.*

► Доведення. Змістовність оцінки $\hat{\theta} = \frac{m}{n}$ безпосередньо випливає з теореми Бернуллі.

Доведемо незміщеність оцінки $\hat{\theta} = \frac{m}{n}$ Кількість успіхів m — випадкова величина, розподілена за біноміальним законом. Її математичне сподівання дорівнює $M(m) = np$ Знайдемо

$$M\left(\hat{\theta}\right) = M\left(\frac{m}{n}\right) = \frac{1}{n} = M(m) = \frac{1}{n}np = p .$$

Отже, оцінка $\hat{\theta} = \frac{m}{n}$ незміщена.

Ефективність оцінки буде доведена далі (див. приклад 3.7). ◄

Теорема 3.6. *Емпірична функція розподілу вибірки* $F^*(x)$ *є незміщеною змістовною оцінкою функції розподілу* $F(x)$ *випадкової величини X.*

Теорему приймемо без доведення.

3.3. ЕФЕКТИВНІСТЬ ТОЧКОВИХ ОЦІНОК

У пункті 3.1 було введене поняття ефективної оцінки. Зрозуміло, що, наприклад, з двох незміщених оцінок $\hat{\theta}_1$ і $\hat{\theta}_2$ параметра θ ліпшого є та, варіація якої менша. Але виникають питання: наскільки добрими є оцінки $\hat{\theta}_1$ та $\hat{\theta}_2$ порівняно з нескінченною кількістю інших незміщених оцінок і чи можна відшукати ту з них, яка буде ефективною для даного параметра, та знайти її дисперсію? Яким способом можна ввести кількісний показник ефективності оцінки $\hat{\theta}$ параметра θ, який дасть змогу побудувати шкалу незміщених оцінок у плані їхньої ефективності? Відповіді на поставлені питання дає теорема, яка відома в статистиці як *нерівність Рао-Крамера-Фреше.* Для того щоб сформулювати її, уведемо деякі поняття та позначення.

Нехай для дослідження неперервної випадкової величини X, щільність розподілу якої $f(x;\ \theta)$ залежить від одного параметра, зроблено вибірку $x_1, x_2, \ldots, x_n$. Її результати є реалізаціями незалежних, однаково розподілених випадкових величин $X_1, X_2, \ldots, X_n$ тому щільність їх сумісного розподілу буде

$$L(x_1, x_2, \ldots, x_n, \theta) = f(x_1,\ \theta) \cdot f(x_2,\ \theta) \cdot \ldots f(x_n,\ \theta) = \prod_{i=1}^{n} f(x_i, \theta). \quad (3.17)$$

Ця функція називається функцією правдоподібності.

У разі дискретної випадкової величини функція правдоподібності має вигляд:

$$L(x_1, x_2, \ldots, x_n, \theta) = p(x_1,\ \theta) \cdot p(x_2,\ \theta) \cdot \ldots p(x_n,\ \theta) = \prod_{i=1}^{n} p(x_i,\ \theta), \quad (3.18)$$

де $p(x_i,\ \theta) = P(X = x_i,\ \theta)$.

Функція, визначена рівностями (3.17) та (3.18), задає ймовірність одержання під час здійснення вибірки обсягу n саме значень $x_1, x_2, \ldots, x_n$ у дискретному випадку, або величину, пропорційну ймовірності одержання вибіркових значень близьких до $x_1, x_2, \ldots, x_n$ у неперервному випадку. Тому чим більше значення функції $L(x_1, x_2, \ldots, x_n, \theta)$ за конкретного θ, тим імовірнішою (або більш *правдоподібною*) є поява вибіркових значень $x_1, x_2, \ldots, x_n$. Звідси й назва функції $L(x_1, x_2, \ldots, x_n, \theta)$ — *функція правдоподібності*.

Цікаво простежити, як сильно впливає зміна параметра θ на ймовірності (3.17), (3.18), або, інакше кажучи, наскільки чутливими є результати проведеного спостереження до параметра. Очевидно, що чим більш різко змінюється функція правдоподібності залежно від θ, тим більше інформації про параметр міститься у конкретних результатах вибірки. Тобто інформацію про невідомий параметр θ можна інтерпретувати як міру невизначеності, що стосується його значення, після проведеного спостереження над випадковою величиною. Наприклад, якщо за спостережуваними значеннями $x_1, x_2, \ldots, x_n$ випадкових величин $X_1, X_2, \ldots, X_n$ можна з імовірністю рівною 1 встановити, чому дорівнює параметр θ, то це значить, що випадкова величина містить максимально можливу інформацію про параметр. І навпаки, якщо розподіл випадкової величини не залежить від параметра θ, то це означає, що немає інформації для знаходження його значення на основі вибірки (ситуація з нульовою інформацією). Для того щоб увести міру такої чутливості результатів спостереження до невідомого параметра, сформулюємо таке означення.

О з н а ч е н н я. *Функцією впливу (або функцією внеску) вибірки* $x_1, x_2, \ldots, x_n$ *називається частинна похідна за параметром* θ *від логарифма функції* $L(x_1, x_2, \ldots, x_n, \theta)$ *тобто*

$$U(x_1, x_2, \ldots, x_n, \theta) = \frac{\partial}{\partial\theta} \ln L(x_1, x_2, \ldots, x_n, \theta).$$

Далі вважатимемо, що функція правдоподібності L відповідає певним вимогам регулярності (див., наприклад, [12, с. 363]), які дозволяють диференціювати за параметром під знаком інтеграла, що залежить від параметра.

Розглянемо спочатку випадок неперервної випадкової величини.

Оскільки $L(x_1, x_2, \ldots, x_n, \theta)$ є щільністю розподілу системи випадкових величин $X_1, X_2, \ldots, X_n$, то

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(x_1, x_2, \ldots, x_n, \theta)\, dx_1, \ldots, dx_n = 1. \tag{3.19}$$

За умов регулярності існують перші дві похідні по θ функції правдоподібності L, і для їх знаходження можна змінювати порядок диференціювання та інтегрування. Тому після диференціювання обох частин рівності (3.19) по θ дістанемо:

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} L(x_1, x_2, \ldots, x_n, \theta) dx_1, \ldots, dx_n = 0.$$

Звідси виходить, що функція впливу центрована за будь-якого значення θ, тобто

$$M(U(x_1, x_2, \ldots, x_n, \theta)) = M\left(\frac{\partial \ln L}{\partial\theta}\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{1}{L}\cdot\frac{\partial L}{\partial\theta}\right) L dx_1, \ldots, dx_n = 0. \tag{3.20}$$

Характеристикою, на основі якої вимірюють швидкість зміни розподілу $L(x_1, x_2, \ldots, x_n, \theta)$ за невеликої зміни значення параметра θ, є функція, яка називається *кількістю інформації Фішера*, що міститься у вибірці:

$$I(\theta) = D\big(U(x_1, \ldots, x_n, \theta)\big) = M\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right]; \tag{3.21}$$

Зазначимо, що друга частина рівності у (3.21) випливає з (3.20) та з властивості дисперсії.

Знайдемо похідну обох частин рівності (3.20) по θ. Змінивши порядок диференціювання та інтегрування, дістанемо рівність

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\left(\frac{1}{L}\cdot\frac{\partial L}{\partial\theta}\right)\frac{\partial L}{\partial\theta} + L\frac{\partial}{\partial\theta}\left(\frac{1}{L}\cdot\frac{\partial L}{\partial\theta}\right)\right] dx_1, \ldots, dx_n = 0,$$

яку можна переписати у вигляді

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\left(\frac{1}{L}\cdot\frac{\partial L}{\partial\theta}\right)^2 + \frac{\partial^2 \ln L}{\partial\theta^2}\right] L dx_1 \ldots dx_n = 0.$$

Звідси випливає тотожність, яку зручно використовувати для обчислення кількості інформації $I(\theta)$:

$$M\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right] = -M\left(\frac{\partial^2 \ln L}{\partial\theta^2}\right). \tag{3.22}$$

Величина

$$i(\theta) = M\left[\left(\frac{\partial \ln f(X,\theta)}{\partial\theta}\right)^2\right] = \int_{-\infty}^{\infty}\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)dx \tag{3.23}$$

називається *кількістю інформації Фішера в одному спостереженні.*

Для дискретної випадкової величини X, яка може набувати значень x_k, $k = 1, 2, \ldots, l$ з імовірностями $p(x_k\ \theta) = P(X = x_k, \theta)$ кількість інформації Фішера *в одному спостереженні*

$$i(\theta) = M\left[\left(\frac{\partial \ln p(X,\theta)}{\partial\theta}\right)^2\right] = \sum_{k=1}^{l}\left(\frac{\partial \ln p(x_k,\theta)}{\partial\theta}\right)^2 p(x_k,\theta). \tag{3.24}$$

Ураховуючи те, що $X_1, X_2, \ldots, X_n$ однаково розподілені й незалежні, дістанемо

$$I(\theta) = n \cdot i(\theta) . \tag{3.25}$$

Сформулюємо основну в цьому підрозділі теорему.

Теорема 3.7. (*Нерівність Рао–Крамера–Фреше*). *Нехай параметрична модель, що розглядається, відповідає вимогам регулярності й* $\hat{\theta}\,(x_1, \ldots, x_n)$ *— незміщена оцінка невідомого параметра* θ. *Тоді існує нерівність*

$$D(\hat{\theta}) \geq \frac{1}{I(\theta)}, \tag{3.26}$$

де $I(\theta)$ *— кількість інформації Фішера що міститься у вибірці.*

Виходячи з рівностей (3.22) і (3.25) нерівність (3.26) можна подати так:

$$D(\hat{\theta}) \geq \frac{1}{M\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]}, D(\hat{\theta}) \geq \frac{1}{M\left[\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^2\right]} \text{ або } D(\hat{\theta}) \geq \frac{1}{ni(\theta)}. \tag{3.27}$$

Отже, нерівність (3.26) визначає нижню межу дисперсій незміщених оцінок параметра θ для регулярних моделей. З означення ефективної оцінки випливає: якщо існує незміщена оцінка $\hat{\theta}$, для якої нерівність (3.26) перетворюється на рівність, то вона ефективна для параметра θ. Тобто нерівність Рао—Крамера—Фреше дає змогу ввести кількісну міру ефективності незміщених оцінок для регулярних моделей. Природно визначити показник ефективності $e(\hat{\theta})$ оцінки $\hat{\theta}$ невідомого параметра θ як відношення мінімально можливого значення дисперсії, що задається правою частиною нерівності (3.26), до дисперсії даної конкретної оцінки $\hat{\theta}$, тобто

$$e(\hat{\theta}) = \frac{1}{I(\theta) D\left(\hat{\theta}(x_1, \ldots, x_n)\right)}.$$

Тож у разі виконання висунутих вимог оцінка $\hat{\theta}$ ефективна, якщо $e(\hat{\theta}) = 1$. Очевидно, що для будь-якої незміщеної оцінки $\hat{\theta}$ параметра θ виконується нерівність $0 < e(\hat{\theta}) \leq 1$.

Якщо $e(\hat{\theta}) < 1$ і $e(\hat{\theta}) = e\left(\hat{\theta}(x_1, \ldots, x_n)\right) \to 1$ при $n \to \infty$, то оцінка $\hat{\theta}$ називається *асимптотично ефективною*.

Зауваження. Нерівність (3.26) перетворюється на рівність тоді й тільки тоді, коли функція впливу є лінійною функцією оцінки, тобто

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)(\hat{\theta} - \theta), \tag{3.28}$$

де перший множник $A(\theta)$ не залежить від результатів вибірки $x_1, x_2, \ldots, x_n$,. Отже, рівність (3.28) є *критерієм ефективності для регулярних моделей.*

У разі виконання умови (3.28) дістанемо

$$I(\theta) = M\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = M\left(\left(A(\theta)(\hat{\theta} - \theta)\right)^2\right) = (A(\theta))^2 D(\hat{\theta}) . \tag{3.29}$$

Оскільки за умови (3.28) нерівність (3.26) перетворюється на рівність, то, підставивши в неї (3.29), дістанемо вираз для дисперсії ефективної оцінки

$$D(\hat{\theta}) = \frac{1}{A(\theta)} . \tag{3.30}$$

Приклад 3.4. Оцінити параметр θ для нормально розподіленої випадкової величини, щільність розподілу якої

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\theta)^2}{2\sigma^2}}, -\infty < x < +\infty,$$

якщо середнє квадратичне відхилення σ відоме.

Р о з в ' я з а н н я. Побудуємо функцію правдоподібності (3.17):

$$L(x_1, x_2, \ldots, x_n, \theta) = \frac{1}{\sigma^n \sqrt{(2\pi)^n}} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}}.$$

Її логарифм

$$\ln L = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}}\right) = -n\ln\sigma - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 .$$

Оскільки частинна похідна

$$\frac{\partial \ln L}{\partial\theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) = \frac{n}{\sigma^2}(\bar{x}_в - \theta),$$

має вигляд (3.28), то, узявши $\hat{\theta} = \bar{x}_в$, $A(\theta) = \frac{n}{\sigma^2}$ дістанемо, що $\bar{x}_в$ буде оцінкою параметра θ з мінімально можливою дисперсією. З рівності (3.30) випливає, що дисперсія цієї оцінки дорівнює $\frac{\sigma^2}{n}$.

Отже, ми довели, що $\bar{x}_в$ є ефективною оцінкою математичного сподівання нормально розподіленої випадкової величини, якщо середнє квадратичне відхилення σ відоме.

Приклад 3.5. Вибірку обсягу *n* зроблено з нормально розподіленої випадкової величини, щільність розподілу якої

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, -\infty < x < +\infty.$$

Знайти кількість інформації Фішера, яка міститься у вибірці, та довести, що $\bar{x}_в$ є ефективною оцінкою математичного сподівання генеральної сукупності.

Р о з в ' я з а н н я. Обчислимо кількість інформації Фішера в одному спостереженні $i(\theta)$. Для цього знайдемо логарифм щільності та його частинні похідні по θ:

$$\ln f = -\ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x-\theta)^2}{2\sigma^2};\ \frac{\partial \ln f}{\partial\theta} = \frac{x-\theta}{\sigma^2};\ \frac{\partial^2 \ln f}{\partial\theta^2} = -\frac{1}{\sigma^2}.$$

Кількість інформації Фішера в одному спостереженні

$$i(\theta) = -M\left(\frac{\partial^2 \ln f}{\partial\theta^2}\right) = -M\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

Кількість інформації Фішера, яка міститься у вибірці,

$$I(\theta) = ni(\theta) = \frac{n}{\sigma^2}.$$

Запишемо нерівність Рао—Крамера—Фреше (3.26):

$$D(\hat{\theta}) \ge \frac{1}{I(\theta)} = \frac{\sigma^2}{n}.$$

Тож, мінімальне можливе значення дисперсії оцінки дорівнює

$$D(\hat{\theta})_{\min} = \frac{\sigma^2}{n}.$$

Дисперсія $D(\bar{x}_{\text{в}}) = \frac{\sigma^2}{n}$ дорівнює $D(\hat{\theta})_{\min}$ тому $\hat{\theta} = \bar{x}_{\text{в}}$ є ефективною оцінкою математичного сподівання генеральної сукупності.

Приклад 3.6. Оцінити параметр θ для випадкової величини, розподіленої за законом Пуассона $P(X = k) = \frac{\theta^k}{k!} e^{-\theta}, k = 0, 1, 2, 3, ...$

Р о з в ' я з а н н я. Побудуємо функцію правдоподібності (3.18)

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta} = \frac{e^{-n\theta}}{x_1! \ldots x_n!} \theta^{\sum_{i=1}^{n} x_i}$$

та обчислимо її логарифм

$$\ln L = -n\theta + \ln\theta \cdot \sum_{i=1}^{n} x_i - \ln(x_1! \ldots x_n!).$$

Частинна похідна

$$\frac{\partial \ln L}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i = \frac{n}{\theta}(\bar{x}_{\text{в}} - \theta)$$

має вигляд (3.28), тому $\bar{x}_{\text{в}}$ є ефективною оцінкою параметра θ для випадкової величини, розподіленої за законом Пуассона. З рівності (3.30) випливає, що дисперсія цієї оцінки дорівнює $\frac{\theta}{n}$.

Приклад 3.7. Проведено n незалежних повторних випробувань Бернуллі. Довести, що відносна частота $\frac{m}{n}$ появи події A є ефективною оцінкою ймовірності p появи події A в одному випробувань.

Р о з в ' я з а н н я. Розглянемо одне випробування Бернуллі. Випадкова величина X_k може набувати значень 0 або 1. Закон її розподілу

$$p(x_k, \theta) = P(X = x_k, \theta) = \theta^{x_k} (1 - \theta)^{1 - x_k}, \; x_k = 0, 1.$$

Обчислимо кількість інформації Фішера в одному випробуванні $i(\theta)$. Для цього знайдемо

$$\ln p(x_k, \theta) = x_k \ln\theta + (1 - x_k) \ln(1 - \theta), \;\; x_k = 0, 1;$$

та

$$\frac{\partial \ln p(x_k, \theta)}{\partial \theta} = \frac{x_k}{\theta} - \frac{1 - x_k}{1 - \theta}, \;\; x_k = 0, 1.$$

За формулою (3.24)

$$i(\theta) = \sum_{k=1}^{n} \left(\frac{\partial \ln p(x_k, \theta)}{\partial \theta} \right)^2 p(x_k, \theta) = \left(\frac{1}{1 - \theta} \right)^2 (1 - \theta) + \left(\frac{1}{\theta} \right)^2 \theta = \frac{1}{\theta(1 - \theta)}.$$

Запишемо нерівність Рао—Крамера—Фреше (3.26):

$$D(\hat{\theta}) \geq \frac{1}{I(\theta)} = \frac{\theta(1 - \theta)}{n}.$$

Отже, мінімальне можливе значення дисперсії оцінки дорівнює

$$D(\hat{\theta})_{\min} = \frac{\theta(1 - \theta)}{n}.$$

Обчислимо дисперсію оцінки $\hat{\theta} = \frac{m}{n}$.

$$D(\hat{\theta}) = D\left(\frac{m}{n}\right) = \frac{1}{n^2}D(m) = \frac{1}{n^2}\cdot n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

Дисперсія $D(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$ дорівнює $D(\hat{\theta})_{\min}$, тому $\hat{\theta} = \frac{m}{n}$ є ефективною оцінкою математичного сподівання генеральної сукупності.

3.4. МЕТОДИ РОЗРАХУНКУ ТОЧКОВИХ СТАТИСТИЧНИХ ОЦІНОК

3.4.1. Метод моментів

Метод моментів був запропонований К. Пірсоном 1894 року.

Нехай досліджується випадкова величина X, закон розподілу якої $f(x;\Theta)$ залежить від параметра, який у загальному випадку може бути векторним $\Theta(\theta_1, \theta_2, ..., \theta_k)$. Тут $f(x;\Theta)$ — щільність розподілу, якщо випадкова величина X неперервна, або ймовірність $P(X = x;\Theta)$. якщо X — дискретна.

Теоретичні моменти, обчислені за законом розподілу $f(x;\Theta)$ випадкової величини X очевидно будуть функціями параметра $\Theta(\theta_1, \theta_2, ..., \theta_k)$. За результатами вибірки (x_1, x_2, …, x_n) можна обчислити емпіричні моменти. Якщо закон розподілу випадкової величини X містить k параметрів $\theta_1, \theta_2, ..., \theta_k$ то, прирівнявши перші k відповідні теоретичні та емпіричні моменти, дістанемо систему k рівнянь з k невідомими параметрами:

$$\int_{-\infty}^{+\infty} x^l f(x;\Theta)dx = \frac{1}{n}\sum_{i=1}^{n} x_i^l,\ \ (l = 1,2,\ldots,k). \tag{3.31}$$

У разі дискретної випадкової величини інтеграли у лівих частинах (3.31) слід замінити відповідними сумами. Розв’язавши систему рівнянь (3.31), дістанемо оцінки $\theta_1, \theta_2, ..., \theta_k$ відповідних параметрів. Питання про те, які саме моменти треба використовувати у системі рівнянь (3.31) — початкові, центральні або їхні модифікації, типу коефіцієнтів асиметрії та ексцесу, треба вирішувати, зважаючи на конкретну мету дослідження та простоту залежності відповідних теоретичних моментів від оцінюваних параметрів $\theta_1, \theta_2, ..., \theta_k$.

Якщо закон розподілу випадкової величини $f(x;\theta)$ залежить від одного параметра, то для його знаходження треба розв’язати одне рівняння

$$M(X) = \bar{x}_{\text{в}}. \tag{3.32}$$

Якщо закон розподілу випадкової величини $f(x;\theta_1,\theta_2)$ залежить від двох параметрів, то для їх знаходження треба розв’язати одну з систем рівнянь:

$$\begin{cases} M(X) = \bar{x}_{\text{в}}, \\ D(X) = D_{\text{в}}; \end{cases} \text{або} \begin{cases} M(X) = \bar{x}_{\text{в}} = v_1, \\ M(X^2) = v_2. \end{cases} \tag{3.33}$$

Оцінки, одержані методом моментів, зазвичай *змістовні*, але їхня ефективність часто значно менша за одиницю. Такі оцінки також можуть бути зміщеними, але часто за допомогою простих поправок зміщення можна усунути. Можна показати, що середнє значення оцінок, одержаних методом моментів, відрізняється від справжнього значення параметра на величину порядку $\frac{1}{n}$, а їхнє середнє квадратичне відхилення $\sigma(\theta)$ асимптотично має вигляд $\frac{c}{\sqrt{n}}$, де c — деяка стала.

Метод моментів часто використовують на практиці завдяки відносній простоті обчислень, пов’язаних з цим методом.

Приклад 3.8. Випадкова величина розподілена за законом Пуассона. Оцінити параметр λ методом моментів.

Р о з в ' я з а н н я. Для знаходження точкової оцінки єдиного параметра λ треба розв'язати рівняння (3.32)

$$M(X) = \bar{x}_{\text{в}}.$$

Ураховуючи, що для випадкової величини, розподіленої за законом Пуассона $M(X) = \lambda$, дістанемо точкову оцінку $\lambda = \bar{x}_{\text{в}}$.

Приклад 3.9. Випадкова величина розподілена за нормальним законом. Оцінити параметри a та σ методом моментів.

Р о з в ' я з а н н я. Для знаходження двох невідомих параметрів прирівняємо теоретичні та емпіричні моменти першого та другого порядків, тобто запишемо систему рівнянь (3.33):

$$\begin{cases} M(X) = \bar{x}_{\text{в}}, \\ D(X) = D_{\text{в}}. \end{cases}$$

Ураховуючи, що для нормально розподіленої випадкової величини $M(X) = a$, $D(X) = \sigma^2$, дістанемо систему рівнянь

$$\begin{cases} a = \bar{x}_{\text{в}}, \\ \sigma^2 = D_{\text{в}}. \end{cases}$$

Отже, точкові оцінки параметрів a та σ нормального розподілу $\hat{a} = \bar{x}_{\text{в}}$ $\hat{\sigma} = \sqrt{D_{\text{в}}}$.

3.4.2. Метод максимальної правдоподібності

З теоретичного погляду найважливішим загальним методом знаходження точкових оцінок параметрів є *метод максимальної правдоподібності*. В окремих випадках цей метод застосовувався ще К.Ф. Гаусом, але як загальний метод знаходження оцінок був запропонований 1912 року Р. Фішером.

Нехай досліджується випадкова величина X, вид закону розподілу якої відомий, але він містить невідомий параметр θ. Цей параметр треба оцінити за результатами вибірки $x_1, x_2, \ldots, x_n$.

Розглянемо функцію правдоподібності, котра для неперервної випадкової величини, щільність розподілу якої $f(x;\theta)$, визначається рівністю (3.17):

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

а для дискретної випадкової величини — рівністю (3.18):

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

де $p(x_i,\theta) = P(X = x_i,\theta)$.

За фіксованих значень $x_1, x_2, \ldots, x_n$. функція правдоподібності залежить від параметра θ. Метод максимальної правдоподібності полягає в тому, що за оцінку невідомого параметра θ генеральної сукупності вибирають таке його значення $\hat{\theta}$, за якого функція правдоподібності L набуває максимального значення.

При розв'язуванні задач зручно від функції правдоподібності L перейти до її логарифма $\ln L$. Функція $\ln L$ називається логарифмічною функцією правдоподібності. Оскільки L набуває максимального значення за того самого значення θ, що й $\ln L$, то для знаходження оцінки потрібно розв'язати *рівняння правдоподібності*

$$\frac{d \ln L}{d\theta} = 0 \tag{3.34}$$

та вибрати такий його розв'язок $\hat{\theta}(x_1, x_2, \ldots, x_n)$, за якого $\ln L$ має максимум. Цей розв'язок називатимемо *оцінкою максимальної правдоподібності* параметра θ.

Оцінки, одержані методом максимальної правдоподібності, за деяких, достатньо широких, умов, накладених на закон розподілу випадкової величини X, є *змістовними*. Вони можуть бути зміщеними, але, як правило, зміщеність можна легко усунути. Крім того, такі оцінки не обов'язково ефективні, однак можна показати, що в разі виконання певних умов (див., наприклад, [18, ст. 544]) вони *асимптотично ефективні*. Також важливість методу максимальної правдоподібності обумовлена теоремою, яку наведемо без доведення.

Теорема 3.8. *Якщо для параметра* θ *існує ефективна оцінка* $\hat{\theta}$ *то рівняння правдоподібності* (3.34) *має єдиний розв'язок* $\hat{\theta}$.

Наведені щойно міркування можна узагальнити на випадок кількох невідомих параметрів розподілу та на випадок вибірок, одержаних із багатовимірних розподілів. Так, для неперервної випадкової величини, щільність розподілу якої $f(x, \theta_1, \theta_2)$·залежить від двох параметрів θ_1 та θ_2, функція правдоподібності має вигляд

$$L(x_1, x_2, \ldots, x_n, \theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2), \tag{3.35}$$

і оцінки максимальної правдоподібності для θ_1 та θ_2 дістанемо, розв'язавши систему рівнянь

$$\begin{cases} \dfrac{\partial \ln L}{\partial \theta_1} = 0, \\ \dfrac{\partial \ln L}{\partial \theta_2} = 0 \end{cases} \tag{3.36}$$

відносно θ_1 і θ_2, та вибравши той її розв'язок ($\hat{\theta}_1(x_1, x_2, \ldots, x_n)$; $\hat{\theta}_2(x_1, x_2, \ldots, x_n)$), за якого $\ln L$ має максимум.

Якщо існує пара сумісно ефективних оцінок $\hat{\theta}_1$ та $\hat{\theta}_2$, то система (3.36) має єдиний розв'язок $\theta_1 = \hat{\theta}_1$, $\theta_2 = \hat{\theta}_2$.

Приклад 3.10. Нехай вибірку $x_1, x_2, \ldots, x_n$ зроблено з генеральної сукупності, ознака X якої розподілена за біноміальним законом $P(X = k) = C_N^k p^k (1-p)^{N-k}$, $k = 0, 1, \ldots, N$, де кількість випробувань N відома. Знайти оцінку ймовірності p появи події в одному випробуванні методом максимальної правдоподібності.

Розв'язання. Біноміальний закон залежить від одного параметра $\theta = p$. Тому $p(x_1, \theta) = P(X = x_i, \theta) = C_N^{x_i} \theta^{x_i} (1-\theta)^{N-x_i}$,. де $x_i = 0, 1, \ldots, N$. Складемо функцію правдоподібності (3.18):

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} p(x_i, \theta) = \prod_{i=1}^{n} C_N^{x_i} \theta^{x_i} (1-\theta)^{N-x_i}.$$

Її логарифм

$$\ln L = \ln \prod_{i=1}^{n} C_N^{x_i} \theta^{x_i} (1-\theta)^{N-x_i} = \sum_{i=1}^{n} \ln\left(C_N^{x_i} \theta^{x_i} (1-\theta)^{N-x_i}\right) =$$

$$= \sum_{i=1}^{n} \left(\ln C_N^{x_i} + x_i \ln\theta + (N - x_i)\ln(1-\theta)\right) =$$

$$= \sum_{i=1}^{n} \ln C_N^{x_i} + \ln\theta \sum_{i=1}^{n} x_i + \ln(1-\theta) \cdot \left(nN - \sum_{i=1}^{n} x_i\right).$$

Знайдемо

$$\frac{d \ln L}{d\theta} = \frac{1}{\theta} \sum_{i=1}^{n} x_i - \frac{1}{1-\theta} \cdot \left(nN - \sum_{i=1}^{n} x_i\right) = \frac{n\bar{x}_{\text{в}}}{\theta} - \frac{n(N - \bar{x}_{\text{в}})}{1-\theta} =$$

$$\frac{n\bar{x}_{\text{в}} - \theta n\bar{x}_{\text{в}} - \theta nN + \theta n\bar{x}_{\text{в}}}{\theta(1-\theta)} = \frac{n\bar{x}_{\text{в}} - \theta nN}{\theta(1-\theta)}.$$

Запишемо рівняння правдоподібності (3.34):

$$\frac{n\bar{x}_{\text{в}} - \theta nN}{\theta(1-\theta)} = 0, \quad \bar{x}_{\text{в}} - \theta N = 0.$$

Звідси $\theta = \frac{\bar{x}_в}{N}$. Доведемо, що при $\theta = \frac{\bar{x}_в}{N}$ логарифм функції правдоподібності ln L має максимум. Для цього знайдемо його другу похідну:

$$\frac{d^2 \ln L}{d\theta^2} = -\frac{n\bar{x}_в}{\theta^2} - \frac{n(N-\bar{x}_в)}{(1-\theta)^2};$$

$$\left.\frac{d^2 \ln L}{d\theta^2}\right|_{\theta=\frac{\bar{x}_в}{N}} = -\frac{n\bar{x}_в N^2}{\bar{x}_в{}^2} - \frac{n(N-\bar{x}_в)}{\left(1-\frac{\bar{x}_в}{N}\right)^2} = -\frac{nN^2}{\bar{x}_в} - \frac{nN^2}{N-\bar{x}_в} = -nN^2\left(\frac{1}{\bar{x}_в} + \frac{1}{N-\bar{x}_в}\right) < 0.$$

Отже, при $\theta = \frac{\bar{x}_в}{N}$ функція ln L має максимум. Тому оцінкою максимальної правдоподібності ймовірності p появи події в одному випробуванні є $\hat{\theta} = \hat{p} = \frac{\bar{x}_в}{N}$.

Приклад 3.11. Нехай вибірку $x_1, x_2, \ldots, x_n$ зроблено з генеральної сукупності, розподіленої за нормальним законом, щільність якого $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$ $-\infty < x < +\infty$. Знайти оцінки параметрів a та σ методом максимальної правдоподібності.

Р о з в ' я з а н н я. Щільність нормального закону залежить від двох параметрів $\theta_1 = a$ та $\theta_2 = \sigma$:

$$f(x, \theta_1\, \theta_2) = \frac{1}{\theta_2\sqrt{2\pi}} e^{\frac{(x-\theta_1)^2}{2\theta_2^2}}.$$

Складемо функцію правдоподібності (3.17):

$$L(x_1, x_2, \ldots, x_n, \theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2) = \frac{1}{\theta_2^n\sqrt{(2\pi)^n}} e^{-\frac{\sum_{l=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2}}.$$

Знайдемо її логарифм

$$\ln L = \ln \frac{1}{\theta_2^n\sqrt{(2\pi)^n}} e^{-\frac{\sum_{l=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2}} = -n\ln\theta_2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2\theta_2^2}\sum_{l=1}^{n}(x_i-\theta_1)^2.$$

Система рівнянь методу максимальної правдоподібності (3.36) має вигляд:

$$\begin{cases} \frac{\partial \ln L}{\partial\theta_1} = \frac{1}{\theta_2^2}\sum_{l=1}^{n}(x_i-\theta_1) = 0, \\ \frac{\partial \ln L}{\partial\theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{l=1}^{n}(x_i-\theta_1)^2 = 0; \end{cases} \Leftrightarrow \begin{cases} \sum_{l=1}^{n} x_i - n\theta_1 = 0, \\ -n\theta_2^2 + \sum_{l=1}^{n}(x_i-\theta_1)^2 = 0; \end{cases} \Leftrightarrow \begin{cases} \theta_1 = \frac{1}{n}\sum_{l=1}^{n} x_i = \bar{x}_в, \\ \theta_2^2 = \frac{1}{n}\sum_{l=1}^{n}(x_i-\bar{x}_в)^2 = D_в. \end{cases}$$

Доведемо, що в точці ($\bar{x}_в; \sqrt{D_в}$) функція ln L має максимум. Обчислимо її другі частинні похідні:

$$\frac{\partial^2 \ln L}{\partial\theta_1^2} = -\frac{n}{\theta_2^2},\quad \frac{\partial^2 \ln L}{\partial\theta_1\partial\theta_2} = -\frac{2}{\theta_2^3}\sum_{l=1}^{n}(x_i-\theta_1) = -\frac{2n(\bar{x}_в-\theta_1)}{\theta_2^3},$$

$$\frac{\partial^2 \ln L}{\partial\theta_2^2} = \frac{n}{\theta_2^2} - \frac{3}{\theta_2^4}\sum_{l=1}^{n}(x_i-\theta_1)^2.$$

Обчислимо вираз $\Delta = AC - B^2$.

$$A = \left.\frac{\partial^2 \ln L}{\partial\theta_1^2}\right|_{(\bar{x}_в;\sqrt{D_в})} = -\frac{n}{D_в},$$

$$B = \frac{\partial^2 \ln L}{\partial\theta_1 \partial\theta_2}\bigg|_{(\bar{x}_в;\sqrt{D_в})} = -\frac{2n(\bar{x}_в - \bar{x}_в)}{\sqrt{D_в^3}} = 0,$$

$$C = \frac{\partial^2 \ln L}{\partial\theta_2^2}\bigg|_{(\bar{x}_в;\sqrt{D_в})} = \frac{n}{D_в} - \frac{3}{D_в^2}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2 = \frac{n}{D_в} - \frac{3}{D_в^2}\cdot nD_в = -\frac{2n}{D_в}.$$

Отже, $\Delta = AC - B^2 = \frac{2n^2}{D_в^2} > 0$ і $A < 0$ тому в точці ($\bar{x}_в;\sqrt{D_в}$) функція $\ln L$ має максимум. Тому оцінками максимальної правдоподібності параметрів нормального закону є $\hat{\theta}_1 = \hat{a} = \bar{x}_в$ та $\hat{\theta}_2 = \hat{\sigma} = \sqrt{D_в}$.

§ 4. ІНТЕРВАЛЬНІ ОЦІНКИ ПАРАМЕТРІВ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

У § 3 було розглянуто методи знаходження *точкових оцінок* невідомих параметрів генеральної сукупності за вибірковими даними (оцінкою математичного сподівання μ сукупності була середня вибіркова $\bar{x}_в$, а дисперсії сукупності σ^2 — виправлена вибіркова дисперсія s^2). Проте стверджувати, що $\mu = M(X) \approx \bar{x}_B$, доволі ризиковано, оскільки $\bar{x}_в$ є точковим значенням випадкової величини $\bar{X}_в$ і для вибірки малого обсягу може значно відрізнятися від μ.

Параметр θ генеральної сукупності можна оцінити двома способами. Як відомо, вибіркова оцінка $\hat{\theta}$, виражена одним числом, тобто точкова оцінка, має суттєві недоліки. Насамперед така оцінка є наближеним значенням параметра. Звідси виникає потреба знати, наскільки оцінка $\hat{\theta}$ відхиляється від оцінюваного параметра θ. Цілком природно сподіватись, що вибірка великого обсягу дасть точніші результати, бо містить більше інформації про оцінювану величину. Надійнішим є інтервальний тип статистичної оцінки, тобто інтервал, ширина якого суттєво залежить від обсягу вибірки.

У цьому параграфі розглянемо методи побудови *інтервальних оцінок* параметрів генеральної сукупності.

4.1. ТОЧНІСТЬ І НАДІЙНІСТЬ ОЦІНКИ. ДОВІРЧИЙ ІНТЕРВАЛ

Довірчим інтервалом для оцінки невідомого параметра θ сукупності називається інтервал ($\hat{\theta}_1$; $\hat{\theta}_2$), який з надійністю γ накриває параметр θ. Це означає, що ймовірність подвійної нерівності

$$\hat{\theta}_1(x_1, x_2, ..., x_n) < \theta < \hat{\theta}_1(x_1, x_2, ..., x_n \;|$$

має бути не меншою від γ.

Кінці інтервалу $\hat{\theta}_1$ та $\hat{\theta}_2$ є випадковими величинами, знайденими за даними вибірки $x_1, x_2, ..., x_n$.

Задача побудови довірчих інтервалів має велике практичне значення, особливо для обробки результатів вимірювань.

У виборі $\hat{\theta}_1$ і $\hat{\theta}_2$ є деяка невизначеність. Інтервал ($\hat{\theta}_1$ і $\hat{\theta}_2$) може бути симетричним або, навпаки, однобічним.

Нехай за даними ***повторної*** вибірки знайдена точкова оцінка $\hat{\theta}$ невідомого параметра θ (θ вважаємо сталою величиною). $\hat{\theta}$ відрізняється від θ і є тим ліпшою оцінкою для θ, чим менша величина $|\hat{\theta} - \theta|$.

Характеристикою ***точності*** оцінки називається таке число $\delta > 0$, що $|\hat{\theta} - \theta| < \delta$. Чим менше δ, тим більша точність оцінки $\hat{\theta}$. Але про виконання нерівності $|\hat{\theta} - \theta| < \delta$ можна твердити не категорично, а лише з певним рівнем довіри (надійністю).

Надійністю оцінки $\hat{\theta}$ для параметра θ сукупності називається ймовірність γ, з якою виконується нерівність $\left|\hat{\theta}-\theta\right|<\delta$. Надійність має ще іншу назву — довірча ймовірність. Як правило, надійність γ має бути близькою до 1. Зокрема, надійність задають як 0,95, 0,99 або 0,999 і т. ін.

Отже, надійність γ дорівнює ймовірності випадкової події $\left|\hat{\theta}-\theta\right|<\delta$:

$$P\left(\left|\hat{\theta}-\theta\right|<\delta\right)=\gamma.$$

Оскільки нерівність $\left|\hat{\theta}-\theta\right|<\delta$ рівносильна нерівності $\hat{\theta}-\delta<\theta<\hat{\theta}+\delta$, то

$$P\left(\left|\hat{\theta}-\theta\right|<\delta\right)=P\left(\hat{\theta}-\delta<\theta<\hat{\theta}+\delta\right)=\gamma.$$

Це співвідношення слід тлумачити так: імовірність того, що інтервал $\left(\hat{\theta}-\delta,\hat{\theta}+\delta\right)$ накриє оцінюваний параметр θ, дорівнює γ. *Довірчим інтервалом* є симетричний інтервал $\left(\hat{\theta}-\delta,\hat{\theta}+\delta\right)$, який з надійністю γ накриває оцінюваний параметр θ. Оскільки $\hat{\theta}$ є випадковою величиною, то кінці цього інтервалу є випадковими величинами $\hat{\theta}_1=\hat{\theta}-\delta$ і $\hat{\theta}_2=\hat{\theta}+\delta$, і їх називають ***довірчими межами***.

Метод довірчих інтервалів розробили американський статистик Дж. Нейман та англійський статистик Р. Фішер.

Події $\hat{\theta}-\delta<\theta<\hat{\theta}+\delta$ та $\theta-\delta<\hat{\theta}<\theta+\delta$ рівносильні, тому

$$P\left(\theta-\delta<\hat{\theta}<\theta+\delta\right)=P\left(\hat{\theta}-\delta<\theta<\hat{\theta}+\delta\right)=\gamma.$$

У цій рівності лише $\hat{\theta}$ є випадковою величиною, і ймовірність події $\theta-\delta<\hat{\theta}<\theta+\delta$ залежить лише від розподілу величини $\hat{\theta}$. Якщо функцією розподілу величини $\hat{\theta}$ є $F_{\hat{\theta}}(x)$, то

$$P\left(\hat{\theta}-\delta<\theta<\hat{\theta}+\delta\right)=P\left(\theta-\delta<\hat{\theta}<\theta+\delta\right)=F_{\hat{\theta}}(\theta+\delta)-F_{\hat{\theta}}(\theta-\delta).$$

На відміну від точкових оцінок, для знаходження яких достатньо лише даних однієї вибірки, для побудови інтервальних оцінок треба знати закон розподілу досліджуваної величини X генеральної сукупності, закон розподілу оцінки $\hat{\theta}$, деякі параметри розподілу величини X. За досить великого обсягу вибірки на основі центральної граничної теореми часто припускають асимптотично-нормальний закон розподілу $\overline{X}_\text{в}$ за будь-якого закону розподілу величини X.

Для побудови інтервальних оцінок використовують спеціальні випадкові величини (статистики), закон розподілу яких відомий, але не залежить від оцінюваних параметрів генеральної сукупності.

Надійність γ — зазвичай задане заздалегідь, близьке до 1, число. Тоді $\alpha=1-\gamma$ — величина, близька до 0, називається рівнем значущості. Величина α характеризує ризик, що похибка вибірки вийде за допустимі межі.

4.2. ДОВІРЧИЙ ІНТЕРВАЛ ДЛЯ МАТЕМАТИЧНОГО СПОДІВАННЯ НОРМАЛЬНО РОЗПОДІЛЕНОЇ ВЕЛИЧИНИ ЗА ВІДОМОГО σ_Γ

Припустимо, що ознака X генеральної сукупності розподілена нормально і середнє квадратичне відхилення $\sigma(X)=\sigma$ відоме ($\sigma=\sigma_\Gamma$). Найліпшою незміщеною точковою оцінкою для математичного сподівання $M(X)=a$ є середня вибіркова $\overline{X}_\text{в}=\frac{1}{n}\sum_{i=1}^{n}X_i$, знайдена за вибіркою обсягу n (тут $\hat{\theta}=\overline{X}_\text{в}$, $\theta=a$). Тоді

$$P\left(\overline{X}_\text{в}-\delta<a<\overline{X}_\text{в}+\delta\right)=P\left(a-\delta<\overline{X}_\text{в}<a+\delta\right)=F_{\overline{X}_\text{в}}(a+\delta)-F_{\overline{X}_\text{в}}(a-\delta). \quad (4.1)$$

Визначимо функцію $F_{\overline{X}_в}(x)$ для величини $\overline{X}_в = \frac{1}{n}\sum_{i=1}^{n} X_i$. Величини $X_1, X_2, ..., X_n$ — незалежні нормально розподілені з $M(X_i) = a,\ D(X_i) = \sigma_г^2,\ i = \overline{1,n}$. Оскільки $\overline{X}_в = \frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n$, тобто є лінійною функцією від нормально розподілених величин X_i, то вона теж є нормально розподіленою і має два параметри:

$$M(\overline{X}_в) = M\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n\right) = \frac{1}{n}(M(X_1) + M(X_2) + ... + M(X_n)) = \frac{1}{n}\cdot n\cdot a = a;$$

$$D(\overline{X}_в) = D\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n\right) = \frac{1}{n^2}D(X_1) + \frac{1}{n^2}D(X_2) + ... + \frac{1}{n^2}D(X_n) = \frac{1}{n^2}n\cdot\sigma_г^2 = \frac{\sigma_г^2}{n};$$

$$\sigma(\overline{x}_в) = \frac{\sigma_г}{\sqrt{n}}. \tag{4.2}$$

Для безповторної вибірки

$$\sigma(\overline{x}_в) = \frac{\sigma_г}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx \frac{\sigma_г}{\sqrt{n}}\sqrt{1-\frac{n}{N}}. \tag{4.3}$$

Отже,

$$F_{\overline{X}_в}(x) = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma(\overline{X}_в)}\right).$$

Тоді за формулою (4.1) маємо

$$P(\overline{X}_в - \delta < a < \overline{X}_в + \delta) = P(a - \delta < \overline{X}_В < a + \delta) =$$
$$= \left(\frac{1}{2} + \Phi\left(\frac{a+\delta-a}{\sigma/\sqrt{n}}\right)\right) - \left(\frac{1}{2} + \Phi\left(\frac{a-\delta-a}{\sigma/\sqrt{n}}\right)\right) = \Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right).$$

Формула, яка зв'язує граничну точність δ, середню вибіркову $\overline{x}_в$, обсяг вибірки n, має вигляд $P(\overline{x}_в - \delta < a < \overline{x}_в + \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right)$, $\gamma = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t)$, де $\overline{x}_в$ — значення величини $\overline{X}_в$ для даної вибірки обсягу n, а $t = \frac{\delta\sqrt{n}}{\sigma}$.

Знайшовши з цієї рівності $\delta = \frac{t\sigma}{\sqrt{n}}$, запишемо довірчий інтервал, який з надійністю γ накриває невідоме математичне сподівання a нормально розподіленої ознаки X генеральної сукупності:

$$\overline{x}_в - \frac{t\sigma_Г}{\sqrt{n}} < a < \overline{x}_в + \frac{t\sigma_Г}{\sqrt{n}}. \tag{4.4}$$

Число t визначаємо з рівності $\gamma = 2\Phi(t)$ за таблицею функції Лапласа $\Phi(t)$; значення t відповідає аргументу, для якого $\Phi(t) = \frac{\gamma}{2}$ (дод. 2).

Аналізуючи формулу (4.4), можемо зробити такі висновки:

1) у разі збільшення обсягу вибірки (числа n) δ зменшується, тобто інтервал звужується й оцінка має більшу точність;

2) у разі збільшення надійності γ збільшується значення $\Phi(t)$, а отже, збільшується також t, бо $\Phi(t)$ — зростаюча функція; збільшення t приводить до збільшення δ, а отже, до розширення довірчого інтервалу й до зменшення точності інтервальної оцінки.

Формула $t = \frac{\delta\sqrt{n}}{\sigma}$ зв'язує три величини: t — коефіцієнт надійності, δ — точність оцінки та n — обсяг вибірки. Знаючи дві з цих величин, можна обчислити третю величину.

Імовірність того, що інтервал $\left(\bar{x}_{\text{в}} - \frac{t\sigma}{\sqrt{n}}; \quad \bar{x}_{\text{в}} + \frac{t\sigma}{\sqrt{n}}\right)$ не накриє оцінюваний параметр, дорівнює $\alpha = 1 - \gamma$; α є рівнем значущості (α зазвичай невелике число — 0,01 або 0,001);

$\bar{x}_{\text{в}} - \frac{t\sigma}{\sqrt{n}}$ — нижня межа довірчого інтервалу;

$\bar{x}_{\text{в}} + \frac{t\sigma}{\sqrt{n}}$ — верхня межа довірчого інтервалу.

Отже, для побудови довірчого інтервалу для математичного сподівання *a* нормально розподіленої ознаки нам потрібно знати чотири величини: $\bar{x}_{\text{в}}$, t, σ, n.

Приклад 4.1. Нехай $X \sim N(a;3)$. Знайти довірчий інтервал для математичного сподівання *a* при $\bar{x}_{\text{в}} = 11{,}92$, знайденій за вибіркою обсягу $n = 36$, з надійністю $\gamma = 0{,}95$.

Р о з в ' я з а н н я. Оскільки ознака X нормально розподілена та має відоме $\sigma_Г = 3$, то довірчий інтервал для її невідомого математичного сподівання *a* знайдемо за формулою (4.4), де $\bar{x}_{\text{в}} = 11{,}92$, $n = 36$.

Коефіцієнт довіри t знайдемо за таблицею значень функції Лапласа: $\Phi(t) = \frac{0{,}95}{2} = 0{,}475 \Rightarrow t = 1{,}96$.

Тоді $\delta = \frac{1{,}96 \cdot 3}{\sqrt{36}} = 0{,}98$.

Використавши формулу (4.4) , дістанемо довірчий інтервал

$$11{,}92 - 0{,}98 < a < 11{,}92 + 0{,}98 \text{ або } 10{,}94 < a < 12{,}90.$$

Отже, *a* перебуває в інтервалі (10,94; 12,90) з надійністю 0,95.

Насправді було здійснено лише одну вибірку, і тому було обчислено лише одне значення $\bar{x}_{\text{в}}$. Відповідна інтервальна оцінка буде або коректно включати параметр, або неправильно його виключати. На жаль, фахівці-практики зі статистики не знають, чи праві вони кожного разу; відомо лише, що періодично вони будуть некоректно оцінювати параметр, тобто помилятись.

Крім того, широкий довірчий інтервал дає мало інформації. Збільшення обсягу вибірки в 4 рази звужує довірчий інтервал удвічі. Це логічно, бо більший обсяг вибірки забезпечує більше інформації. Але збільшення обсягу вибірки збільшує й вартість її опрацювання.

Приклад 4.2. Компанія, яка виготовляє комп'ютери, доставляє їх безпосередньо до клієнтів, які замовляють машини через Інтернет. Компанія конкурує перш за все за ціною і швидкістю постачання. Щоб досягти мети щодо швидкості доставки, компанія кожен зі своїх найпопулярніших комп'ютерів постачає на склади в усій країні, де вони зберігаються і можуть бути доставлені клієнту за 1 день. Ця стратегія вимагає високого рівня запасів, що значно збільшує вартість зберігання. Для зменшення витрат менеджер хоче з'ясувати оптимальний рівень запасів. При цьому він відзначив, що щоденний попит є випадкова змінна, розподілена за нормальним законом із середнім квадратичним відхиленням $\sigma = 75$ комп'ютерів. Менеджер провів спостереження протягом 25-ти періодів і зафіксував дані попиту:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

За цими даними менеджер планує оцінити середній попит 95 %-вим довірчим інтервалом.

Р о з в ' я з а н н я. Для визначення оптимального рівня запасів менеджер повинен знати середній попит, який для нормально розподіленої величини позначаємо літерою *a*. Оцінимо *a* довірчим інтервалом, межами якого будуть числа $\bar{x}_{\text{в}} - \frac{t\sigma}{\sqrt{n}}$ та $\bar{x}_{\text{в}} + \frac{t\sigma}{\sqrt{n}}$. Тепер обчислимо ці межі при $n = 25$, $\sigma = 75$ і $t = 1{,}96$ $(2\Phi(t) = 0{,}95)$.

Перш за все знайдемо $\bar{x}_{\text{в}}$ за наведеною вибіркою:

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{25}(235 + 374 + ... + 332 + 334) = \frac{1}{25} \cdot 9254 = 370{,}16.$$

Підставивши $\bar{x}_в$, n, σ і t у формули довірчих меж, знайдемо

$$\bar{x}_в \mp \frac{t\sigma}{\sqrt{n}} = 370{,}16 \mp \frac{1{,}96 \cdot 75}{\sqrt{25}} = 370{,}16 \mp 29{,}40 \text{. Отже, } 340{,}76 < a < 399{,}56 .$$

Цю оцінку менеджер може використовувати як початкову інформацію для створення політики запасів. Підкреслимо, що дехто помилково трактує одержаний довірчий інтервал як такий, що з імовірністю 0,95 середнє потрапляє в інтервал (340,76; 399,56). Ця інтерпретація неправильна, бо середнє сукупності не є змінною, про яку складене ймовірнісне твердження. Середнє *а* — фіксоване, але невідоме. Одержаний результат означає, що ми *неодноразово* робимо вибірки обсягу 25 одиниць з досліджуваної сукупності, і 95 % знайдених $\bar{x}_в$ будуть такими, що інтервал ($\bar{x}_в - 29{,}4$; $\bar{x}_в + 29{,}4$) накриє невідоме *а* і 5 % знайдених $\bar{x}_в$ будуть такими, що цей інтервал не вміщуватиме в себе *а*. Уявіть, що ми робимо 40 вибірок по 25 спостережень. Тоді 5 % від 40 — це 2 вибірки, знайдені за якими довірчі інтервали не міститимуть *а*.

Важливо зрозуміти, що навіть із досконало виконаних експериментів певна частка (у нашому прикладі 5 %) надаватиме некоректні оцінки.

На ширину довірчого інтервалу має вплив і середнє квадратичне відхилення σ. У разі збільшення σ інтервал стає ширшим, а оцінка менш точною. Це абсолютно логічний результат, бо при збільшенні σ випадкова величина має більшу варіабельність і тому важче оцінити середнє по сукупності.

Інтервальна оцінка (340,76; 399,56), як основа для створення політики запасів, дуже широка, а потрібна більша точність. На щастя, статистики можуть управляти шириною інтервалу, збільшуючи обсяг вибірки для одержання більш вузьких інтервалів. Щоб зрозуміти, як ми можемо визначити необхідний розмір вибірки, розглянемо помилку вибірки. Помилка вибірки — це різниця між оцінкою $\hat{\theta}$ і параметром θ. У нашому випадку помилка $|\mathrm{X}_в - a| < \frac{t\sigma}{\sqrt{n}}$, де $\frac{t\sigma}{\sqrt{n}} = \delta$ — це максимальна помилка, яку ми готові прийняти.

Формула

$$\delta = \frac{t\sigma}{\sqrt{n}} \tag{4.5}$$

за відомого $\sigma = \sigma_Г$ поєднує три величини:

1) δ — точність оцінки;
2) t — коефіцієнт надійності;
3) n — обсяг вибірки.

Знаючи дві з цих величин, можемо знайти третю. Зокрема коефіцієнт довіри

$$t = \frac{\delta\sqrt{n}}{\sigma}. \tag{4.6}$$

За значенням t і таблицею функції $\Phi(t)$ знайдемо надійність: $\gamma = 2\Phi(t)$.

Необхідний обсяг вибірки знаходимо за формулою

$$n = \left(\frac{t\sigma}{\delta}\right)^2. \tag{4.7}$$

Повернемося до прикладу 4.2 і визначимо необхідний обсяг вибірки. Коефіцієнт надійності $t = 1{,}96$.

Припустимо, що менеджер вирішив оцінити середній попит з гранично допустимою похибкою $\delta = 16$ одиниць. Тоді з рівняння $\frac{1{,}96 \cdot 75}{\sqrt{n}} = 16$ знайдемо необхідний для такої точності обсяг вибірки: $n = \left(\frac{1{,}96 \cdot 75}{16}\right)^2 = 84{,}41$.

Оскільки n має бути цілим числом, то округлимо цей результат до 85. Це означає, що для одержання оцінки середнього a з граничною похибкою не більшою від 16 за 95 %-вого рівня довіри, потрібно провести не 25, а 85 спостережень.

При $n = 85$ довірчий інтервал для *а* буде $\bar{x}_в - 16 < a < \bar{x}_в + 16$, де $\bar{x}_в$ треба обчислити за вибіркою з 85 спостережень попиту.

4.3. ДОВІРЧИЙ ІНТЕРВАЛ ДЛЯ МАТЕМАТИЧНОГО СПОДІВАННЯ НОРМАЛЬНО РОЗПОДІЛЕНОЇ ВЕЛИЧИНИ ЗА НЕВІДОМОГО σ_{Γ}

Нехай ознака X нормально розподілена в генеральній сукупності. Тепер розглянемо більш реалістичну ситуацію, а саме коли невідомі середнє значення $M(X)=a$ і середнє квадратичне відхилення $\sigma(X)=\sigma_{\Gamma}$. Тому не можна використовувати формули з попереднього параграфа для оцінки a.

Побудуємо випадкову величину $\mathrm{T}=\dfrac{\overline{X}-a}{s/\sqrt{n}}$, яка має розподіл Стьюдента з $\nu=n-1$ степенями вільності, де $\overline{X}$ — середня вибіркова; s — виправлене середнє квадратичне відхилення вибіркове, n — обсяг вибірки. Щільність розподілу величини T:

$$f(t,n)=B_n\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}},$$

де $B_n=\dfrac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\cdot\Gamma\left(\frac{n-1}{2}\right)}$, $\Gamma(x)=\int\limits_0^{+\infty}u^{x-1}e^{-u}du$ — гамма-функція.

Функція $f(t,n)$ не залежить ні від a, ні від $\sigma_{\text{г}}$. Оскільки $f(t,n)$ — парна функція від t, то

$$P\left(\left|\frac{\overline{X}-a}{s/\sqrt{n}}\right|<t_{\gamma}\right)=2\int\limits_0^{t_{\gamma}}f(t,n)dt=\gamma$$

або

$$P\left(-t_{\gamma}<\frac{\overline{X}-a}{s/\sqrt{n}}<t_{\gamma}\right)=P\left(\overline{X}-\frac{t_{\gamma}s}{\sqrt{n}}<a<\overline{X}+\frac{t_{\gamma}s}{\sqrt{n}}\right)=\gamma.$$

Звідси маємо довірчий інтервал для $M(X)=\mu$:

$$\overline{x}_B-\frac{t_{\gamma}s}{\sqrt{n}}<a<\overline{x}_B+\frac{t_{\gamma}s}{\sqrt{n}}, \tag{4.8}$$

який з надійністю γ накриває невідоме математичне сподівання a; число t_{γ} знаходимо за таблицею розподілу Стьюдента (дод. 6).

Приклад 4.3. За даними вибірки з 5 одиниць маємо $\overline{x}_{\text{в}}=175,\ s=30$. Оцінити математичне сподівання $M(X)$ нормально розподіленої ознаки X генеральної сукупності з 90 %-вою надійністю.

Р о з в ' я з а н н я. Знайдемо інтервальну оцінку за формулою (4.8). Оскільки відомі $\overline{x}_{\text{в}}=175$, $s=30$ та $n=5$, то за таблицею розподілу Стьюдента при $\gamma=0{,}9$ та числі степенів вільності $v=n-1=5-1=4$ знайдемо лише $t_{\gamma}=2{,}132$.

Тому довірчий інтервал для $M(X)=a$ має межі $175\mp\dfrac{2{,}132\cdot 30}{\sqrt{5}}=175\mp 28{,}6.$

Отже, з надійністю 0,90 (90 %) невідоме математичне сподівання a буде накрите довірчим інтервалом $146{,}4<a<203{,}6$.

Інтервал дуже широкий, і для його звуження (підвищення точності оцінки) необхідно взяти вибірку більшого обсягу, бо вибірка малого обсягу вміщує мало інформації про досліджувану ознаку.

Зауваження. $\lim\limits_{n\to\infty}f(t,n)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. Тому за необмеженого зростання n розподіл Стьюдента наближається до нормального. Практично при $n>30$ розподіл Стьюдента можна замінити нормальним законом і довірчий коефіцієнт t_{γ} знаходити так, як і для нормального закону за таблицею значень функції Лапласа, тобто з рівняння $2\Phi(t_{\gamma})=\gamma$.

Проте слід зазначити, що для малих вибірок ($n \le 30$) заміна розподілу Стьюдента нормальним розподілом призводить до грубих помилок — до невиправданого звуження довірчого інтервалу, тобто до невиправданого підвищення точності оцінки.

Так, у нашому прикладі заміна розподілу Стьюдента на нормальний дає t_γ з рівняння $2\Phi(t_\gamma) = 0{,}9$, тобто $t_\gamma = 1{,}645$ і відповідний довірчий інтервал буде вужчий, а саме (152,93;197,07) — це неправильно!

Те, що розподіл Стьюдента для малої вибірки дає широкий довірчий інтервал, свідчить не про слабкість цього методу, а лише про те, що малого обсягу вибірка містить мало інформації про ознаку X.

Приклад 4.4. Представники товариства захисту споживачів для контролю розфасованої в пакети продукції здійснили вибіркове обстеження 18 пакетів, на яких зазначалося, що вміст пакета важить 800 грамів, зважили і зафіксували їхню вагу:

775	780	794	799	793	791
789	797	782	806	795	779
791	792	805	798	792	787

Знайти 99 %-вий довірчий інтервал для ваги пакетів.

Р о з в ' я з а н н я. За умовою $n = 18$, $a = 800$. Нехай вага розфасованої продукції нормально розподілена величина X. Знайдемо 99 %-вий довірчий інтервал для a за даними вибірки. Обчислимо

$$\bar{x}_в = \frac{1}{18}(x_1 + x_2 + ... + x_{18}) = \frac{14245}{18} = 791{,}39;$$

$s^2 = \frac{18}{17}\left(\frac{775^2 + 780^2 + ... + 787^2}{18} - (791{,}39)^2\right) = 71{,}74$. Отже, $s = \sqrt{71{,}74} \approx 8{,}47$. Коефіцієнт довіри $t_{\gamma=0{,}99,\ \nu=17} = 2{,}898$ (за розподілом Стьюдента). Довірчі межі: $791{,}39 \mp \frac{8{,}47 \cdot 2{,}898}{\sqrt{18}} = 791{,}39 \mp 5{,}79$. Отже, 99 %-вий довірчий інтервал для $a: 785{,}6 < a < 797{,}18$.

4.4. ПОБУДОВА ДОВІРЧИХ ІНТЕРВАЛІВ ЗА БЕЗПОВТОРНОЇ ВИБІРКИ

Формули, одержані для повторної вибірки, справедливі і для безповторної вибірки за умови, що в тих випадках, де використовувалась формула (4.2)

$$\sigma(\bar{X}_в) = \frac{\sigma_Г}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{\sigma_в}{\sqrt{n-1}},$$

потрібно використати формулу

$$\sigma(\bar{X}_в) \approx \frac{\sigma_в}{\sqrt{n-1}}\sqrt{1 - \frac{n}{N}}, \tag{4.3}$$

де $\sigma_Г$ — середнє квадратичне відхилення генеральне; $\sigma(\bar{X}_в)$ — середнє квадратичне відхилення середньої вибіркової безповторної вибірки; n та N — обсяги вибірки та генеральної сукупності. Ця формула показує, що похибка безповторної вибірки менша, ніж похибка повторної вибірки, за однакового обсягу n обох типів вибірок. Змінюється і формула для необхідного обсягу вибірки.

Довірчий інтервал для математичного сподівання a за відомого $\sigma_г$ для безповторної вибірки матиме вигляд:

$$\bar{x}_в - \frac{t\sigma_г}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} < a < \bar{x}_в + \frac{t\sigma_г}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}. \tag{4.9}$$

Знайдемо формулу, за якою визначається необхідний обсяг безповторної вибірки. З формули (4.5) маємо

$$\delta = \frac{t\sigma_г}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \Rightarrow \delta^2 = \frac{(t\sigma_г)^2}{n}\left(1 - \frac{n}{N}\right) \Rightarrow n\left(\delta^2 + \frac{t^2\sigma_г^2}{N}\right) = t^2\sigma_г^2 \Rightarrow n\left(N\delta^2 + t^2\sigma_г^2\right) = Nt^2\sigma_г^2 \Rightarrow$$

$$n = \frac{Nt^2\sigma_г^2}{N\delta^2 + t^2\sigma_г^2}. \qquad (4.10.)$$

Зауваження 1. Якщо за умовою задачі генеральне середнє квадратичне відхилення σ невідоме, то у формулі для довірчого інтервалу для a необхідно замість $\sigma_г$ взяти $s = \sqrt{\frac{n}{n-1}}\sigma_в$. При цьому $\delta = \frac{ts}{\sqrt{n}} = \frac{t\sigma_в}{\sqrt{n-1}}\sqrt{1-\frac{n}{N}}$.

Для безповторної вибірки межі довірчого інтервалу будуть такі:

$$\bar{x}_в \pm \frac{t\sigma_в}{\sqrt{n-1}}\sqrt{1-\frac{n}{N}}.$$

Зауваження 2. Якщо обсяг n безповторної вибірки малий порівняно з обсягом N усієї сукупності ($1-\frac{n}{N} \approx 1$) або якщо обсяг генеральної сукупності взагалі невідомий, то в цих випадках формула довірчого інтервалу для a буде такою самою, як і для повторної вибірки.

Приклад 4.5. Із 1500 деталей було відібрано 250, розмір яких згрупований у таблиці:

Розмір деталей	7,8—8,0	8,0—8,2	8,2—8,4	8,4—8,6	8,6—8,8	8,8—9,0
Кількість деталей	5	20	80	95	40	10

Знайти середню похибку вибірки під час визначення середнього розміру в усій сукупності за такої умови:

а) вибірка повторна;

б) вибірка безповторна.

Знайти довірчий інтервал для середнього розміру деталей з надійністю 95 %. Відомо, що розмір деталей є нормально розподіленою величиною.

Р о з в ' я з а н н я. За умовою $n = 250$, $N = 1500$. Похибка вибірки:

а) для повторної вибірки знаходиться за формулою (4.2);

б) для безповторної вибірки знаходиться за формулою (4.3).

Дисперсія $\sigma_г^2$ невідома. Тому знайдемо її точкову оцінку $\sigma_в^2$ з вибірки. Для цього перейдемо від інтервального розподілу до точкового розподілу та, для зручності обчислень вручну, до умовних варіант:

i	Інтервал $\tilde{x}_{i-1} - \tilde{x}_i$	Частота m_i	Середина інтервалу, x_i	$u_i = \frac{x_i - c}{h}$	$u_i m_i$	$u_i^2 m_i$
1	7,8—8,0	5	7,9	–3	–15	45
2	8,0—8,2	20	8,1	–2	–40	80
3	8,2—8,4	80	8,3	–1	–80	80
4	8,4—8,6	95	8,5 = c	0	0	0
5	8,6—8,8	40	8,7	1	40	40
6	8,8—9,0	10	8,9	2	20	40
сума		250			–75	285

Ширина інтервалу $h = 0{,}2$. Тоді $\bar{u} = \frac{-75}{250} = -0{,}3$, $\overline{u^2} = \frac{285}{250} = 1{,}14$, $\bar{x}_в = \bar{u}\cdot h + c = (-0{,}3)\cdot 0{,}2 + 8{,}5 = 8{,}5 - 0{,}06 = 8{,}44$.

$$\sigma_в^2 = \left(\overline{u^2} - (\bar{u})^2\right)h^2 = \left(1{,}14 - (-0{,}3)^2\right)\cdot(0{,}2)^2 = 1{,}05\cdot 0{,}04 = 0{,}042.$$

а) для повторної вибірки похибку оцінимо за формулою (4.2):

$$\sigma(\bar{X}_в) = \frac{\sigma_в}{\sqrt{n-1}} = \frac{\sqrt{0{,}042}}{\sqrt{249}} = \sqrt{\frac{0{,}042}{249}} \approx 0{,}013.$$

б) для безповторної вибірки похибку вибірки оцінюємо за формулою (4.3):

$$\sigma\left(\overline{X}_{\text{в}}\right)=\frac{\sigma_{в}}{\sqrt{n-1}}\sqrt{1-\frac{n}{N}}=0{,}013\cdot\sqrt{1-\frac{250}{1500}}\approx 0{,}012\,.$$

Отже, як і належить, похибка безповторної вибірки менша, ніж похибка повторної вибірки.

З рівня довіри $\gamma=0{,}95$ маємо коефіцієнт довіри $t=1{,}96$. Розрахунки меж довірчих інтервалів подано в таблиці

Вибірка	Гранична похибка $\delta=t\sigma(\overline{x}_{\text{в}})$	межі довірчого інтервалу	
		нижня $\overline{x}_{\text{в}}-\delta$	верхня $\overline{x}_{\text{в}}+\delta$
Повторна	$\delta=1{,}96\cdot 0{,}013=0{,}02548$	$8{,}44-0{,}02548=8{,}41452$	$8{,}44+0{,}02548=8{,}46548$
Безповторна	$\delta=1{,}96\cdot 0{,}012=0{,}02352$	$8{,}44-0{,}02352=8{,}41648$	$8{,}44+0{,}02352=8{,}46352$

Отже, для повторної вибірки $8{,}41452<a<8{,}46548$, для безповторної вибірки — $8{,}41648<$ $<a<8{,}46352$.

Можемо зробити висновок, що ширина довірчого інтервалу для повторної вибірки дорівнює 0,05096, а для безповторної — 0,04704, тобто за однакової надійності безповторна вибірка дає більш точну оцінку для середнього значення $M(X)=a$.

4.5. ІНТЕРВАЛЬНА ОЦІНКА ДЛЯ СЕРЕДНЬОГО КВАДРАТИЧНОГО ВІДХИЛЕННЯ σ НОРМАЛЬНО РОЗПОДІЛЕНОЇ ВЕЛИЧИНИ

Нехай ознака X генеральної сукупності — нормально розподілена величина. Необхідно оцінити невідоме середнє квадратичне відхилення $\sigma_{\text{г}}(X)=\sigma$. Оцінимо σ довірчим інтервалом із заданою надійністю γ. З цією метою скористаємось випадковою величиною χ.

Точковою незміщеною оцінкою для σ є виправлене середнє квадратичне відхилення $s=\sqrt{\frac{n}{n-1}}\sigma_{\text{в}}$. Для побудови довірчого інтервалу розглянемо різницю $\sigma-s$ і висунемо вимогу, щоб

$$P\left(\left|\sigma-s\right|<\delta\right)=\gamma \text{ або } P\left(s-\delta<\sigma<s+\delta\right)=\gamma\,.$$

Нерівність $s-\delta<\sigma<s+\delta$ рівносильна нерівності $s\left(1-\frac{\delta}{s}\right)<\sigma<s\left(1+\frac{\delta}{s}\right)$.

Позначивши $\frac{\delta}{s}=q$, дістанемо нерівність:

$$s(1-q)<\sigma<s(1+q), \tag{4.11}$$

з якої знайдемо довірчий інтервал, якщо обчислимо q. З цією метою введемо випадкову величину $Y=\chi^2=\frac{(n-1)s^2}{\sigma^2}$, яка розподілена за законом χ^2 з $n-1$ ступенями вільності (див. п. 6). Квадратний корінь із цієї величини позначимо $\chi=\frac{s}{\sigma}\sqrt{n-1}$, де n — обсяг вибірки.

Щільність розподілу величини Y, розподіленої за законом χ^2 з числом ступенів вільності $v=n-1$, має вигляд:

$$f(y)=\begin{cases}0, & y\le 0,\\ \dfrac{e^{-\frac{y}{2}}y^{\frac{v}{2}-1}}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}, & y>0,\end{cases}\qquad \Gamma(x)=\int_0^{\infty}u^{x-1}e^{-u}du$$

або $f(y)=\begin{cases} 0, & y\le 0 \\ \dfrac{e^{-\frac{y}{2}}y^{\frac{n-3}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}, & y>0 \end{cases}$.

Величина $\chi=\sqrt{Y}$, тому щільність її розподілу знайдемо за формулою:

$$g(\chi)=f(\chi^2)\cdot 2\chi=\begin{cases} 0, & \chi\le 0 \\ \dfrac{e^{-\frac{\chi^2}{2}}(\chi^2)^{\frac{n-3}{2}}\cdot 2\chi}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}, & \chi>0 \end{cases}=\begin{cases} 0, & \chi\le 0 \\ \dfrac{e^{\frac{-\chi^2}{2}}\chi^{n-2}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}, & \chi>0 \end{cases}.$$

Як бачимо, цей розподіл не залежить від оцінюваного параметра σ, а залежить лише від обсягу вибірки n.

Перетворимо нерівність (4.11) на рівносильну $\frac{\sqrt{n-1}}{1+q}<\frac{s\sqrt{n-1}}{\sigma}<\frac{\sqrt{n-1}}{1-q}$, поетапно здійснивши перехід до обернених величин $\frac{1}{s(1+q)}<\frac{1}{\sigma}<\frac{1}{s(1-q)}$ (нехай $q<1$) та помноживши всі члени цієї нерівності на $s\sqrt{n-1}$.

Оскільки нерівності $s-\delta<\sigma<s+\delta$ та $\frac{\sqrt{n-1}}{1+q}<\chi<\frac{\sqrt{n-1}}{1-q}$ рівносильні, то

$$P(s-\delta<\sigma<s+\sigma)=P\left(\frac{\sqrt{n-1}}{1+q}<\chi<\frac{\sqrt{n-1}}{1-q}\right)=\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}}g(\chi)d\chi=\gamma.$$

Звідси можна знайти q за заданих n та γ і записати довірчий інтервал для σ за нерівністю (4.11). Оскільки цей інтеграл знайти непросто, то складені таблиці для знаходження $q(\gamma,n)$ (табл. 10).

Приклад 4.6. Кількісна ознака X генеральної сукупності розподілена нормально. За вибіркою обсягом $n=35$ знайдено виправлене середнє квадратичне відхилення $s=0{,}8$. Знайдіть інтервал, який з надійністю $\gamma=0{,}99$ накриє невідоме середнє квадратичне відхилення σ генеральної сукупності.

Р о з в ' я з а н н я. Знайдемо q при $n=35$ та $\gamma=0{,}99$. Для цього звернемось до таблиці, з якої маємо $q(0{,}99;35)=0{,}38$. Підставивши $s=0{,}8$ та $q=0{,}38$ у формулу (4.11), дістанемо шуканий довірчий інтервал:

$$0{,}8(1-0{,}38)<\sigma<0{,}8(1+0{,}38) \text{ або } 0{,}496<\sigma<1{,}104.$$

Формула (4.11) застосовувалася за умови $q<1$. Якщо $q>1$ то, враховуючи, що $\sigma>0$, довірчий інтервал будемо знаходити за нерівністю:

$$0<\sigma<s(1+q). \qquad (4.12)$$

Приклад 4.7. Кількісна ознака X генеральної сукупності розподілена нормально. За вибіркою обсягу $n=15$ знайдено виправлене середнє квадратичне відхилення $s=0{,}18$. Знайдіть довірчий інтервал, який з надійністю 0,999 накриває невідоме середнє квадратичне відхилення генеральної сукупності σ.

Р о з в ' я з а н н я. За таблицею знайдемо $q(0{,}999;15)=1{,}15$. Підставляючи $s=0{,}18$, $q=1{,}15>1$ в нерівність (4.12), знайдемо довірчий інтервал для σ:

$$0<\sigma<0{,}18(1+1{,}15) \text{ або } 0<\sigma<0{,}387.$$

4.6. ДОВІРЧИЙ ІНТЕРВАЛ ДЛЯ ДИСПЕРСІЇ ЗА НЕВІДОМОГО МАТЕМАТИЧНОГО СПОДІВАННЯ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

Припустимо, що ознака X генеральної сукупності нормально розподілена. Побудуємо довірчий інтервал для σ^2, скориставшись випадковою величиною χ^2.

Найкращою точковою оцінкою для дисперсії σ^2 є виправлена вибіркова дисперсія $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_в)^2$. Відомо, що величина $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ має χ^2— розподіл з $\nu = n-1$ степенями вільності.

Обчислимо

$$P\left(u_1 < \chi^2 < u_2\right) = F_{\chi^2}(u_2, \nu) - F_{\chi^2}(u_1, \nu), \tag{4.13}$$

де $F_{\chi^2}(u,\nu)$ — функція розподілу χ^2. Перейдемо в нерівності $u_1 < \chi^2 < u_2$ до обернених величин

$$\frac{1}{u_2} < \frac{1}{\chi^2} < \frac{1}{u_1}$$

або

$$\frac{1}{u_2} < \frac{\sigma^2}{s^2(n-1)} < \frac{1}{u_1}.$$

Домноживши всі частини цієї нерівності на $s^2(n-1) > 0$, дістанемо довірчий інтервал для дисперсії σ^2

$$\frac{s^2(n-1)}{u_2} < \sigma^2 < \frac{s^2(n-1)}{u_1}, \tag{4.14}$$

де u_1 та u_2 визначимо з надійністю γ. А саме, нерівність (4.14) рівносильна нерівності $u_1 < \chi^2 < u_2$. Тому за формулою (4.13) маємо:

$$P\left(\frac{s^2(n-1)}{u_2} < \sigma^2 < \frac{s^2(n-1)}{u_1}\right) = F_{\chi^2}(u_2, \nu) - F_{\chi^2}(u_1, \nu) = P\left(\chi^2 < u_2\right) - P\left(\chi^2 < u_1\right) =$$
$$\left(1 - P\left(\chi^2 > u_2\right)\right) - \left(1 - P\left(\chi^2 > u_1\right)\right) = P\left(\chi^2 > u_1\right) - P\left(\chi^2 > u_2\right) = \gamma.$$

Візьмемо $P(\chi^2 > u_2) = \frac{\alpha}{2}$, $P(\chi^2 > u_1) = 1 - \frac{\alpha}{2}$. Тоді $\gamma = 1 - \alpha$ і $\frac{\alpha}{2} = \frac{1}{2}(1-\gamma)$.

Знайдемо u_1 та u_2 за табл. 7 для $P(\chi^2 > u_1)$ і $P(\chi^2 > u_2)$. Ця таблиця містить значення u_1 і u_2 лише для максимального числа ступенів вільності $\nu = n - 1 = 100$, тобто для вибірки обсягу $n \le 101$.

Припустимо, що за результатами 22 вимірювань деякої величини, розподіленої нормально, знайдемо точкову оцінку для дисперсії $\sigma^2 \approx s^2 = 0{,}0303$. Знайдемо довірчий інтервал для σ^2 з надійністю $\gamma = 0{,}98$. За табл. 7 розподілу χ^2 при $\gamma = 0{,}98$ і $\nu = 21$ (степенями вільності) маємо

$$\alpha = 1 - 0{,}98 = 0{,}22; \quad \frac{\alpha}{2} = 0{,}01; \quad 1 - \frac{\alpha}{2} = 0{,}99;$$
$$P\left(\chi^2 > \chi_1^2\right) = 0{,}99 \Rightarrow \chi_1^2 = 8{,}897;$$
$$P\left(\chi^2 > \chi_2^2\right) = 0{,}01 \Rightarrow \chi_2^2 = 38{,}932.$$

Тоді довірчий інтервал для σ^2 має такі межі:

$\frac{(n-1)s^2}{\chi_2^2} = \frac{21 \cdot 0{,}0303}{38{,}932} \approx 0{,}016$ — ліва межа довірчого інтервалу,

$\frac{(n-1)s^2}{\chi_1^2} = \frac{21 \cdot 0{,}0303}{8{,}897} \approx 0{,}075$ — права межа довірчого інтервалу.

Але, якщо обсяг вибірки $n > 101$, то можемо скористатись асимптотичною властивістю χ^2 — розподілу. А саме, уже при $n > 30$ величина $Z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ є нормованою нормально розподіленою випадковою величиною з $M(Z) = 0$ і $\sigma(Z) = 1$ За заданою величиною α знайдемо:

$$P(|Z| < z) = P(-z < Z < z) = 2\Phi(z) = 1 - \alpha .$$

Звідки маємо:

$$P\left(-z < \sqrt{2\chi^2} - \sqrt{2\nu - 1} < z\right) = P\left(\sqrt{2\nu - 1} - z < \sqrt{2\chi^2} < \sqrt{2\nu - 1} + z\right) =$$

$$= P\left(\left(\sqrt{2\nu - 1} - z\right)^2 < 2\chi^2 < \left(\sqrt{2\nu - 1} + z\right)^2\right) = P\left(\frac{\left(\sqrt{2\nu - 1} - z\right)^2}{2} < \chi^2 < \frac{\left(\sqrt{2\nu - 1} + z\right)^2}{2}\right) =$$

$$= 2\Phi(z) = 1 - \alpha .$$

За таблицею функції Лапласа з рівняння $2\Phi(z) = 1 - \alpha = \gamma$ знаходимо значення *z,* а потім обчислюємо $u_1 = \frac{1}{2}\left(\sqrt{2\nu - 1} - z\right)^2$ та $u_2 = \frac{1}{2}\left(\sqrt{2\nu - 1} + z\right)^2$. Підставляючи знайдені u_1 та u_2 в нерівність (4.14.), дістанемо довірчий інтервал для σ^2.

Із формули (4.14) маємо ще одну формулу для знаходження довірчого інтервалу для середнього квадратичного відхилення σ:

$$\frac{\sqrt{n-1} \cdot s}{\sqrt{u_2}} < \sigma < \frac{\sqrt{n-1} \cdot s}{\sqrt{u_1}} .$$

Приклад 4.7. Із генеральної сукупності здобута вибірка обсягом n = 150, з якої знайдено $\bar{x}_в = 10$ і $s^2 = 0{,}09$. За рівня значущості $\alpha = 0{,}05$ знайти: а) довірчий інтервал для генеральної дисперсії σ^2; б) з якою надійністю можна стверджувати, що похибка δ під час оцінювання *a* не перевищує 0,05? в) за якого обсягу вибірки можна стверджувати з надійністю $\gamma = 0{,}95$, що гранична похибка $\delta = 0{,}05$?

Р о з в ’ я з а н н я. а) Для побудови довірчого інтервалу для σ^2 маємо $n = 150$, $\gamma = 1 - 0{,}05 = 0{,}95$, $s^2 = 0{,}09$. Знайдемо коефіцієнт довіри t з рівняння $2\Phi(t) = 0{,}95$. Отже, $t = 1{,}96$. Тепер обчислимо

$$u_1 = \frac{1}{2}\left(\sqrt{2 \cdot 149 - 1} - 1{,}96\right)^2 = 116{,}64;$$

$$u_2 = \frac{1}{2}\left(\sqrt{2 \cdot 149 - 1} + 1{,}96\right)^2 = 184{,}21.$$

За формулою (4.14) дістанемо довірчий інтервал для дисперсії:

$$\frac{0{,}09 \cdot 149}{184{,}21} < \sigma^2 < \frac{0{,}09 \cdot 149}{116{,}64} .$$

Остаточно маємо: $0{,}073 < \sigma^2 < 0{,}115$.

б) Обчислимо ймовірність того, що гранична похибка вибірки $\delta = 0{,}05$. Із формули (4.8) маємо:

$$\delta = \frac{ts}{\sqrt{n}} . \qquad (4.14^{I})$$

Підставивши $\delta = 0{,}05$, $s = \sqrt{0{,}09} = 0{,}3$, $n = 150$ дістанемо коефіцієнт довіри $t = \frac{0{,}05 \cdot \sqrt{150}}{0{,}3} \approx 2{,}04$.
За коефіцієнтом довіри $t = 2{,}04$ знайдемо довірчу ймовірність: $2\Phi(2{,}04) \approx 0{,}96$.

Отже, ймовірність того, що *a* знаходиться в інтервалі (9,95; 10,05), дорівнює 0,96.

в) Для визначення необхідного обсягу вибірки *n*, за якого забезпечується з надійністю $\gamma = 0{,}95$ гранична похибка $\delta = 0{,}05$, застосуємо формулу (4.14^{I}). З цієї формули маємо:

$$n = \frac{t^2 s^2}{\sigma^2} = \frac{(1{,}96)^2 \cdot 0{,}09}{(0{,}05)^2} \approx 139 .$$

4.7. ОЦІНКА ЧАСТКИ *p* ЗА ПОВТОРНОЮ ВИБІРКОЮ

У § 3 показано, що точковою оцінкою частки p ознаки в генеральній сукупності є вибіркова частка $w = \frac{m}{n}$. Для оцінки граничної похибки $|p - w| < \delta$ необхідно знайти інтервальну оцінку для p:

$$w - \delta < p < w + \delta. \tag{4.15}$$

Частка w ознаки у повторній вибірці точно описується біноміальним законом розподілу:

$$P_n(m) = C_n^m p^m (1-p)^{n-m}.$$

Оскільки при $p \neq 0{,}5$ розподіл несиметричний, то для заданого рівня довіри γ знайдемо найменший інтервал (p_1; p_2), для якого ймовірність знаходження p лівіше p_1 чи правіше p_2 дорівнює $\frac{1-\gamma}{2} = \frac{\alpha}{2}$. Значення p_1 та p_2 можна знайти, розв’язавши такі рівняння відносно p:

$$\sum_{m=0}^{\tilde{m}-1} C_n^m p^m q^{n-m} = \frac{\alpha}{2};$$
$$\sum_{m=\tilde{m}}^{n} C_n^m p^m q^{n-m} = \frac{\alpha}{2},$$

де $\tilde{m} = wn$.

З першого рівняння знайдемо p_1, а з другого — p_2. У результаті дістанемо інтервальну оцінку $p_1 < p < p_2$ — довірчий інтервал, який з надійністю γ накриває частку p ознаки генеральної сукупності. Для знаходження p_1 та p_2 є спеціальні таблиці, за якими за заданих n, α, w визначаємо інтервал (p_1; p_2). Зауважимо, що цей спосіб знаходження p_1 та p_2 ефективний лише для вибірок невеликого обсягу ($n < 20.$).

Нехай обсяг повторної вибірки досить великий $(n \geq 20)$, а частка p не дуже мала $(np \geq 10)$. За цих умов біноміальний розподіл добре апроксимується нормальним розподілом з параметрами (див. § 3):

$$M(w) = M\left(\frac{m}{n}\right) = p; \quad D\left(\frac{\mathrm{m}}{\mathrm{n}}\right) = \frac{pq}{n} \text{ або } \sigma\left(\frac{m}{n}\right) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}} = \sigma_{\text{г}}.$$

Нормальний розподіл симетричний і тому розрахунок меж довірчого інтервалу для p можна здійснити за функцією Лапласа.

Обчислимо

$$P(|w-p| < \delta) = P(w - \delta < p < w + \delta) = P(p - \delta < w < p + \delta) = \Phi\left(\frac{p+\delta-p}{\sqrt{\frac{pq}{n}}}\right) -$$
$$-\Phi\left(\frac{p-\delta-p}{\sqrt{\frac{pq}{n}}}\right) = \Phi\left(\frac{\delta\sqrt{\mathrm{n}}}{\sqrt{pq}}\right) - \Phi\left(-\frac{\delta\sqrt{\mathrm{n}}}{\sqrt{pq}}\right) = 2\Phi\left(\delta\sqrt{\frac{n}{pq}}\right) = 2\Phi(t) = \gamma.$$

Отже, $P(w - \delta < p < w + \delta) = 2\Phi(t) = \gamma$, де $t = \delta\sqrt{\frac{n}{pq}}$

$$\text{або } P(|w-p| < \delta) \approx 2\Phi\left(\delta\sqrt{\frac{n}{pq}}\right). \tag{4.16}$$

Гранична похибка $\delta = t\sqrt{\frac{pq}{n}} = t\sqrt{\frac{p(1-p)}{n}}$. Оскільки p — невідоме, то ми не можемо знайти довірчий інтервал, бо δ залежить від p. Тому нерівність

$$|w-p|<t\sqrt{\frac{p(1-p)}{n}} \tag{4.17}$$

розв'яжемо відносно p. Нерівність (4.17) рівносильна нерівності

$$|w-p|^2<t^2\,\frac{p(1-p)}{n}\Leftrightarrow\left(t^2+n\right)p^2-\left(2nw+t^2\right)p+nw^2<0.$$

Розв'язком цієї нерівності є довірчий інтервал $p_1<p<p_2$,

де

$$p_1=\frac{2nw+t^2-t\sqrt{4nw(1-w)+t^2}}{2\left(n+t^2\right)},$$

$$p_2=\frac{2nw+t^2+t\sqrt{4nw(1-w)+t^2}}{2\left(n+t^2\right)}.$$

Інтервал $\left(p_1;p_2\right)$ з надійністю $\gamma=2\Phi(t)$ накриває частку p.

Якщо обсяг вибірки n досить великий, то у співвідношенні $\delta=t\sqrt{\frac{p(1-p)}{n}}$ можна p замінити його точковою оцінкою $p\approx w$. Тоді

$$\delta=t\sqrt{\frac{w(1-w)}{n}}\,. \tag{4.18}$$

Підставивши δ у формулу (4.15), дістанемо довірчий інтервал для частки p ознаки генеральної сукупності:

$$w-t\sqrt{\frac{w(1-w)}{n}}<p<w+t\sqrt{\frac{w(1-w)}{n}}. \tag{4.19}$$

Формула (4.18) поєднує три величини: рівень довіри γ, граничну похибку δ та обсяг вибірки n. Знаючи дві з цих величин, можна знайти третю.

Проте для визначення необхідного обсягу вибірки з (4.18) ми маємо формулу:

$$n=\frac{t^2w(1-w)}{\delta^2}, \tag{4.20}$$

до якої входить w, яке визначається за вибіркою обсягу n. А ми маємо визначити n до організації вибірки. Оскільки w невідоме, то визначимо w з розрахунку, щоб обсяг вибірки був максимальним. За умовою екстремуму $n(w)$, як квадратна функція відносно w, матиме максимум при $w=\frac{1}{2}$ (вершина параболи $y=w(t-w)$).

Отже, підставивши у формулу (4.20) значення $w=\frac{1}{2}$, дістанемо

$$n_{\max}=\frac{t^2}{4\delta^2}. \tag{4.21}$$

Приклад 4.8. Знайти обсяг повторної вибірки, створеної для визначення частки виробів другого сорту в партії із 10 000 виробів. Гранична похибка вибірки не повинна перевищувати 5 % з надійністю 0,99.

Р о з в ' я з а н н я. За умовою $\delta=0{,}05,\gamma=0{,}99$ і за формулою (4.21) маємо

$$n_{\max}=\frac{\frac{1}{4}\cdot(2{,}575)^2}{(0{,}05)^2}=663{,}06\,.$$

Оскільки n — ціле число, то необхідний обсяг вибірки $n_{\max}=664$.

Приклад 4.9. Знайти довірчий інтервал для частки виробів другого сорту задачі 4.8, якщо у вибірці із 664 виробів частка другосортних виробів становить 80 %.

Р о з в ’ я з а н н я. За умовою $w = 0{,}8$, $n = 664$, $t = 2{,}575$.

За формулою (4.19) маємо :

$$0{,}8 - 2{,}575\sqrt{\frac{0{,}8\cdot 0{,}2}{664}} < p < 0{,}8 + 2{,}575\sqrt{\frac{0{,}8\cdot 0{,}2}{664}} \Rightarrow$$

$$0{,}8 - 0{,}04 < p < 0{,}8 + 0{,}04 \Rightarrow$$

$$0{,}76 < p < 0{,}84.$$

Приклад 4.10. Під час розбракування партії товару було відібрано методом випадкової вибірки 200 одиниць товару. Частка браку у вибірці склала 3 %. Обчислити з надійністю 0,96 граничну похибку вибіркової частки браку. Знайти довірчий інтервал для середнього квадратичного відхилення $\sigma_г$ частки браку з надійністю 0,95.

Р о з в ’ я з а н н я. Гранична похибка обчислюється за формулою (4.18):

$$\delta = t\sqrt{\frac{w(1-w)}{n}} = 2{,}055\sqrt{\frac{0{,}03\cdot 0{,}97}{200}} = 0{,}024788 \approx 0{,}025.$$

Отже, частка браку в генеральній сукупності з надійністю 0,96 знаходиться в межах (3 % – 2,5 %; 3 % + 2,5 %), тобто від 0,5 % до 5,5 % .

Для знаходження довірчого інтервалу для $\sigma_г$ скористаємося формулою (4.11). За умовою n = 200, $\gamma = 0{,}95$. З табл. 10 маємо $q(0{,}95;200) = 0{,}1$. Знайдемо значення s:

$$s = \sqrt{\frac{n}{n-1}w(1-w)} \Rightarrow s = \sqrt{\frac{200}{199}\cdot 0{,}03\cdot 0{,}97} \approx 0{,}171.$$

Отже, $0{,}171(1-0{,}1) < \sigma_г < 0{,}171(1+0{,}1) \Rightarrow 0{,}154 < \sigma_г < 0{,}188.$

4.8. ОЦІНКА ЧАСТКИ *p* ЗА БЕЗПОВТОРНОЮ ВИБІРКОЮ

Формули (4.18—4.21), одержані в п. 6 для повторної вибірки, мають місце і для безповторної вибірки за умови (див. § 3), що:

$$\sigma(w) = \sigma\left(\frac{m}{n}\right) \approx \sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)} = \sqrt{\frac{p(1-p)}{n}}\cdot\sqrt{1-\frac{n}{N}}. \qquad (4.22)$$

Формула для визначення необхідного обсягу безповторної вибірки з урахуванням (4.22) матиме вигляд

$$n \approx \frac{p(1-p)t^2 N}{N\delta^2 + p(1-p)\,t^2}\,. \qquad (4.23)$$

Це випливає з того, що (див. § 3) похибка

$\delta = \frac{t\sigma_г}{\sqrt{n}} = \frac{t\sqrt{p(1-p)}}{\sqrt{n}}$ — для повторної вибірки;

$\delta = t\frac{\sqrt{p(1-p)}}{\sqrt{n}}\cdot\sqrt{1-\frac{n}{N}}$ — для безповторної вибірки.

Формула (4.23) потрібна для оцінки p, але p, як правило, невідоме. Тому визначимо $n_{\max}$. Величина n як функція p досягне максимуму за умови, що

$$\frac{dn}{dp} = \frac{t^2N^2\delta^2(1-2p)}{(N\delta^2 + \left(p-p^2\right)\ t^2)^2} = 0 \Rightarrow 1-2p = 0 \Rightarrow p = \frac{1}{2}\,.$$

Звідси маємо формулу для визначення обсягу безповторної вибірки

$$n_{\max} = \frac{t^2N}{4N\delta^2 + t^2}\,. \qquad (4.24)$$

Зауваження 1. Якщо обсяг n безповторної вибірки значно менший, ніж обсяг N усієї сукупності, то $1-\frac{n}{N}\approx 1,$ і можна користуватися відповідними формулами для повторної вибірки.

Зауваження 2. Якщо $\sigma_{г}$ невідоме, то у формулах повторної і безповторної вибірки будемо користуватись оцінкою

$$\sigma_{г}\approx s=\sqrt{\frac{n}{n-1}}\,\sigma_{в}=\frac{\sqrt{w(1-w)}}{\sqrt{n-1}} \quad . \tag{4.25}$$

Приклад 4.11. Знайти необхідний обсяг вибірки у прикладі 4.8 за умови, що вибірка безповторна.

Р о з в ' я з а н н я. Оскільки p невідоме, то знайдемо максимальний обсяг вибірки за формулою (4.24):

$$n_{\max}=\frac{10\,000\cdot(2{,}575)^2}{4\cdot 10\,000(0{,}05)+(2{,}575)^2}=\frac{10\,000\cdot 6{,}630625}{40\,000\cdot 0{,}0025+6{,}630625}=\frac{66306{,}25}{100+6{,}630625}=\frac{66306{,}25}{106{,}630625}\approx 621{,}18.$$

Оскільки $n_{\max}$ ціле число, то $n_{\max}=622$.

Приклад 4.12. Для визначення якості насіння соняшника була здійснена вибірка 600 насінин методом випадкового відбору. Частка якісних насінин у вибірці склала 95 %. З якою надійністю можна вважати частку якісного насіння в усій сукупності рівною приблизно 95 %, якщо допустима похибка у визначенні цієї частки дорівнює 2 %.

Розв'язання. За умовою $w=0{,}95, n=600, \delta=0{,}02.$ Оскільки обсяг генеральної сукупності невідомий, то застосуємо формулу (4.16) для повторної вибірки $P(|w-p|<0{,}02)=2\Phi\left(0{,}02\sqrt{\frac{600}{0{,}95\cdot 0{,}05}}\right)=$ $=2\Phi(2{,}25)=0{,}9756$.

Отже, можна вважати, що якість насіння соняшника в усій сукупності дорівнює 95 % з надійністю 0,9756. Такий самий результат матимемо для безповторної вибірки, бо за умови невідомого обсягу генеральної сукупності використовуємо формулу (4.16) як для повторної, так і для безповторної вибірки.

Приклад 4.13. (Розв'яжіть самостійно). На заводах корпорації працює 2500 робітників. Для визначення середньої заробітної плати здійснили повторну вибірку обсягом 401 робітник. На підставі цього обстеження встановили:

1) середня заробітна плата становить 1500 грн;
2) середнє квадратичне відхилення по зарплаті становить 120 грн.

Визначити можливі межі середньої заробітної плати робітників з надійностями 0,6827; 0,9545; 0,9973. Поясніть одержані результати.

Відповіді: (1494; 1506); (1488; 1512); (1482; 1516).

§ 5. ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

5.1. ЗАГАЛЬНІ ПОНЯТТЯ ПЕРЕВІРКИ СТАТИСТИЧНИХ ГІПОТЕЗ

Перевірка гіпотез має різноманітне застосування, як у бізнесі та економіці, так і в багатьох інших сферах. Перевірка гіпотез — основна процедура підбиття підсумків про поведінку генеральної сукупності.

Інформація, одержана у результаті дослідження вибірки генеральної сукупності, може використовуватися для деяких висновків щодо всієї генеральної сукупності. Наприклад, на підприємстві про якість продукції роблять висновки за результатами вибіркового контролю. Якщо вибіркова частка браку не перевищує заздалегідь установленої (нормативної) величини p_0, то партія продукції приймається. Оскільки висновки про відповідність якості продукції встановленим вимогам робляться на підставі вибіркової перевірки, судження про якість продукції не можна розглядати як категоричне. Тут ідеться лише про припущення (гіпотезу), що частка браку в усій партії (генеральній сукупності) менша або дорівнює p_0.

Часто буває необхідним знати закон розподілу досліджуваної ознаки генеральної сукупності. Якщо закон розподілу невідомий, але є міркування для припущення певного його вигляду (назвемо його А, де як А може виступати рівномірний, показниковий, нормальний розподіл тощо), тоді висувають ***гіпотезу Н: ознака генеральної сукупності розподілена за законом А.*** У цій гіпотезі йдеться про вигляд невідомого розподілу.

Іноді закон розподілу ознаки генеральної сукупності відомий, але його параметри (числові характеристики) невідомі. Якщо є міркування припустити, що невідомий параметр θ дорівнює певному значенню θ_0, то висувають *гіпотезу Н:* $\theta = \theta_0$. Ця гіпотеза вказує на припущену величину параметра відомого розподілу.

Можливі також інші гіпотези: про рівність параметрів ознак двох різних розподілів, про незалежність вибірок тощо.

Означення 1. ***Статистичною гіпотезою*** називається будь-яке припущення стосовно вигляду розподілу або числового значення параметрів невідомого закону розподілу генеральної сукупності, зроблене внаслідок дослідження вибірки з цієї генеральної сукупності.

Статистичні гіпотези поділяють на ***параметричні та непараметричні.*** У ***параметричних*** статистичних гіпотезах містяться твердження про значення параметрів генеральної сукупності. Усі інші статистичні гіпотези називають ***непараметричними***.

Наприклад, висувається гіпотеза про числове значення середньої генеральної сукупності μ, генеральної дисперсії σ^2 тощо. Це приклади параметричних гіпотез.

Прикладом непараметричних статистичних гіпотез є гіпотези, які тестують дані, що не мають стосунку до числових значень параметрів генеральної сукупності. Наприклад, непараметричні статистичні гіпотези містять припущення про закони розподілу ознак генеральної сукупності, про незалежність вибіркових сукупностей тощо.

Гіпотези можуть містити припущення стосовно одного або більше одного числових значень параметра генеральної сукупності.

Означення 2. Гіпотезу називають ***простою***, якщо вона містить припущення стосовно одного і тільки одного числового значення параметра генеральної сукупності.

Наприклад, якщо λ — параметр показникового розподілу, то гіпотеза $H:\lambda = 5$ — проста.

Означення 3. Гіпотезу називають ***складною***, якщо вона складається із скінченної або нескінченної кількості простих гіпотез.

Наприклад***, гіпотеза Н: математичне сподівання нормального розподілу дорівнює 2*** — складна гіпотеза, бо середнє квадратичне відхилення σ невідоме і може набувати будь-якого значення.

Гіпотеза H ***: показниковий розподіл має параметр*** $\lambda > 2$ складна гіпотеза, оскільки складається із нескінченної множини простих гіпотез $H_k:\lambda = c_k$, де $c_k > 2, k = 1,2,...$.

Розв'язування кожної задачі перевірки статистичних гіпотез завжди починається з визначення ***основної та альтернативної гіпотез***. Основну гіпотезу називають ***нульовою гіпотезою*** та позначають H_0. Одночасно з основною гіпотезою H_0 завжди розглядають ***альтернативну їй гіпотезу*** H_1, яка конкурує з нульовою гіпотезою.

Означення 4. ***Основною (нульовою) статистичною гіпотезою*** H_0 називають припущене твердження, яке вважають правильним, якщо не доведене інше (альтернативне).

Означення 5. ***Альтернативною (конкуруючою) статистичною гіпотезою*** H_1 називають твердження, яке вважається правильним, якщо нульова гіпотеза відхилена.

Наприклад, для нульової гіпотези $H_0 : \mu = 6$ (середня генеральної сукупності дорівнює 6) можна вказати один із трьох варіантів визначення альтернативної гіпотези: $H_1:\mu \neq 6,; H_1:\mu > 6$, $H_1:\mu < 6.;$

Після визначення основної та альтернативної гіпотез, за домовленістю, робиться припущення про те, що нульова гіпотеза є справедливою, після чого відбувається перевірка достовірності саме основної нульової гіпотези. Після закінчення процесу перевірки достовірності основної гіпотези H_0 робиться висновок стосовно достатності або недостатності існуючих статистичних фактів на користь підтримки альтернативної гіпотези H_1.

Перевірка нульової гіпотези H_0 проводиться статистичними методами, тому її називають ***статистичною перевіркою гіпотези***. Основна нульова гіпотеза, яка на початковому етапі перевірки завжди вважається правильною, насправді може бути як правильною, так і помилковою. Тому за результатами статистичної перевірки нульової гіпотези може бути прийнято як правильне, так і помилкове рішення. Правильне рішення може бути прийнято у двох випадках: коли, за результатами перевірки не відхиляється правильна нульова гіпотеза та відхиляється хибна нульова гіпотеза. Помилкове рішення може бути прийнято теж у двох випадках: коли відхиляється правильна нульова гіпотеза та не відхиляється хибна. Або, інакше кажучи, у результаті прийняття помилкового рішення можуть бути допущені ***помилки двох типів:***

1) буде відхилено правильну нульову гіпотезу (помилка першого типу);

2) не буде відхилено хибну нульову гіпотезу (помилка другого типу).

Отже, доцільно відрізняти як гіпотези H_0 *і* H_1, так і наслідки помилок першого та другого типів, які виникають під час перевірки статистичних гіпотез.

Розглянемо ці питання на прикладі перевірки достовірності гіпотези H_0 у сфері світового кримінального права.

Коли особа звинувачується у злочині, вона постає перед судом. Звинувачення передає випадок злочину на розгляд суду присяжних, який має винести вирок на підставі наданих доказів. Фактично, присяжними здійснюється перевірка двох гіпотез. Перша гіпотеза: відповідач винен, друга гіпотеза: відповідач невинен. Яку з них позначити за нульову, а яку за альтернативну? Оскільки в кримінальному судочинстві завжди діє презумпція невинуватості, основною припущеною гіпотезою буде така гіпотеза:

H_0*: відповідач не винен.*

Альтернативною або досліджуваною буде гіпотеза:

H_1 *відповідач винен.*

Звичайно, присяжні не знають, котра з гіпотез справедлива. Вони мають винести рішення, спираючись на докази, надані як звинуваченням, так і захистом. Існує лише два можливих рішення — засудити чи виправдати відповідача. Мовою статистики, засудження відповідача рівнозначне відхиленню нульової гіпотези на користь альтернативної, або підтвердження присяжними достатності доказів, наданих звинуваченням для висновку про те, що відповідач винен. За умови недостатності доказів для визнання відповідача винним присяжні ухвалюють рішення про його виправдання, що мовою статистики означає «недостатність статистичних доказів для відхилення нульової гіпотези на користь альтернативної».

Існує два можливих типи помилок.

Помилка першого типу має місце за умови відхилення істинної нульової гіпотези. Згідно з наведеним прикладом, зазначена вище помилка матиме місце за умови засудження невинної особи.

Помилка другого типу має місце за умови невідхилення помилкової нульової гіпотези, або виправдання винного відповідача.

Імовірність помилки першого типу позначається α та називається ***рівнем значущості.*** Імовірність помилки другого типу позначається β. Імовірності помилок α та β є взаємопов'язаними, спроба знизити одну з них призведе до збільшення іншої.

У кримінальному судочинстві помилка першого типу трапляється, коли невинну особу помилково засуджують. Помилка другого типу трапляється, коли винний відповідач виправдовується.

У табл. 5.1 наведено узагальнену термінологію та поняття перевірки гіпотез.

Таблиця 5.1

ТЕРМІНОЛОГІЯ ТА ПОНЯТТЯ ПЕРЕВІРКИ ГІПОТЕЗ

ВИСНОВОК	H_0 істинна (відповідач невинен)	H_0 помилкова (відповідач винен)
ВІДХИЛИТИ H_0 (Засудити відповідача)	Помилка першого типу P (імовірність помилки першого типу) = α	Правильне рішення (Справедливий вирок)
НЕ ВІДХИЛЯТИ H_0 (Виправдати відповідача)	Правильне рішення (Справедливий вирок)	Помилка другого типу P (імовірність помилки другого типу) = β

У системі правосуддя засудження невинної особи (помилка першого типу) вважається більш серйозною. Як наслідок, система встановлює, що ймовірність наявності помилки першого типу має бути як можна меншою. Це забезпечується тим, що обов'язок доведення вини покладається на звинувачення (звинувачення зобов'язане довести вину, захист не повинен доводити нічого) та інструкціями, які судді дають присяжним визнавати відповідача винним тільки якщо «доказ переважає розумний сумнів». За відсутності достатніх доказів, присяжні мають винести вердикт виправдання, навіть якщо можуть бути деякі докази вини. Як наслідок таких заходів, імовірність виправдання винних людей досить велика.

Взаємозв'язок між імовірностями помилок першого та другого типу системою правосуддя трактується так: «Краще виправдати 100 винних осіб, ніж засудити одну невинну», тому ймовірність помилки першого типу має становити 1/100 від імовірності помилки другого типу.

Означення 6. ***Імовірність припущення помилки першого типу*** — це ймовірність невідхилення альтернативної гіпотези за умови, що нульова гіпотеза справедлива:

$$\alpha = P(H_1/H_0), \tag{5.1}$$

$$\updownarrow$$

$$\alpha = P(\text{невідхилення } H_1/H_0 \text{ правильна.}$$

Найчастіше рівень значущості α задається наперед. У статистичних дослідженнях найчастіше використовують такі його значення: 0,001; 0,005; 0,01; 0,05. Наприклад, прийняття рівня значущості 0,05 означає, що в п'яти випадках із ста ми ризикуємо отримати помилку першого типу.

Імовірність не припуститися помилки першого типу називається ***рівнем надійності*** та позначається γ.

$$\gamma = 1 - \alpha = 1 - P(H_1/H_0) = P(H_0/H_0), \tag{5.2}$$

$$\updownarrow$$

$$\gamma = P(\text{невідхилення } H_0/H_0 \text{ правильна}).$$

Зауваження 1. Під час контролю якості продукції ймовірність визнати нестандартними стандартні вироби називають ризиком виробника, а ймовірність визнати придатними браковані вироби називають ризиком споживача.

Означення 7. ***Імовірність припущення помилки другого типу*** — це ймовірність невідхилення хибної гіпотези H_0:

$$\beta = P(H_0/H_1), \tag{5.3}$$

$$\updownarrow$$

$$\beta = P(\text{невідхилення } H_0/H_1 \text{ правильна}).$$

Зауваження 2. Єдиним способом одночасного зменшення ймовірностей похибок першого та другого типу є збільшення обсягу вибірки.

Основні положення перевірки гіпотез:

1. Існує два типи гіпотез. Одна з них називається нульовою гіпотезою, а інша — альтернативною або досліджуваною гіпотезою.
2. Процес перевірки гіпотези завжди починається з припущення, що нульова гіпотеза справедлива. Метою процесу перевірки є визначення достатності або недостатності статистичних фактів на користь підтримки альтернативної гіпотези.
3. Існує два типи можливих висновків:
   - статистичних фактів для підтвердження альтернативної гіпотези достатньо;
   - статистичних фактів для підтвердження альтернативної гіпотези недостатньо.
4. Під час перевірки нульової гіпотези можливі два типи помилок. Помилка першого типу має місце за умови відхилення достовірної нульової гіпотези. Помилка другого типу має місце за умови прийняття рішення про невідхилення помилкової нульової гіпотези.

Статистичні критерії для перевірки гіпотез

Правило, за яким приймається рішення про відхилення або невідхилення статистичної гіпотези H_0, називається ***статистичним критерієм***.

Перевірку статистичної гіпотези можна здійснити лише з використанням даних вибірки. Для цього слід обрати деяку випадкову статистичну характеристику (вибіркову функцію), точний або наближений розподіл якої відомий, і за допомогою цієї характеристики здійснити перевірку основної нульової гіпотези. Позначимо обрану випадкову статистичну характеристику (вибіркову функцію) через U.

Означення 8. ***Статистичним критерієм перевірки гіпотези*** (або просто критерієм) називають випадкову величину U, розподіл якої (точний або наближений) є відомим, та застосовується для перевірки справедливості основної гіпотези.

Іншими словами, визначення статистичного критерію — це визначення правила перевірки основної статистичної гіпотези.

Означення 9. ***Потужність критерію*** — це ймовірність відхилення хибної гіпотези H_0, або ймовірність запобігання помилки другого типу:

$$1 - \beta = P(H_1/H_1), \tag{5.4}$$

$$\Updownarrow$$

$$1 - \beta = P(\text{невідхилення } H_1/H_1 \text{ правильна}).$$

Зауваження 3. Якщо випадкова величина U розподілена за нормальним законом, то критерій позначають не літерою U, а літерою Z. Якщо статистична характеристика розподілена за законом Фішера — Снедекора, то її позначають F. У разі розподілу статистичної характеристики за законом Стьюдента її позначають T, а у разі розподілу за законом «*хі*— квадрат» — χ^2.

Зауваження 4. За великих обсягів вибірки ($n > 30$) закони розподілу статистичних критеріїв T, F, χ^2 наближаються до нормального розподілу.

Після визначення статистичного критерію перевірки гіпотези, обчислюється спостережуване значення u^* статистичного критерію U за даними вибірки.

Означення 10. ***Спостережуваним значенням u* статистичного критерію U*** називають значення відповідного критерію, обчислене за даними вибірки.

Критична область та критичні точки

Множину всіх можливих значень статистичного критерію U поділяють на дві підмножини A та $\overline{A}$, для яких виконується умова: $A \cap \overline{A} = \varnothing,\ A \cup \overline{A} = \Omega$.

Означення 11. ***Областю допустимих значень*** A називають множину значень критерію, за яких основна гіпотеза H_0 не відхиляється.

Означення 12. ***Критичною областю*** $\overline{A}$ називають сукупність значень критерію, за яких основна гіпотеза H_0 відхиляється.

Означення 13. ***Критичними точками*** (межами) критерію U називають точки $u_{кр}$, які відокремлюють критичну область $\overline{A}$ від області допустимих значень A.

Якщо спостережуване за даними вибірки значення u^* потрапляє в критичну область $\overline{A}$, то гіпотеза H_0 відхиляється.

Якщо спостережуване за даними вибірки значення u^* потрапляє в область допустимих значень A, то гіпотеза H_0 не відхиляється.

Порядок визначення критичних точок та критичних областей

Критичні точки — межі критичних областей — знаходять із таблиць розподілу ймовірностей випадкової величини U, відповідно до заданого рівня значущості.

Розрізняють три види критичних областей: ***правостороння, лівостороння і двостороння.***

Означення 14. ***Правосторонньою критичною областю*** є критична область, яка задається нерівністю: $U > u_{кр}$ (рис. 5.1).

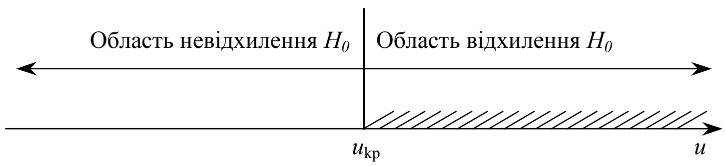


Рис. 5.1

Критичну точку $u_{кр}$ цієї області за обраного рівня значущості α визначають зі співвідношення:

$$P(U > u_{кр}) = \alpha. \tag{5.5}$$

Означення 15. ***Лівосторонньою критичною областю*** є критична область, яка задається нерівністю: $U < u_{кр}$ (рис. 5.2).

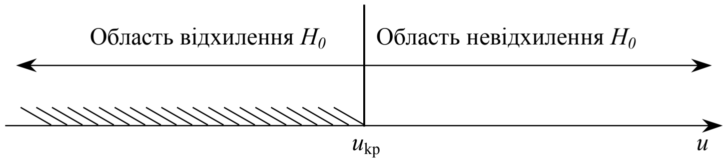


Рис 5.2

Значення $u_{кр}$ знаходять за умовою:

$$P(U < u_{кр}) = \alpha. \tag{5.6}$$

Означення 16. ***Двосторонньою критичною областю*** є критична область, яка задається двома нерівностями : $U < u_{\text{кр}}^{\text{лів}}\, u_{\text{кр}}$, $U > u_{\text{кр}}^{\text{пр}}$ (рис. 5.3).

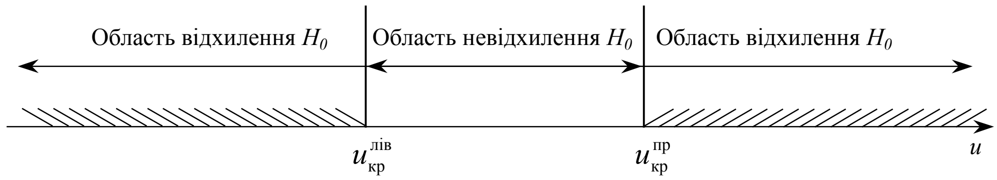


Рис. 5.3

Критичні точки $u_{\text{кр}}^{\text{лів}}$, $u_{\text{кр}}^{\text{пр}}$ знаходять за умови:

$$P(U < u_{\text{кр}}^{\text{лів}}) + P(U > u_{\text{кр}}^{\text{пр}}) = \alpha. \tag{5.7}$$

У разі, коли двостороння критична область симетрична відносно нуля:

$$P\left(U < u_{\text{кр}}^{\text{лів}}\right) = P\left(U > u_{\text{кр}}^{\text{пр}}\right) = \frac{\alpha}{2} \text{ і } u_{кр}^{пр} = -u_{\text{кр}}^{\text{лів}} = u_{\text{кр}},$$

де $u_{кр}$ визначається з умови:

$$P\left(U > u_{кр}\right) = \frac{\alpha}{2}. \tag{5.8}$$

Загальний алгоритм перевірки статистичної гіпотези

Розглянемо загальний підхід до перевірки правильності нульової гіпотези.

Крок перший. Формулюються гіпотези: нульова H_0 та альтернативна H_1.

Крок другий. Вибирається статистичний критерій $U \in \{\text{Z}, \text{T}, F, \chi^2, ...\}$, який відповідає сформульованій гіпотезі, тобто визначається закон розподілу для обраної вибіркової статистики (обрана вибіркова статистика — це статистична вибірка).

Крок третій. За даними вибірки обчислюється спостережуване значення u^* критерію *U*.

Крок четвертий. Залежно від змісту альтернативної гіпотези за вибраним статистичним критерієм U та рівнем значущості α визначається критична область (правостороння, лівостороння або двостороння) та визначаються критичні точки розподілу.

Наприклад, для перевірки нульової гіпотези: $\text{H}_0 : \mu = \mu_0$ (*середнє значення генеральної сукупності дорівнює* μ_0*)*, залежно від вигляду альтернативної гіпотези можливі три варіанти побудови різних критичних областей:

1) якщо: $\text{H}_1 : \mu > \mu_0$, будується правостороння критична область;

2) якщо $\text{H}_1 : \mu < \mu_0$, будується лівостороння критична область;

3) якщо $\text{H}_1 : \mu \neq \mu_0$, будується двостороння критична область.

Крок п'ятий. Формулюється висновок.

Якщо u^* — спостережуване значення критерію *U* належить критичній області, можемо стверджувати про достатність статистичних фактів для підтвердження альтернативної гіпотези H_1 (гіпотеза H_0 відхиляється).

Якщо u^* — спостережуване значення критерію *U* належить області допустимих значень, стверджуємо про недостатність статистичних фактів для підтвердження альтернативної гіпотези H_1 (гіпотеза H_0 не відхиляється).

Прикладні задачі перевірки статистичних гіпотез буде розглянуто у наступних пунктах. До основних категорій перевірки статистичних гіпотез можна віднести такі:

1) про числові значення параметрів генеральних сукупностей (математичного сподівання, дисперсії, імовірності тощо);

2) про порівняння числових характеристик генеральних сукупностей;

3) про закони розподілу генеральних сукупностей тощо.

Вибір конкретного критерію перевірки в кожному із наведених випадків залежить від обсягу вибірки (великий чи малий) і наявності додаткової інформації щодо ознаки, яка досліджується.

5.2. ПЕРЕВІРКА ДОСТОВІРНОСТІ ГІПОТЕЗИ ПРО ЗНАЧЕННЯ СЕРЕДНЬОЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ, ЯКЩО ДИСПЕРСІЯ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ ВІДОМА. ІНТЕРПРЕТАЦІЯ ВЕЛИЧИНИ *P*-ЗНАЧЕННЯ (*P-VALUE*) ПІД ЧАС ПЕРЕВІРКИ СТАТИСТИЧНИХ ГІПОТЕЗ

Нехай X — нормально розподілена генеральна сукупність із генеральною середньою μ. Припустимо, що її генеральна дисперсія σ^2 — відома (із попередніх досліджень, знайдена теоретично, або обчислена за вибіркою великого обсягу, оскільки за вибіркою великого обсягу можна отримати якісну оцінку дисперсії генеральної сукупності). При цьому генеральна середня μ — невідома, але є вагомі причини стверджувати, що її гіпотетичне значення дорівнює μ_0. Для перевірки цього твердження слід обрати вибірку обсягом n із даної генеральної сукупності, обчислити її середнє значення $\overline{x}_в$, та визначити, наскільки значущою (за заданим рівнем значущості α) є різниця між припущеним (гіпотетичним) значенням μ_0 та середньою вибірковою $\overline{x}_в$.

Такі задачі трапляються під час розв'язання технологічних та економічних проблем, які характеризуються будь-яким середнім показником, наприклад: середній вміст нітратів у продукції, середня вологість сировини, середня продуктивність праці, середній обсяг реалізації продукції, середній прибуток підприємства тощо.

Розглянемо у загальному вигляді задачу про перевірку достовірності нульової гіпотези про рівність генеральної середньої μ ***та її гіпотетичного значення*** μ_0***, якщо дисперсія генеральної сукупності відома.***

Крок 1. Обираємо нульову гіпотезу

$$H_0 : \mu = \mu_0,$$

та одну із трьох альтернативних гіпотез:

$$H_1 : \mu > \mu_0,\ H_1 : \mu < \mu_0,\ H_1 : \mu \neq \mu_0.$$

Крок 2. Обираємо статистичний критерій (прогнозований розподіл вибіркової статистики).

Для ознаки X генеральної сукупності розподіленої за нормальним законом $X \sim N(\mu,\sigma)$, відомим є факт нормального розподілу відповідної випадкової величини: вибіркової середньої $\overline{X}_в \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ (див. §4).

Зауваження 5. Якщо закон розподілу генеральної сукупності невідомий, але обсяг вибірки, обраної із цієї сукупності, великий (містить понад 30 спостережень), то згідно з центральною граничною теоремою вибіркова середня $\overline{X}_в$ має наближено нормальний закон розподілу $\overline{X}_в \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Для формалізації процедури перевірки основної статистичної гіпотези $H_0 : \mu = \mu_0$ доцільно зробити перехід від випадкової величини $\overline{X}_в \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ до її стандартної форми $Z \sim N(0;1)$.

Для цього використаємо відоме правило: якщо від значень нормально розподіленої випадкової величини відняти її середню (математичне сподівання) та результат поділити на середнє квадратичне відхилення, то отримаємо нормовану нормальну випадкову величину (стандартну форму випадкової величини $\overline{X}_в$) : $Z = \frac{\overline{X}_в - \mu}{\sigma(\overline{X}_в)}$, $Z \sim N(0;1)$.

Правило визначення статистичного критерію.

Для перевірки достовірності гіпотези про значення середньої генеральної сукупності за відомої дисперсії обирається критерій Z, якщо:

1) обсяг вибірки великий: n > 30 (закон розподілу генеральної сукупності X може бути довільним),

2) обсяг вибіркиm n ≤ 30 (закон розподілу генеральної сукупності X обов'язково має бути нормальним).

Отже, за статистичний критерій обрано випадкову величину

$$Z = \frac{\overline{X}_в - \mu_0}{\sigma(\overline{X}_в)} = \frac{\overline{X}_в - \mu_0}{\sigma(X)}\sqrt{n}\text{, оскільки } \sigma(\overline{X}_в) = \frac{\sigma(X)}{\sqrt{n}}. \tag{5.9}$$

Випадкова величина Z має нормований нормальний закон розподілу ймовірностей: $Z \sim N(0;1)$, оскільки $M(Z) = 0$ і $\sigma(Z) = 1$.

Крок 3. За наявною вибіркою обчислюється спостережуване значення z* критерію Z:

$$z^* = \frac{\bar{x}_в - \mu_0}{\frac{\sigma(X)}{\sqrt{n}}} = \frac{(\bar{x}_в - \mu_0)}{\sigma(X)}\sqrt{n}\,. \tag{5.10}$$

Крок 4. Залежно від змісту альтернативної гіпотези за вибраним статистичним критерієм Z і рівнем значущості α визначається критична область (правостороння, лівостороння або двостороння) та визначаються критичні точки розподілу. При цьому можливі три випадки.

Випадок 1. Якщо альтернативна гіпотеза має вигляд: $H_1: \mu > \mu_0$, ***то будується правостороння критична область*** (рис. 5.4).

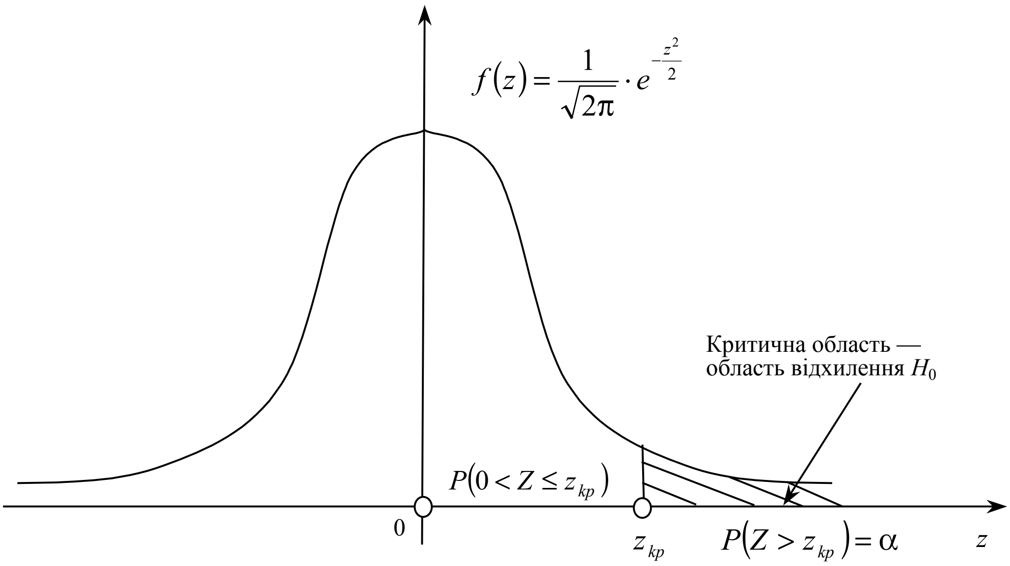


Рис. 5.4

Оскільки розподіл величини Z симетричний відносно нуля (рис. 5.4), то

$$P(0 < Z < \infty) = \frac{1}{2}, \text{ або } P(0 < Z \leq z_{кр}) + P(Z > z_{кр}) = \frac{1}{2}\,. \tag{5.11}$$

Ураховуючи, що для правосторонньої критичної області $P(Z > z_{кр}) = \alpha$, маємо:

$$P(0 < Z \leq z_{кр}) + \alpha = \frac{1}{2} \text{ або } P(0 < Z \leq z_{кр}) = \frac{1 - 2\alpha}{2}\,. \tag{5.12}$$

Знайдемо ймовірність потрапляння нормально розподіленої випадкової величини на проміжок $(0; z_{кр}]$.

Оскільки $\Phi(0) = 0$, маємо:

$$\Phi(z_{кр}) = \frac{1 - 2\alpha}{2}\,. \tag{5.13}$$

За таблицею значень функції Лапласа

знаходимо аргумент $z_{кр} = \Phi^{-1}\left(\frac{1 - 2\alpha}{2}\right)$. (5.14)

Випадок 2. За альтернативної гіпотези $H_1: \mu < \mu_0$ ***будується лівостороння критична область*** (рис. 5.5).

$z_{кр}$ знаходять з рівності $P(Z < z_{кр}) = \alpha$.

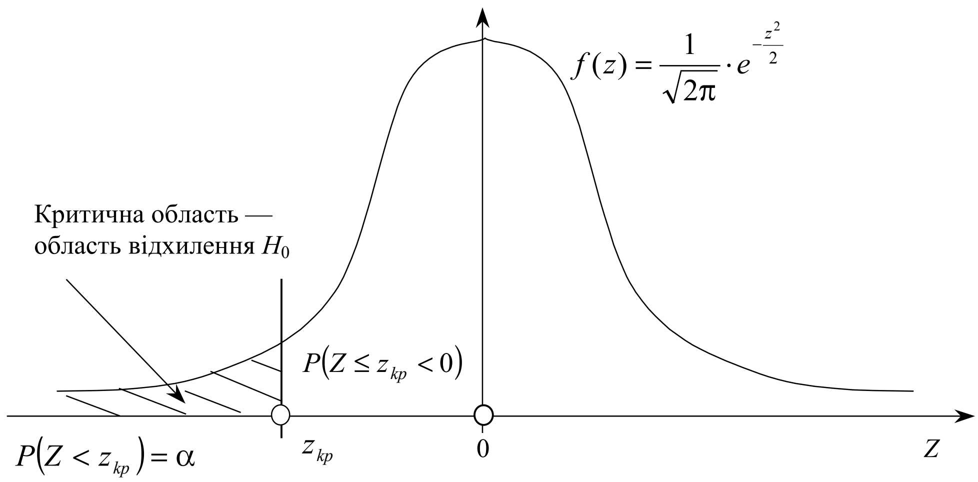


Рис. 5.5

Використовуючи рівність

$$P(Z < z_{кр}) + P(z_{кр} \le Z < 0) = \frac{1}{2}, \tag{5.15}$$

дістаємо:

$$P(z_{кр} \le Z < 0) = \frac{1-2\alpha}{2}, \tag{5.16}$$

або

$$\Phi(z_{кр}) = -\frac{1-2\alpha}{2}, \tag{5.17}$$

звідки

$$z_{кр} = \Phi^{-1}\left(-\frac{1-2\alpha}{2}\right)). \tag{5.18}$$

Критична точка $z_{кр}$ — межа лівосторонньої критичної області — завжди набуватиме від'ємних значень, оскільки вона знаходиться зліва від центру розподілу (математичного сподівання) випадкової величини.

За таблицею значень функції $\Phi(z)$ знаходимо значення z_1 із рівняння $\Phi(z_1) = \frac{1-2\alpha}{2}$. Ураховуючи непарність функції Лапласа, маємо $z_{кр} = -z_1 = -\Phi^{-1}\left(\frac{1-2\alpha}{2}\right)$.

Випадок 3. За альтернативної гіпотези $H_1 : \mu \ne \mu_0$ ***будується двостороння критична область*** (рис. 5.6).

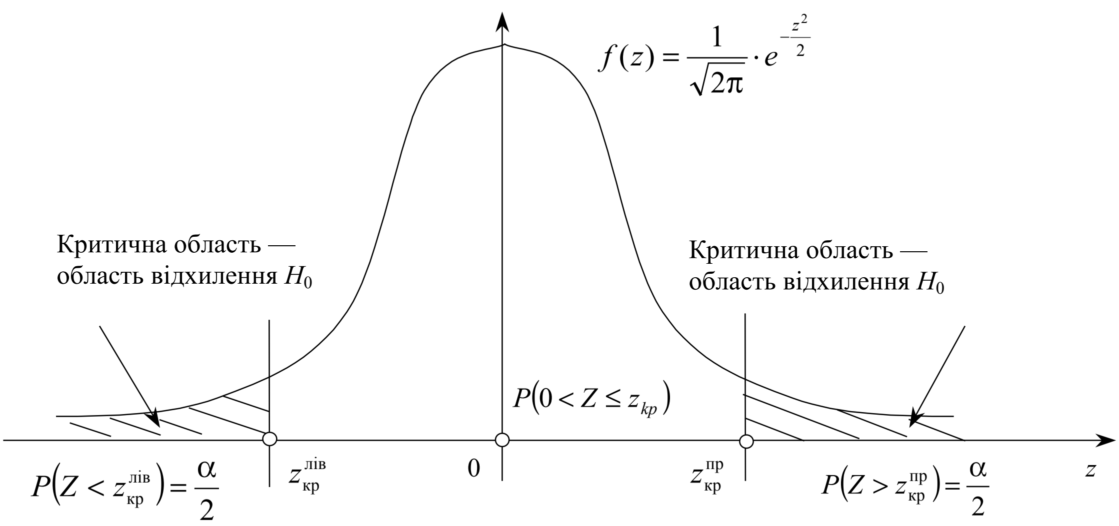


Рис. 5.6

Знаходимо дві критичні точки $z_{\text{кр}}^{\text{лів}} < z_{\text{кр}}^{\text{пр}}$ з умови:

$$P\left(Z < z_{\text{кр}}^{\text{лів}}\right) = P\left(Z > z_{\text{кр}}^{\text{пр}}\right) = \frac{\alpha}{2}, \text{ де } z_{\text{кр}}^{\text{пр}} = -z_{\text{кр}}^{\text{лів}} = z_{\text{кр}}. \tag{5.19}$$

Для знаходження значення $z_{кр}$ скористаємось рівністю:

$$P\left(0 < Z \le z_{кр}\right) + P\left(Z > z_{кр}\right) = \frac{1}{2}. \tag{5.20}$$

Ураховуючи, що $P(Z > z_{\text{кр}}) = \dfrac{\alpha}{2}$, дістаємо:

$$P\left(0 < Z \le z_{кр}\right) = \frac{1-\alpha}{2}, \text{ або } \Phi\left(z_{\text{кр}}\right) - \Phi(0) = \frac{1-\alpha}{2}. \tag{5.21}$$

Відповідно

$$\Phi\left(z_{\text{кр}}\right) = \frac{1-\alpha}{2}. \tag{5.22}$$

З останньої рівності та таблиці значень функції $\Phi(z)$ знаходимо

$$z_{\text{кр}} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right), \text{ де } z_{\text{кр}}^{\text{пр}} = -z_{\text{кр}}^{\text{лів}} = z_{\text{кр}};\ z_{\text{кр}}^{\text{пр}} = z_{\text{кр}};\ z_{\text{кр}}^{\text{лів}} = -z_{\text{кр}}. \tag{5.23}$$

Крок 5. Висновок. Якщо z* — спостережуване значення критерію Z — належить критичній області, то робимо висновок, що є достатньо статистичних фактів для невідхилення альтернативної гіпотези H_1 (достатньо статистичних фактів для відхилення гіпотези H_0). Якщо z* — спостережуване значення критерію Z — не належить критичній області, то робимо висновок, що статистичних фактів для підтвердження альтернативної гіпотези H_1 недостатньо (достатньо статистичних фактів для невідхилення гіпотези H_0).

Приклад 5.1 *(приклад побудови двосторонньої критичної області). Фахівець зі статистики отримав завдання перевірити дані, оголошені в доповіді: «середня заробітна плата робітника металургійної промисловості країни А склала 28 000 грош. од. за попередній рік».*

Р о з в ' я з а н н я. Висловлюється твердження про значення параметра генеральної сукупності. Однак проблема в тому, що ми не знаємо точного значення цього параметра насправді. Отже, щоб перевірити справедливість цього твердження, слід обрати вибірку із генеральної сукупності ($n > 30$) та дослідити, чи підтверджують висунуту гіпотезу отримані з вибірки дані.

Після створення та обробки результатів вибірки фахівець зі статистики порахував середню вибіркову $\bar{x}_{\text{в}}$, що дорівнює 26 500 грош. од.

За *відомим стандартним (середнім квадратичним) відхиленням* генеральної сукупності він знайшов стандартну помилку вибіркової середньої $\sigma(\bar{x}_{\text{в}})$ яка склала 1000 грош. од., та почав перевірку.

Крок 1. Обираються нульова та альтернативна гіпотези:

H_0: середня заробітна плата робітника за рік $\mu = \mu_0 = 28\,000$ грош. од.

H_1: середня заробітна плата робітника за рік $\mu \ne 28\,000$ грош. од.

Крок 2. Вибирається статистичний критерій для перевірки нульової гіпотези.

Оскільки середнє квадратичне відхилення генеральної сукупності відоме, статистичний критерій обираємо за формулою (5.9):

$$Z = \frac{\overline{X}_{\text{в}} - \mu_0}{\sigma(X_{\text{в}})},$$

де $Z \sim N(0; 1)$ за умови правильності нульової гіпотези.

Крок 3. За обраною статистикою обчислюється спостережуване значення статистичного критерію, як відношення різниці між середнім вибірки та припущеним значенням середнього генеральної сукупності до стандартного відхилення вибіркової середньої:

$$z^* = \frac{\bar{x}_{\text{в}} - \mu_0}{\sigma\left(\overline{X}_{\text{в}}\right)} = \frac{26\,500 - 28\,000}{1000} = -1{,}5.$$

Крок 4. Залежно від змісту альтернативної гіпотези за вибраним статистичним критерієм Z і рівнем значущості α будується критична область (правостороння, лівостороння або двостороння) та визначаються критичні точки розподілу.

Оскільки рівень значущості α не задано за умовою задачі, обираємо його самостійно. Зазвичай обирають рівень значущості 5 % .

Зауваження стосовно обраного рівня значущості. *Весь сенс перевірки гіпотези полягає в тому, щоб зрозуміти, наскільки великим є ризик отримання висновку про відхилення нульової гіпотези тоді, коли вона дійсно є правильною. Оскільки для висновку за результатами перевірки гіпотези використовується вибірка, ми повинні звернути увагу на те, що наявна помилка вибірки. Наприклад, стосовно країни А: ми не можемо відкинути нульову гіпотезу просто тому, що у вибірці середня заробітна плата робітника за рік не дорівнює 28 000 грош. од. Але тоді виникає питання: з якою граничною величиною похибки ми бажаємо провести аналіз: з тим, щоб визначити, чи є суттєвим відхилення між значенням середньої зарплати, отримане за результатами вибірки та її припущеним значенням, чи це лише наслідок помилки вибірки? Це питання, на яке дає відповідь вибір рівня значущості.*

Після визначення розподілу вибіркової статистики та вибору рівня значущості обчислюється діапазон значень спостережуваного значення критерію, для якого нульова гіпотеза не відхиляється. Проведемо аналогію діапазону невідхилення нульової гіпотези з футбольною сіткою та м'ячем. Обраний розподіл вибіркової статистики в поєднанні з вибором рівня значущості встановлює діапазон, в якому нульова гіпотеза, можливо, істинна (тобто ширина сітки). Тоді спостережуване значення z^* для вибірки буде м'ячем. Якщо м'яч потрапляє в сітку, тоді нульова гіпотеза, можливо, істинна і тому не може бути відхилена.

Якщо м'яч (спостережуване значення z^*) виходить за межі сітки, то робиться висновок, що нульова гіпотеза відхиляється для даного рівня значущості.

Пам'ятаємо, що ширина сітки побудована зі значною похибкою, тому, коли м'яч виходить за межі сітки, причиною є вже не помилка вибірки, а те, що нульова гіпотеза просто не відповідає дійсності.

Використовуючи обраний рівень значущості 5 % для альтернативної гіпотези H_1: середня заробітна плата робітника за рік $\mu \neq 28\,000$ грош. од., будуємо двосторонню критичну область.

Критичні значення, що розділяють області відхилення та невідхилення нульової гіпотези, знаходимо за формулами знаходження критичних точок для двосторонньої критичної області (5.22, 5.23):

$$\Phi\left(z_{\text{кр}}\right)=\frac{1-\alpha}{2}=\frac{1-0{,}05}{2}=0{,}475\,;\;\; z_{\text{кр}}=\Phi^{-1}\left(\frac{1-\alpha}{2}\right)=1{,}96\,.$$

Отже, –1,96 і 1,96 — критичні значення, що розділяють області відхилення та невідхилення нульової гіпотези (рис. 5.7).

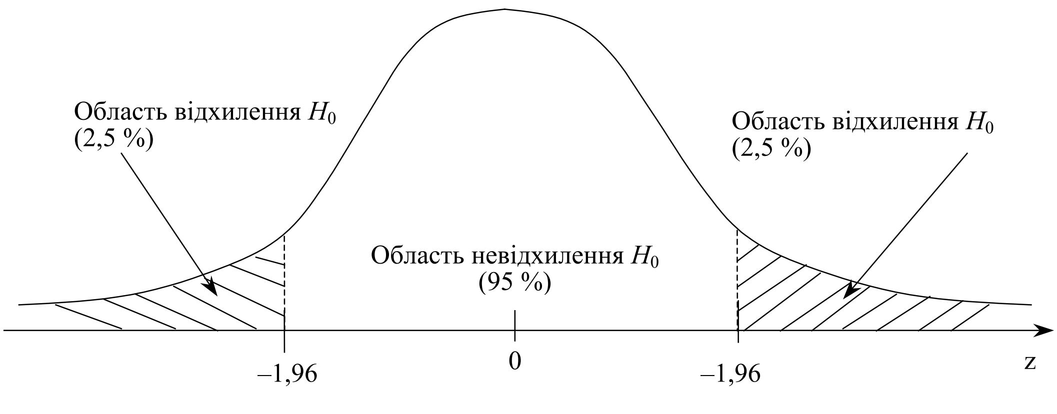


Рис. 5.7

Спостережуване значення $z^* = -1{,}5$ знаходиться на відстані 1,5 стандартної помилки ліворуч від середнього значення випадкової величини Z за умови, що нульова гіпотеза правильна. Для обраного рівня значущості 5 % межа відхилення нульової гіпотези буде знаходитися на відстані 1,96 стандартної помилки з обох боків від середнього значення Z. Отже, дане спостережуване значення статистичного критерію знаходиться в межах діапазону невідхилення нульової гіпотези. Отримане відхилення вибіркової середньої $\bar{x}_{\text{в}} = 26\,500$ від припущеного в нульовій гіпотезі значення $\mu_0 = 28\,000$ пояснюється випадковістю або помилкою вибірки.

Крок 5. Оскільки z^* — спостережуване значення критерію Z належить області його допустимих значень, то робимо висновок, що за рівня значущості 5 % статистичних фактів для підтвердження альтернативної гіпотези H_1 недостатньо.

Примітка (до розв'язування попереднього прикладу). Якби вибіркова середня склала б 22 000 грош. од., то спостережуване значення z^* вибірки дорівнювало б (22 000 – 28 000)/1000 = = –6,0, і це призвело б до відхилення нульової гіпотези. Іншими словами, відхилення вибіркової середньої від припущеного значення дуже велике, щоб пояснити це випадковістю чи помилкою вибірки. Такий великий розрив був би свідченням того, що значення параметра генеральної сукупності μ було іншим, ніж припущене в нульовій гіпотезі.

Отже, мале спостережуване значення z^* критерію показує, що результати вибірки не набагато відрізняються від передбачуваного значення μ_0, тим самим роблячи відмову від нульової гіпотези малоймовірною. Також відзначимо, що в міру збільшення обсягу вибірки стандартна помилка вибіркової середньої зменшується. Це приводить до збільшення значення z^*, тим самим збільшуючи ймовірність відхилення нульової гіпотези. Тому, коли розмір вибірки дуже великий, відмовитись від нульової гіпотези можна, навіть якщо вибіркова середня не набагато відрізняється від передбачуваного середнього значення μ_0. Іншими словами, коли розмір вибірки дуже великий, помилки вибірки будуть дуже малі, що дає нам упевненість, що вибірка є майже репрезентативною.

У наведеному вище прикладі два критичних значення: –1,96 і 1,96 визначили двосторонню область відхилення нульової гіпотези. Це сталося тому, що в нульовій гіпотезі стверджувалося, що значення сукупності точно дорівнювало 28 000 грош. од. Тому, якщо середня вибіркова набагато більша або набагато менша, ніж 28 000, то ми повинні відкинути нульову гіпотезу (див. рис. 5.7).

Приклад 5.2. (приклад побудови лівосторонньої критичної області).

Фахівець зі статистики отримав завдання перевірити дані, оголошені в доповіді: «середня заробітна плата робітника металургійної промисловості країни А склала не менше ніж 28 000 грош. од. за попередній рік».

Р о з в ' я з а н н я. Для перевірки даних, оголошених у доповіді, будемо використовувати дані, отримані з дослідження вибірки прикладу 1:

$$x_{\text{в}} = 26\,500 \text{ грош. од.}, \ \sigma(\bar{x}_{\text{в}}) = 1000 \text{ грош. од.}$$

Стандартне (середнє квадратичне) відхилення генеральної сукупності відоме.

Крок 1. Обираються нульова та альтернативна гіпотези:

H_0: середня заробітна плата робітника за рік $\mu \geq$ 28 000 грош. од. ($H_0 : \mu \geq \mu_0$);

H_1: середня заробітна плата робітника за рік $\mu <$ 28 000 грош. од. ($H_1 : \mu < \mu_0$).

***Крок 2**:* Визначається статистичний критерій для перевірки нульової гіпотези.

Оскільки середнє квадратичне відхилення генеральної сукупності відоме, статистичний критерій обираємо за формулою

$$Z = \frac{\overline{X}_{\text{в}} - \mu_0}{\sigma(\overline{X}_{\text{в}})}.$$

Крок 3. Обчислюється спостережуване значення статистичного критерію, як відношення різниці між середньою вибіркою та припущеним нульовою гіпотезою значенням середнього з генеральної сукупності до стандартного відхилення вибіркової середньої:

$$z^* = \frac{\bar{x}_{\text{в}} - \mu_0}{\sigma(\overline{X}_{\text{в}})} = \frac{26\,000 - 26\,000}{1000} = -1{,}5.$$

***Крок 4**:* Обираємо рівень значущості, критичну область (правостороння, лівостороння або двостороння), критичну точку.

Оскільки рівень значущості не задано за умовою задачі, обираємо його самостійно — 5 %.

У нульовій гіпотезі зазначено, що середня генеральної сукупності більша або дорівнює 28 000 грош. од. Відтак щоразу, коли середня вибіркова буде більшою або дорівнюватиме 28 000 грош. од., це буде просто підтвердженням нульової гіпотези і, отже, ми не зможемо її відхилити. Однак, якщо вибіркова середня буде меншою за 28 000 грош. од., це може бути причиною відхилення нульової гіпотези. Маємо лівосторонню критичну область (рис. 5.8).

Рис. 5.8

У цьому випадку всього 5 % від усіх можливих результатів, які, можливо, приведуть до відмови від нульової гіпотези, знаходяться на крайній лівій частині розподілу. Межу лівосторонньої критичної області знаходимо за формулами (5.17 і 5.18):

$$\Phi(z_{кр}) = -\frac{1-2\alpha}{2} = -\frac{1-0{,}1}{2} = -0{,}45 \text{, звідки } z_{кр} = \Phi^{-1}\left(-\frac{1-2\alpha}{2}\right) = -1{,}645 .$$

Крок 5. Оскільки z^* — спостережуване значення критерію Z — належить області невідхилення H_0, то робимо висновок, що статистичних фактів для підтвердження альтернативної гіпотези H_1 недостатньо.

Приклад 5.3. (приклад побудови правосторонньої критичної області).

Фахівець зі статистики отримав завдання перевірити дані, оголошені в доповіді: «середня заробітна плата робітника металургійної промисловості країни А склала не більше ніж 28 000 грош. од. за попередній рік».

Р о з в ' я з а н н я. Спробуємо дати відповідь на поставлене запитання, використовуючи дані, отримані з дослідження вибірки прикладу 1: $\bar{x}_в$ дорівнює 26 500 грош. од., $\sigma(\bar{x}_в)$ = 1000 грош. од. Стандартне (середнє квадратичне) відхилення генеральної сукупності відоме.

Оскільки $\bar{x}_в = 26\,500$, що менше, ніж 28 000, цей факт відразу підтверджує нульову гіпотезу. Розглянемо випадок, якщо за результатами обробки вибірки отримали $\bar{x}_в = 29\,500$.

Крок 1. Обирають нульову та альтернативну гіпотези:

H_0: середня заробітна плата робітника за рік $\mu \le 28000$ грош. од. ($H_0 : \mu \le \mu_0$);

H_1: середня заробітна плата робітника за рік $\mu > 28\,000$ грош. од. ($H_1 : \mu > \mu_0$).

Крок 2: Визначається статистичний критерій для перевірки нульової гіпотези.

Оскільки середнє квадратичне відхилення генеральної сукупності відоме, статистичний критерій обираємо за формулою:

$$Z = \frac{\overline{X}_в - \mu_0}{\sigma(\overline{X}_в)} .$$

Крок 3. Обчислюється спостережуване значення статистичного критерію як відношення різниці між середньою вибірковою та припущеним значенням середньої генеральної сукупності до стандартного відхилення вибіркової середньої:

$$z^* = \frac{\bar{x}_в - \mu_0}{\sigma(\bar{x}_в)} = \frac{29\,500 - 28\,000}{1000} = 1{,}5 .$$

Крок 4. Обираємо рівень значущості, критичну область (правосторонню, лівосторонню або двосторонню) та знаходимо критичну точку (або критичні точки).

Оскільки рівень значущості не задано за умовою задачі, обираємо його самостійно — 5 %. Критичну точку — межу правосторонньої критичної області — знаходимо за формулами (5.13 і 5.14):

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2} = 0{,}45 \Rightarrow z_{кр} = \Phi^{-1}\left(\frac{1-2\alpha}{2}\right) = 1{,}645 .$$

У нульовій гіпотезі зазначено, що середня генеральної сукупності менша або дорівнює 28 000 грош. од. Отже, у цьому випадку всі значення, які менші або дорівнюють 28 000 грош. од., будуть просто підтверджувати нульову гіпотезу. Однак, якщо середня вибіркова буде набагато більшою, ніж 28 000, то нульова гіпотеза може бути відхилена. Тож маємо правосторонню область відхилення (рис. 5.9).

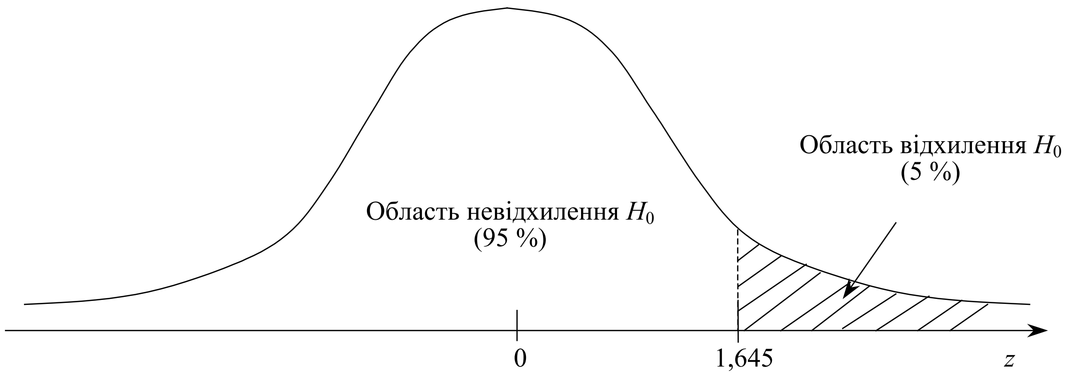


Рис. 5.9

Крок 5. Оскільки спостережуване значення $z^* = 1{,}5$ належить області невідхилення H_0, то робимо висновок, що статистичних фактів для підтвердження альтернативної гіпотези H_1 недостатньо.

Зауваження 6 (щодо визначення нульової та альтернативної гіпотез).

Нульова гіпотеза вважається правильною, поки не доведено, що вона є помилковою. Наведемо три способи визначення нульової та альтернативної гіпотез. Для ілюстрації повернемось до прикладу 5.1, в якому доповідач стверджує, що середня заробітна плата робітника металургійної промисловості країни *A* склала 28 000 грош. од. за попередній рік.

Таблиця 5.2

СПОСОБИ ВИЗНАЧЕННЯ НУЛЬОВОЇ ТА АЛЬТЕРНАТИВНОЇ ГІПОТЕЗ

Примітка: H_0 — нульова гіпотеза; H_1 — альтернативна гіпотеза.	
1) H_0: $\mu = \mu_0$,	H_1: $\mu \neq \mu_0$.
або H_0: $\mu = 28\ 000$,	H_1: $\mu \neq 28\ 000$.
2) H_0: $\mu \geq \mu_0$,	H_1: $\mu < \mu_0$.
або H_0: $\mu \geq 28\ 000$,	H_1: $\mu < 28\ 000$.
3) H_0: $\mu \leq \mu_0$,	H_1: $\mu > \mu_0$.
або H_0: $\mu \leq 28\ 000$,	H_1 $\mu > 28\ 000$

Слід звернути увагу, що умова рівності завжди включається в нульову гіпотезу. Крім того, вибір комбінації має прямий вплив на область відхилення нульової гіпотези.

Вибір між трьома варіантами — це завдання аналітика. У простому випадку якщо твердження зазначає, що середня заробітна плата робітника металургійної промисловості є «точно» 28 000 грош. од. за рік, тоді перший варіант є очевидним вибором.

Однак вибір є трохи складнішим, якщо потрібно перевірити, що середня заробітна плата робітника металургійної промисловості більша або дорівнює 28 000 грош. од. за рік. Перевірку цього твердження можна зробити одним із двох способів.

Можна сформулювати нульову та альтернативну гіпотези, як у випадку 2):

$$H_0\text{: } \mu \geq 28\ 000;\ H_1\text{: } \mu < 28\ 000.$$

Однак, якщо нульова гіпотеза не відкидається, то робимо висновок, що вона може бути правильною. З іншого боку, можна сформулювати нульову та альтернативну гіпотези, як у випадку 3):

$$H_0\text{: } \mu \leq 28\ 000;\ H_1\text{: } \mu > 28\ 000.$$

Тоді, якщо випробування приводить до відмови від нульової гіпотези, ми мали б «переконливі статистичні докази», що середня заробітна плата робітника металургійної промисловості становила більше ніж 28 000 грош. од. за рік. Отже, вибір одного із трьох способів визначення нульової та альтернативної гіпотез (див. табл. 5.2) залежить від того, чи хоче аналітик знати, що твердження нульової гіпотези можливе, або він бажає отримати переконливі статистичні докази того, що сформульоване в нульовій гіпотезі твердження є правильним. Іншими словами, якщо аналітику достатньо інформації про можливість невідхилення деякого припущення, це припущення може бути прийнято за нульову гіпотезу, але якщо аналітик шукає переконливих статистичних доказів справедливості припущеного твердження, це твердження має бути прийнято за альтернативну гіпотезу.

Зауваження 7 (стосовно визначення рівня значущості).

Аналітик самостійно може обирати рівень значущості для випробування. Однак потрібно розуміти, що обираючи рівень значущості 1 %, у результаті перевірки нульової гіпотези рівно 1 % усіх можливих спостережень призведуть до її відхилення. Наприклад, зі 100 обчислених за вибірковими даними спостережуваних значень критерію невідхиленню нульової гіпотези будуть сприяти рівно 99, і лише 1 — відхиленню. Отже, якщо нульова гіпотеза відхиляється навіть на цьому рівні значущості, то статистичні докази за відхилення нульової гіпотези вважаються досить вагомими.

Аналогічно, для 10-відсоткового рівня значущості, 10 зі 100 обчислених спостережуваних значень критерію будуть підтверджувати, що нульова гіпотеза є хибною, тоді як рівно 90 інших спостережуваних значень критерію сприятимуть невідхиленню нульової гіпотези.

Тож якщо нульова гіпотеза не була відхилена з рівнем значущості 1 %, але відхиляється з 10-відсотковим рівнем значущості, це пояснюється лише збільшенням рівня значущості (або збільшенням імовірності відмови від нульової гіпотези від 1 % до 10 %). Іншими словами, зі збільшенням рівня значущості відхиляти нульову гіпотезу стає легше. Тому, якщо нульова гіпотеза відхиляється з меншими рівнями значущості, докази помилковості нульової гіпотези вважаються більш вагомими.

Зауваження 8 (стосовно використання іншого методу розв'язання задач перевірки статистичних гіпотез — методу побудови довірчого інтервалу).

1. Для перевірки статистичних гіпотез можна застосувати інший підхід, а саме, побудову довірчих інтервалів. Висновок про відхилення гіпотези H_0 можна зробити, якщо оцінювальний параметр ознаки генеральної сукупності не потрапляє до довірчого інтервалу з надійністю $\gamma = 1 - \alpha$, а потрапляє до критичної області (див. § 4).
2. Використовуючи метод побудови довірчого інтервалу, крім перевірки гіпотези, ми отримуємо додаткову інформацію про можливі справжні значення параметру генеральної сукупності. Але цей метод можна застосовувати тільки для альтернативних гіпотез, що визначають двосторонню критичну область (див. § 4).
3. Побудова довірчого інтервалу та двосторонньої критичної області приводить до однакового результату, але висновки за результатами використання цих двох методів відрізняються: двостороння критична область визначає межі (критичні точки), між якими знаходяться $\gamma = (1 - \alpha)$ кількості всіх спостережуваних значень критерію, знайдених під час повторних досліджень, що сприяють невідхиленню нульової гіпотези; а довірчий інтервал визначає межі (критичні точки), між якими з імовірністю $\gamma = (1 - \alpha)$ знаходиться дійсне значення оцінювального параметра (див. § 4).

Зауваження 9 (стосовно статистичного висновку проти економічного).

Після завершення перевірки нульової гіпотези аналітик має впевнитися, що результати перевірки мали економічний сенс. Незалежно від відхилення або невідхилення нульової гіпотези за результатами її статистичної перевірки, якщо отримані висновки не мають економічного сенсу, то аналітику може знадобитися проведення додаткових випробувань, можливо, уже із іншою нульовою гіпотезою.

***P*-значення (*P-value*)**

Пастка при перевірці гіпотези полягає в тому, що вона відповідає лише на запитання: чи буде нульова гіпотеза відхилена на даному рівні значущості.

Наприклад, перевірка нульової гіпотези на основі даних конкретної вибірки може призвести до її відхилення з рівнем значущості 10 %, але з 5-відсотковим рівнем значущості вона вже відхилятися не буде. Ми дійсно не знаємо, в якій точці між 10 і 5-відсотковим рівнями значущості нульова гіпотеза перетворилася з відхиленої на не відхилену. Ця точка називається *P*-значенням (або *p-value*).

Означення 17. ***P-значення*** є найменшим значенням рівня значущості, за якого нульова гіпотеза може бути відхилена для даної вибірки.

Наприклад, припустимо, що спостережуване значення u^* (інакше його ще називають статистикою вибірки) знаходиться на відстанні 1,96 стандартної помилки від припущеного значення параметра генеральної сукупності за односторонньої перевірки гіпотези (нагадаємо, що відстань в 1,96 стандартної помилки від припущеного значення параметра генеральної сукупності до спостережуваного значення критерію за одностороньої перевірки гіпотези залишає критичну область у 2,5 %). Тоді саме 2,5 % і буде *p*-значенням. Якщо рівень значущості обрати меншим ніж 2,5 %, то зробити висновок про відхилення нульової гіпотези не буде можливості, оскільки спостережуване значення критерію належатиме області невідхилення нульової гіпотези. Інакше, для рівня значущості більшого за 2,5 %, нульова гіпотеза буде відхилена (тобто спостережуване значення критерію належатиме критичній області). Як бачимо, чим далі u^* — спостережуване значення критерію даної вибірки лежить від припущеного в нульовій гіпотезі значення параметра ознаки генеральної сукупності, тим меншим буде p-значення, і тому зменшується ймовірність отримання висновку про відхилення нульової гіпотези. Але, чим менше p-значення, тим вагомішими вважаються аргументи за відхилення нульової гіпотези на користь альтернативної.

Розглянемо два способи перевірки гіпотези про середнє значення ознаки генеральної сукупності у разі відомого стандартного відхилення ознаки генеральної сукупності:

1) з використанням методу побудови області відхилення;

2) з використанням p-значення (метод p-value approach).

Для розв'язання конкретних виробничих проблем значну допомогу менеджеру може надати якісне проведення перевірки статистичних гіпотез, коли нульова гіпотеза являє собою *status quo* (існуючий стан речей). Часто буває, що висновок, отриманий у результаті перевірки гіпотези, набуває форми деякого напрямку дій. Наприклад, якщо існують докази зростання або скорочення величини певного параметра, може бути прийнято рішення про новий план управління запасами, або про випуск нового виду продукції, перехід на кращі ліки для лікування хвороби тощо.

Щоб проілюструвати цей процес, розглянемо такий приклад:

Приклад 5.4. Нова система розрахунків у супермаркеті.

Менеджер супермаркету обдумує впровадження нової системи розрахунків для клієнтів, які використовують на касі кредитні електронні картки для розрахунків. Після всебічного фінансового аналізу він визначив, що нова система буде рентабельною, тільки якщо середнє значення всіх одноразових електронних платежів за місяць перевищуватиме 170 грн. Випадкова вибірка 400 одноразових електронних платежів за місяць засвідчила, що середній платіж у вибірці становив 178 грн. Менеджер знає, що електронні платежі є приблизно нормально розподіленими зі стандартним відхиленням у 65 грн. Чи можна, виходячи з цих даних, зробити висновок, що впровадження нової системи розрахунків буде рентабельним?

Розв'язання. Цей приклад розглядає сукупність одноразових електронних платежів у супермаркеті. Для перевірки, чи буде рентабельним упровадження нової системи, ми ставимо таке запитання: середнє вибіркове значення 178 є достатньо великим — порівняно зі 170?

Розглянемо два способи отримати відповідь на це запитання.

Перший спосіб — це метод побудови області відхилення або метод перевірки статистичної гіпотези. Цей метод використовується у статистичних розрахунках, які здійснюються вручну.

Другий спосіб — метод дослідження за допомогою p-значення (інакше його називають *p-value approach*). Цей метод використовується здебільшого для розрахунків з використанням комп'ютера та статистичного програмного забезпечення.

Розглянемо кожний спосіб окремо.

Перший спосіб.

Крок перший. Визначаємо нульову та альтернативну гіпотези.

Для того щоб зробити висновок, що система буде рентабельною, менеджеру потрібно довести, що середнє значення всіх одноразових електронних платежів за місяць перевищуватиме 170 грн. Якщо середнє значення всіх одноразових електронних платежів за місяць буде меншим, або дорівнюватиме 170 грн, система не буде рентабельною. Отже, нульова та альтернативна гіпотези можуть бути виражені у такий спосіб:

$H_0: \mu = 170$ — нова система розрахунків не установлюється, оскільки вона нерентабельна;

$H_1: \mu > 170$ — нова система розрахунків встановлюється, оскільки вона є рентабельною.

Здається, раціонально відхилити нульову гіпотезу на користь альтернативної, якщо середнє значення по вибірці значно більше, ніж 170. Якщо, наприклад, обчислене середнє значення по вибірці дорівнювало би 500, то висновок про помилковість нульової гіпотези був би абсолютно очевидний. З іншого боку, значення $\bar{x}_в$ близьке до 170, наприклад 171, не дозволяє нам відхилити нульову гіпотезу, оскільки цілком можливо отримати середнє значення по вибірці 171, якщо середнє по сукупності, з якої була здійснена вибірка, становить 170. На жаль, висновок про значення середньої генеральної сукупності не завжди є настільки очевидним. У прикладі, який ми розглядаємо, середнє значення по вибірці було обчислене і становить 178. Така оцінка, очевидно, не досить віддалена, і не досить близька до 170. Для висновку про значення середньої генеральної сукупності встановлюємо область відхилення — діапазон значень критичної області. Якщо статистичні дані, що перевіряються (спостережуване значення u^*), потрапляють до цього діапазону, ми робимо висновок про можливість відхилення нульової гіпотези на користь альтернативної.

Крок другий. Визначення статистичного критерію (визначення правильного розподілу статистики вибірки).

Оскільки середнє квадратичне відхилення генеральної сукупності відоме, статистичний критерій обираємо за формулою (5.9):

$$Z = \frac{\bar{X}_в - \mu_0}{\sigma(\bar{X}_в)}, \text{ де } \sigma(\bar{X}_в) = \frac{\sigma}{\sqrt{n}}.$$

Крок третій. Обчислюється спостережуване значення статистичного критерію як відношення різниці між середнім вибірки та припущеним у нульовій гіпотезі значенням середнього генеральної сукупності до стандартного відхилення вибіркової середньої:

$$z^* = \frac{\overline{x}_в - \mu_0}{\sigma(\overline{X}_в)} = \frac{178 - 170}{65 \big/ \sqrt{400}} = 2{,}46 .$$

Крок четвертий. Обираємо рівень значущості, критичну область (правостороння, лівостороння або двостороння), критичну точку.

Оскільки рівень значущості не задано за умовою задачі, обираємо його самостійно — 5 %.

Цей рівень значущості дає нам правосторонню область відхилення нульової гіпотези $Z > z_{кр}$. За формулами (5.13 і 5.14) знаходимо критичну точку: $\Phi(z_{кр}) = \frac{1-2\alpha}{2} = 0{,}45$,

звідки $z_{кр} = \Phi^{-1}\left(\frac{1-2\alpha}{2}\right) = 1{,}645$.

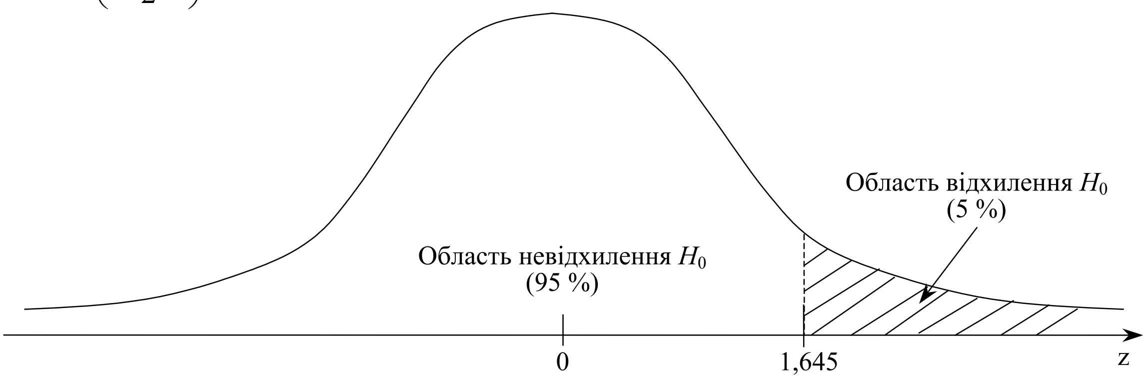


Рис. 5.10

Крок п'ятий. Оскільки спостережуване значення статистичного критерію $z^* = 2{,}46$ є більшим, ніж критичне значення 1,645, та потрапляє до області відхилення нульової гіпотези (рис. 5.10), робимо висновок, що існує достатньо статистичних доказів на користь альтернативної гіпотези про те, що середнє значення всіх одноразових електронних платежів за місяць перевищує 170 грн.

Підкреслимо, що перевірка була проведена для рівня значущості 5 %. Відхилення нульової гіпотези дає підставу для висновку про доцільність установлення нової системи розрахунків.

Другий спосіб.

Статистична процедура — це тільки один із кількох факторів, який розглядається менеджером для прийняття рішення. Використовуючи перший спосіб розв'язання задачі, менеджер з'ясував, що існує достатньо статистичних доказів для висновку, що середнє значення всіх одноразових електронних платежів за місяць перевищує 170 грн. Проте, перед початком здійснення будь-яких дій, потрібно дослідити ще багато інших факторів, таких як вартість, фізична можливість перебудови системи розрахунків, можливість помилки (у нашому випадку — помилки першого типу).

Для отримання повної користі від інформації, яка стала нам доступною у результаті перевірки статистичної гіпотези та прийняття більш обґрунтованого рішення, корисно визначити міру кількості статистичних фактів, які підтримують альтернативну гіпотезу.

Таку міру нам надає перевірка *p*-значення (*p-value*).

Можливість оцінити найменшу ймовірність припущення помилки першого типу надає нам *p*-значення.

Ми відхилили нульову гіпотезу на 5-відсотковому рівні значущості. Знайдемо *p*-значення — найменшу ймовірність припущення помилки першого типу, або найменший рівень значущості, на якому нульова гіпотеза може бути помилково відхилена на користь альтернативної. Оскільки йдеться про наявність помилки першого типу, вважаємо, що нульова гіпотеза $H_0: \ \mu = 170$ є справедлива. Тоді для нашого прикладу *p*-значення є ймовірністю спостереження середнього по вибірці щонайменше у розмірі 178, коли середнє по сукупності становить 170. Так,

$$p - \text{значення} = P(\overline{X}_в > 178) = P\left(\frac{\overline{X}_в - \mu_0}{\sigma/\sqrt{n}} > \frac{178-170}{65/\sqrt{400}}\right) = P(Z > 2{,}46) =$$

$$= 1 - P(Z \le 2{,}46) = 1 - 0{,}9931 = 0{,}0069.$$

Отже, імовірність спостерігати середнє значення щонайменше так само великим, як 178 у сукупності, де середнє дорівнює 170, становить 0,0069, що є дуже малим значенням. Інакше кажучи, ми спостерігали малоймовірну подію — малоймовірну настільки, що маємо серйозні сумніви, аби зробити припущення про те, що нульова гіпотеза правильна. Тобто, ми отримали вагому підставу відхилити нульову гіпотезу і підтримати альтернативну.

Інтерпретація *p*-значення

p-значення надає корисну інформацію, оскільки отримане значення є мірою статистичного доказу, що підтримує альтернативну гіпотезу. Для розуміння цього пояснення повною мірою, звернемося до табл. 5.3, де вказано кілька можливих значень $\overline{X}_в$, відповідних їм спостережуваних даних z^* та *p*-значення для прикладу 4.

Таблиця 5.3

Середнє вибіркове $\overline{x}_в$	Спостережуване значення z^* критерію Z: $z^* = \frac{\overline{X}_в - \mu}{\sigma/\sqrt{n}} = \frac{\overline{x}_в - 170}{65/\sqrt{400}}$.	*p*-значення
170	0	0,5000
172	0,62	0,2676
174	1,23	0,1093
176	1,85	0,0322
178	2,46	0,0069
180	3,08	0,0010

Зауважимо, що чим ближче $\overline{x}_в$ до гіпотетичного значення 170, тим більшим є її *p*-значення.

Що більша різниця між $\overline{x}_в$ та 170, тим меншим є її *p*-значення.

Чим далі знаходиться $\overline{x}_в$ від значення 170, тим легше зробити висновок про відхилення нульової гіпотези на користь альтернативної.

Отже, чим менше *p*-значення, тим більше статистичних фактів на користь підтримки альтернативної гіпотези.

На рис 5.11 графічно зображено інформацію табл. 5.3.

Наведено *p*-значення для прикладу 4:

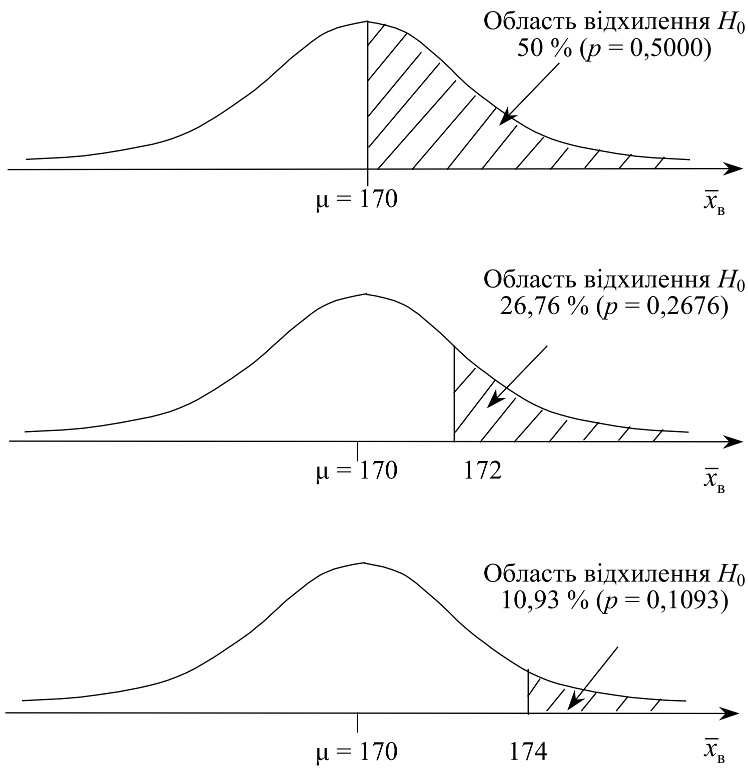
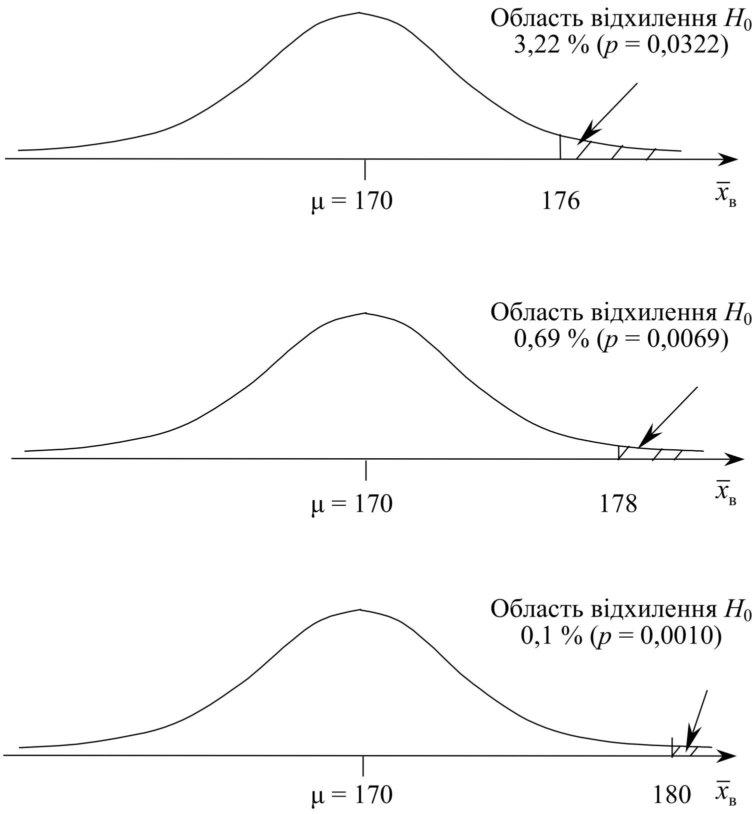


Рис. 5.11



Закінчення рис. 5.11

У зв'язку з цим виникає питання, наскільки малим має бути p-значення для висновку про те, що альтернативна гіпотеза правильна? У загальному випадку відповідь залежить від кількості факторів, включаючи вартість помилок першого та другого типу. У прикладі 4 помилка першого типу трапиться, якщо менеджер прийме рішення на користь нової системи розрахунків, у випадку коли вона нерентабельна. Якщо вартість цієї помилки висока, ми намагаємося зменшити її ймовірність. У випадку з методом області відхилення досягатимемо цього встановленням рівня значущості досить низьким, наприклад 1 %. Використовуючи метод дослідження p-значення, ми повинні наполягати на тому, щоби p-значення було досить малим, надаючи достатній доказ для висновку, що середнє значення всіх електронних рахунків за місяць буде перевищувати 170 грн для встановлення нової системи розрахунків.

Використання отриманих *p*-значень для висновків стосовно альтернативної гіпотези.

Якщо p-значення менше за 0,01, стверджується, що існує неспростовний доказ для прийняття альтернативної гіпотези як істинної. Вважається, що перевірка має високий рівень статистичної значущості.

Якщо p-значення знаходиться у межах від 0,01 до 0,05, існує твердий доказ того, що альтернативна гіпотеза істинна. Результат вважається статистично значущим.

Якщо p-значення знаходиться у межах від 0,05 до 0,10 — докази істинності альтернативної гіпотези вважаються непереконливими.

Якщо p-значення більше 5 %, результат вважається статистично незначущим.

Якщо p-значення перевищує 0,10, робиться висновок, що не існує доказів для визнання альтернативної гіпотези істинною.

Деякі типові помилки, пов'язані з інтерпретацією *p*-значення:

1) p-значення не дорівнює ймовірності істинності нульової гіпотези;
2) p-значення не дорівнює ймовірності істинності альтернативної гіпотези;
3) p-значення не дорівнює ймовірності помилки першого роду;
4) p-значення не дорівнює ймовірності помилки другого роду;
5) p-значення не є ймовірність того, що повторне випробування не приведе до такого самого рішення.

Отже, ***досяжний рівень значущості p-значення — це найменша величина рівня значущості, за якого нульова гіпотеза відхиляється для даного спостережуваного значення*** u^* ***обраного критерію U.***

5.3. ПЕРЕВІРКА ДОСТОВІРНОСТІ ГІПОТЕЗИ ПРО ЗНАЧЕННЯ СЕРЕДНЬОЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ, ЯКЩО ДИСПЕРСІЯ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ НЕВІДОМА

Якщо дисперсія генеральної сукупності σ^2 невідома (наприклад у випадку малих вибірок з обсягом не більше 30 елементів), то за статистичний критерій обирають випадкову величину, яка має розподіл Стьюдента з $\nu = n - 1$ ступенями вільності:

$$T = \frac{\overline{X_в} - \mu_0}{s}\sqrt{n}\ , \tag{5.24}$$

$$s = \sqrt{\frac{n}{n-1} \cdot D_в}\ , \tag{5.25}$$

де s — виправлене середнє квадратичне відхилення вибірки. Спостережуване значення t^* критерію T знаходять за формулою:

$$t^* = \frac{\bar{x}_в - \mu_0}{s} \cdot \sqrt{n}\ . \tag{5.26}$$

Критична область будується залежно від вигляду альтернативної гіпотези H_1, аналогічно розглянутим раніше випадкам, коли відома дисперсія генеральної сукупності. Критичні точки $t_{кр}$ визначаються із таблиць розподілу Стьюдента, таблиця додатку 8, за відомими параметрами α і $\nu = n - 1$:

$$t_{кр} = t(\alpha, \nu)\ .$$

Зауваження 10 (до можливості застосування статистичного критерію Т).

Для перевірки достовірності гіпотези про значення середньої генеральної сукупності за невідомої дисперсії обирається критерій Т, якщо:

1) обсяг вибірки великий: $n > 30$ (закон розподілу генеральної сукупності Х може бути довільним),

2) обсяг вибірки $n \le 30$ (закон розподілу генеральної сукупності Х обов'язково має бути нормальним).

Зауваження 11. За великих обсягів вибірки $(n > 30)$ статистичний критерій (5.24) наближається асимптотично до нормованого нормального закону розподілу. У цьому разі критичні точки визначаються за рівностями (5.11—5.23).

Приклад 5.5. Менеджер відділу кадрів страхової компанії отримав завдання перевірити інформацію: середня кількість часу, що витрачається страховими агентами на роботу з електронною поштою щодня, становить 75 хвилин. Він випадковим чином відібрав 25 страхових агентів і знайшов для них середню кількість часу, що витрачається на роботу з електронною поштою щодня, яка склала 83,4 хвилини із вибірковим стандартним відхиленням, що дорівнює 23,65 хвилин. Визначити на 5-відсотковому рівні значущості, чи може менеджер стверджувати,що отримана інформація хибна? Вважаємо, що кількість часу, який витрачається страховими агентами на роботу з електронною поштою, розподілено за нормальним законом.

Р о з в ' я з а н н я.

Крок перший. Виходячи з умови задачі, сформулюємо гіпотези:

$$H_0 : \mu = 75\ ;$$
$$H_1 : \mu \ne 75\ .$$

Крок другий. За умовою задачі, генеральна сукупність розподілена за нормальним законом із невідомим середнім квадратичним відхиленням σ, тож для перевірки основної статистичної гіпотези H_0 за статистичний критерію обираємо випадкову величину $T = \frac{\overline{X_в} - \mu_0}{s}\sqrt{n}$, що має розподіл Стьюдента з $\nu = n - 1$ ступенями вільності, де $s = \sqrt{\frac{n}{n-1} D_в}$ — виправлене середнє квадратичне відхилення вибірки.

Крок третій. Обчислюємо спостережуване значення критерію:

$$t^* = \frac{83{,}4 - 75}{23{,}65} \cdot \sqrt{25} \approx 1{,}78\ .$$

Крок четвертий.

Для рівня значущості α= 0,05 та двосторонньої критичної області знайдемо відповідні критичні точки $t_{\text{кр}}^{\text{пр}}$ та $t_{\text{кр}}^{\text{лів}}$ (рис. 5.12).

За таблицею додатка 8: $t(\alpha/2; \nu = n-1) = t\,(0{,}025;\ 24\,) = 2{,}064$.

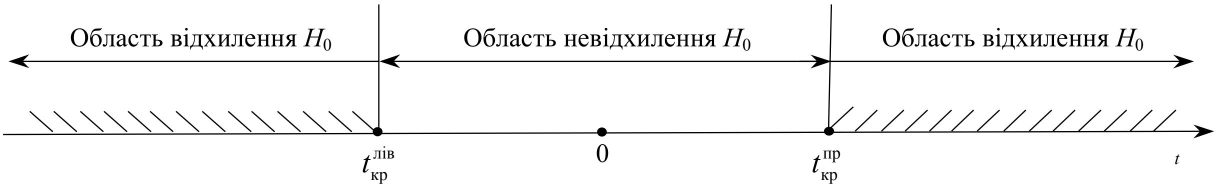


Рис. 5.12

$$t_{\text{кр}}^{\text{лів}} = -t_{\text{кр}} = -2{,}064 \quad t_{\text{кр}} = t_{\text{кр}}^{\text{пр}} = 2{,}064\,,$$

Спостережуване значення $t^* \approx 1{,}78$ не належить області відхилення H_0.

Крок п'ятий.

Оскільки спостережуване значення критерію не потрапляє до критичної області, робимо висновок про невідхилення нульової гіпотези на користь альтернативної. Отже, не має достатніх статистичних фактів для підтвердження, що інформація, отримана менеджером, хибна.

Правило відхилення нульової гіпотези H_0 : μ ≠ μ_0 на користь альтернативної гіпотези H_1 : μ ≠ μ_0 (двостороння критична область). Для перевірки із заданим рівнем значущості α нульової гіпотези H_0: $\mu = \mu_0$ про рівність невідомої середньої μ нормально розподіленої генеральної сукупності з невідомою дисперсією гіпотетичному значенню μ_0 за альтернативної гіпотези H_1 $\mu \neq \mu_0$, необхідно обчислити спостережуване значення t^* критерію T за формулою (5.26) та за таблицею критичних точок розподілу Стьюдента для заданого рівня значущості α і кількості ступенів вільності $\nu = n - 1$ знайти точку $t\left(\alpha/2; \nu = n-1\right)$.

$t_{\text{кр}}^{\text{пр}} = t\left(\alpha/2; \nu = n-1\right)$.

$t_{\text{кр}}^{\text{лів}} = -\,t\left(\alpha/2; \nu = n-1\right)$.

Якщо спостережуване значення критерію $|t^*| < t\left(\alpha/2; \nu = n-1\right)$, то недостатньо статистичних фактів для відхилення нульової гіпотези.

Якщо спостережуване значення критерію $|t^*| > t\left(\alpha/2; \nu = n-1\right)$, то нульову гіпотезу відхиляють.

Приклад 5.6. Інвестор не бажає витрачати свої заощадження на купівлю акцій нафтової компанії, якщо не існує можливості збільшення їх ціни у середньому принаймні на 0,9 відсотка за місяць. Спеціалістом зі статистики проведено 10 незалежних спостережень випадкової величини X — зміни курсу акцій нафтової компанії протягом місяця у відсотках, що має нормальний закон розподілу з невідомими значеннями μ, σ.

Які рекомендації отримає інвестор від спеціаліста зі статистики?

Наслідки спостережень подано у вигляді таблиці:

x_i	2,5	2	–2,3	1,9	–2,1	2,4	2,3	–2,5	1,5	–1,7
n_i	1	1	1	1	1	1	1	1	1	1

Р о з в ' я з а н н я.

Перевіримо правильність нульової гіпотези $H_0 : \mu = 0{,}9$ за альтернативної гіпотези $H_1 : \mu < 0{,}9$ з рівнем значущості α = 0,001.

Запишемо табличні дані у вигляді статистичного ряду розподілу:

x_i	–2,5	–2,3	–2,1	–1,7	1,5	1,9	2	2,3	2,4	2,5
n_i	1	1	1	1	1	1	1	1	1	1

Спочатку обчислимо $\bar{x}_в$, та s:

$$\bar{x}_{\text{в}} = \frac{\sum x_i}{n} = \frac{-2{,}5-2{,}3-2{,}1-1{,}7+1{,}5+1{,}9+2+2{,}3+2{,}4+2{,}5}{10} = 0{,}4\,;$$

$$D_{\text{в}} = \frac{\sum x_i^2}{n} - (x_{\text{в}})^2 =$$

$$\frac{6{,}25 + 5{,}29 + 4{,}41 + 2{,}89 + 2{,}25 + 3{,}61 + 4 + 5{,}29 + 5{,}76 + 6{,}25}{10} - (0{,}4)^2 = 4{,}6 - 0{,}16 = 4{,}44\,;$$

$$s^2 = \frac{n}{n-1} D_{\text{в}} = \frac{10}{9} \cdot 4{,}44 = 4{,}933\,,$$

звідки $s = \sqrt{4{,}933} \approx 2{,}22$.

За альтернативної гіпотези $H_1 : \mu < 0{,}9$ будується лівостороння критична область (рис. 5.13). За таблицею критичних значень розподілу Стьюдента, таблицею додатка 8, знаходимо допоміжну точку:

$$t_{\text{кр}}^{\text{пр}} = t(\alpha = 0{,}001,\ \nu = 9) = 4{,}297\,.$$

Оскільки щільність імовірностей для розподілу Стьюдента парна функція, то $t_{\text{кр}}^{\text{лів}} = -t_{\text{кр}}^{\text{пр}} = 4{,}297$.

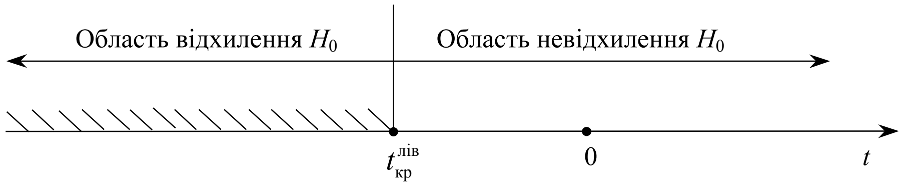


Рис. 5.13

Обчислимо спостережуване значення критерію:

$$t^* = \frac{\overline{x}_{в} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0{,}4 - 0{,}9}{\frac{2{,}22}{\sqrt{10}}} \approx \frac{-0{,}5}{\frac{2{,}22}{3{,}16}} \approx \frac{-0{,}5}{0{,}702} \approx -0{,}712\,.$$

Висновок. Оскільки спостережуване значення критерію не належить критичній області, то немає підстав відхилити нульову гіпотезу $H_0 : \mu = 0{,}9$ на користь альтернативної H_1: $\mu < 0{,}9$. Статистичних фактів недостатньо для прийняття альтернативної гіпотези: ціна акцій нафтової компанії протягом місяця зростає у середньому менше ніж на 0,9 %.

На підставі отриманих результатів перевірки гіпотези спеціаліст зі статистики не отримав достатніх статистичних фактів, щоб не рекомендувати інвестору купівлю акцій нафтової компанії.

Правило відхилення нульової гіпотези $H_0 : \mu = \mu_0$ на користь альтернативної гіпотези H_1: $\mu < \mu_0$ (лівостороння критична область).

Для перевірки із заданим рівнем значущості α нульової гіпотези $H_0 : \mu = \mu_0$ про рівність невідомої середньої μ нормально розподіленої генеральної сукупності з невідомою дисперсією гіпотетичному значенню μ_0 за альтернативної гіпотези $H_1 : \mu < \mu_0$, необхідно обчислити спостережуване значення критерію t^* за формулою (5.26) та за таблицею критичних точок розподілу Стьюдента для заданого рівня значущості α і кількості ступенів вільності $\nu = n - 1$ знайти допоміжну точку $t_{\text{кр}}^{\text{пр}} = t(\alpha, \nu)$ а після цього записати межу лівосторонньої критичної області $t_{\text{кр}}^{\text{лів}} = -t_{\text{кр}}^{\text{пр}}$. Якщо спостережуване значення критерію $t^* < t_{\text{кр}}^{\text{лів}}$ — нульову гіпотезу відхиляють.

Якщо спостережуване значення критерію $t^* > t_{\text{кр}}^{\text{лів}}$ — робимо висновок про недостатність статистичних фактів для відхилення нульової гіпотези.

Приклад 5.7. Банк має намір відкрити філіал у новому районі. Вважається, що новий філіал працюватиме прибутково, якщо середній тижневий дохід мешканців району буде перевищувати $\mu = 500$ грн. Вибіркове опитування 25 мешканців виявило середній дохід $\overline{x}_{\text{в}} = 508$ грн за тиждень, а виправлене середнє квадратичне відхилення доходу $s = 30$ грн. За рівня значущості $\alpha = 0{,}001$ з'ясувати, чи буде філіал працювати прибутково. Вважаємо, що середній тижневий дохід мешканців нового району розподілений за нормальним законом.

Розв'язання.

Крок перший. Визначаємо основну та альтернативну гіпотези:

$H_0 : \mu = 500,$

$H_1 \ \ \mu > 500.$

Крок другий. Значення генеральної дисперсії σ^2 невідоме, закон розподілу генеральної сукупності X — середнього тижневого доходу мешканців нового району — є нормальним за умовою задачі. Для перевірки нульової гіпотези обираємо статистичний критерій:

$$T = \frac{\overline{X}_в - \mu_0}{s}\sqrt{n}\ .$$

Крок третій. Обчислюємо спостережуване значення критерію:

$$t^* = \frac{\overline{x}_в - \mu_0}{s}\sqrt{n} = \frac{508-500}{30}\sqrt{25} \approx 1{,}33\ .$$

Крок четвертий. За заданим рівнем значущості α та числом ступенів вільності $\nu = n - 1 = 24$ для правосторонньої критичної області знаходимо критичну точку $t_{кр}^{пр} = t(\alpha, \nu)$ за таблицею критичних точок розподілу Стьюдента:

$$t_{кр}^{пр} = t(0{,}001;\ 24) = 3{,}467\ .$$

Спостережуване значення критерію не потрапило до області відхилення нульової гіпотези .

Крок п'ятий. Робимо висновок про недостатність статистичних фактів для висновку про прибутковість роботи філіалу банку в новому районі.

Правило відхилення нульової гіпотези $H_0 : \mu = \mu_0$ на користь альтернативної гіпотези $H_1 : \mu > \mu_0$ (правостороння критична область).

Для перевірки з заданим рівнем значущості α нульової гіпотези $H_0 : \mu = \mu_0$ про рівність невідомої середньої μ нормально розподіленої генеральної сукупності з невідомою дисперсією гіпотетичному значенню μ_0 за альтернативної гіпотези $H_1 : \mu > \mu_0$, необхідно обчислити спостережуване значення t^* критерію T за формулою (5.13) та за таблицею критичних точок розподілу Стьюдента для заданого рівня значущості α і кількості ступенів вільності $\nu = n - 1$ знайти критичну точку $t_{кр}^{пр} = t(\alpha, \nu)$.

Якщо спостережуване значення критерію $t^* > t_{кр}^{пр}$ — нульову гіпотезу відхиляють.

Якщо спостережуване значення критерію $t^* < t_{кр}^{пр}$ — робимо висновок про недостатність статистичних фактів для відхилення нульової гіпотези.

Приклад 5.8. Реалізувавши вибірку з генеральної сукупності, елементами якої є однотипні заготовки, довжина яких X є випадковою величиною з нормальним законом розподілу, дістали статистичний розподіл:

x_i	6,5	8,5	10,5	12,5	14,5	16,5
n_i	10	20	30	20	10	10

З рівнем значущості α = 0,001 перевірити правильність гіпотези $H_0 : \mu = 15{,}5$, *за альтернативної гіпотези* $H_1 : \mu < 15{,}5$.

Р о з в ' я з а н н я.

Крок перший. $H_0 : \mu = 15{,}5$; $H_1 : \mu < 15{,}5$.

Крок другий. Оскільки середнє квадратичне відхилення генеральної сукупності невідоме, за статистичний критерій обираємо випадкову величину T, розподілену за законом Стьюдента:

$$T = \frac{\overline{x}_в - \mu_0}{s}\sqrt{n}\ .$$

Крок третій. Для обчислення спостережуваного значення $t^* = \frac{\overline{x}_в - \mu_0}{s}\sqrt{n}$ критерію T обчислимо значення $\overline{x}_в$ та s. Оскільки $n = \sum n_i = 100$, маємо:

$$\overline{x}_в = \frac{\sum x_i n_i}{n} = \frac{6{,}5\cdot 10 + 8{,}5 + 20 + 10{,}5\cdot 30 + 12{,}5\cdot 20 + 14{,}5\cdot 10 + 16{,}5\cdot 10}{100} =$$
$$= \frac{65 + 170 + 315 + 250 + 145 + 165}{100} = \frac{1110}{100} = 11{,}1.$$

$$\frac{x_i^2 n_i}{n} = \frac{42,25 \cdot 10 + 72,25 \cdot 20 + 110,25 \cdot 30 + 156,25 \cdot 20 + 210,25 \cdot 10 + 272,25 \cdot 10}{100} =$$

$$= \frac{422,5 + 144,5 + 3307,5 + 3125 + 2102,5 + 2722,5}{100} = \frac{13125}{100} = 131,25;$$

$$D_{\text{в}} = \frac{\sum x_i^2 n_i}{n} - (\bar{x}_{\text{в}})^2 = 131,25 - (11,1)^2 = 131,25 - 123,21 = 8,04;$$

$$s^2 = \frac{n}{n-1} D_{\text{в}} = \frac{100}{99} \cdot 8,04 \approx 8,12;$$

$$s = \sqrt{8,12 \approx 2,85}\ .$$

Отже , спостережуване значення критерію дорівнює:

$$z^* = \frac{\bar{x}_{\text{в}} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{11,1 - 15,5}{\frac{2,28}{\sqrt{100}}} = -\frac{4,4}{0,285} \approx -15,44.$$

Крок четвертий. Оскільки обсяг вибірки великий (n = 100 > 30), розподіл випадкової величини $T = \frac{\overline{X}_{\text{в}} - \mu_0}{s}\sqrt{n}$ наближатиметься до нормального закону розподілу $N(0;\ 1)$. Тому для визначення критичної точки $t_{\text{кр}} = z_{\text{кр}}$ застосовуємо рівність:

$$\Phi\left(z_{\text{кр}}^{\text{пр}}\right) = \frac{1-2\alpha}{2},$$

$$\Phi\left(z_{\text{кр}}^{\text{пр}}\right) = \frac{1 - 2 \cdot 0,001}{2} = \frac{0,998}{2} = 0,499,$$

звідки $z_{\text{кр}}^{\text{пр}} = 3,2$; $z_{\text{кр}}^{\text{лів}} = -z_{\text{кр}}^{\text{пр}} = -3,2$.

Лівостороння критична область має такий вигляд: $\{z | z < -3,2\}$. Вона зображена на рис. 5.14.

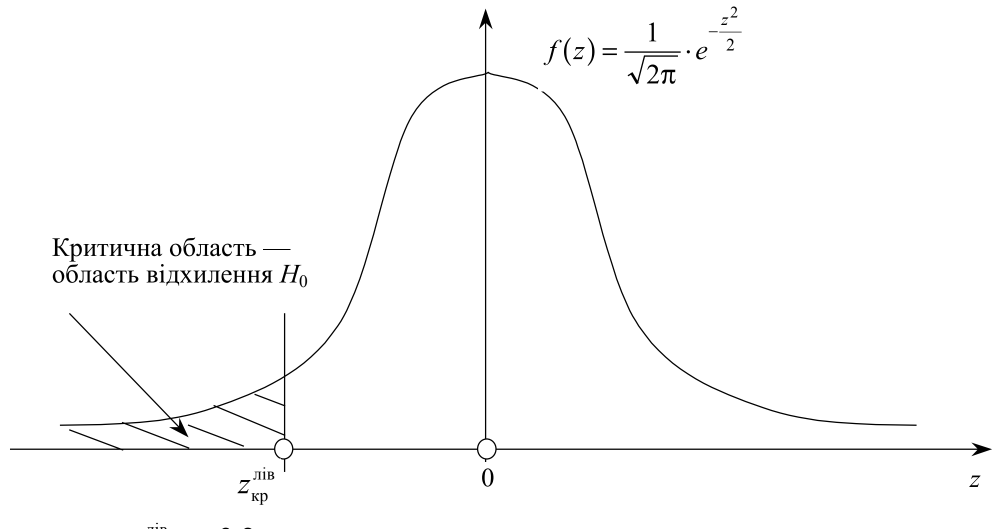


Рис. 5.14

Крок п'ятий. Оскільки спостережуване значення z* критерію належить області відхилення нульової гіпотези $H_0 : \mu = 15,5$, робимо висновок про достатність статистичних фактів для прийняття альтернативної гіпотези $H_1 : \mu < 15,5$.

Приклад 5.9. У зв'язку з величезним обсягом аудиторії глядачів фінальних олімпійських змагань рекламні фірми створюють спеціальні реклами, які містять елементи розваг. Вартість тридцяти секунд такої реклами склала під час останньої зимової олімпіади 2,3 мільйона доларів США. Випадково відібраних 116 осіб, які стежили за фінальними змаганнями, опитали стосовно кількості переглянутих реклам. Чи дозволяють нам отримані дані зробити висновок, що середня кількість переглянутих реклам більша ніж 15, якщо $\bar{x}_{\text{в}}$ = 15,27 , s = 5,72?

Р о з в ' я з а н н я.

Крок перший. $H_0 : \mu = 15$, $H_1 : \mu > 15$.

Крок другий. Оскільки середнє квадратичне відхилення генеральної сукупності невідоме, за статистичний критерій обираємо випадкову величину T, розподілену за законом Стьюдента:

$$T = \frac{\overline{X}_{\text{в}} - \mu_0}{s}\sqrt{n}\,.$$

Крок третій. Обчислюємо спостережуване значення $t^* = \frac{\overline{x}_{\text{в}} - \mu_0}{s}\sqrt{n}$ критерію T : $t^* = \frac{15{,}27 - 15}{5{,}72}\sqrt{116} = 0{,}51$.

Крок четвертий. Оскільки рівень значущості не задано, обираємо його самостійно — 5 %. Обсяг вибірки великий (n = 100 > 30), тоді розподіл статистичного критерію $T = \frac{\overline{X}_{\text{в}} - \mu_0}{s}\sqrt{n}$ наближатиметься до нормального закону розподілу $N(0;\,1)$. Тому для визначення критичної точки застосовуємо рівність:

$$t_{\text{кр}} = z_{\text{кр}}.$$

Крок п'ятий. Оскільки спостережуване значення z^* критерію не належить області відхилення нульової гіпотези $H_0 : \mu = 15$, робимо висновок про недостатність статистичних фактів для прийняття альтернативної гіпотези $H_1 : \mu > 15$.

Зауваження 12. Висновок про недостатність статистичних фактів щодо прийняття альтернативної гіпотези $H_1 : \mu > 15$ можна було зробити після обчислення p-значення:

$$P(t > 0{,}51) \approx P\,(z > 0{,}51) = 0{,}5 - P\,(0 \le z \le 0{,}51) = 0{,}5 - 0{,}195 = 0{,}305.$$

Оскільки p-значення перевищує 10 %, стверджуємо, що не існує доказів для визнання альтернативної гіпотези істинною.

5.4 ПЕРЕВІРКА ГІПОТЕЗИ ПРО ЗНАЧЕННЯ ЙМОВІРНОСТІ (ЧАСТКИ ОЗНАКИ) У ГЕНЕРАЛЬНІЙ СУКУПНОСТІ

Нехай задано генеральну сукупність, імовірність появи ознаки X якої вивчається. Припустимо, що із заданої генеральної сукупності реалізовано вибіркову сукупність обсягом n. Вважаємо, що для кожного з n випадково відібраних для вибірки елементів генеральної сукупності ймовірність появи ознаки X є однаковою, але невідомою величиною p_0. Інакше p_0 називають ***часткою ознаки X генеральної сукупності.*** За досить великим обсягом вибірки знайдено відносну частоту $w = \frac{m^*}{n}$ появи ознаки X у вибірковій сукупності, яка інакше називається ***часткою ознаки вибіркової сукупності***. За заданого рівня значущості α потрібно перевірити достовірність нульової гіпотези $H_0 : \rho = \rho_0$ про рівність спостережуваного значення відносної частоти $w = \frac{m^*}{n}$ появи ознаки X у вибірковій сукупності гіпотетичному значенню ймовірності ρ_0 появи ознаки X у генеральній сукупності.

Зауважимо, що знайдена відносна частота $w = \frac{m^*}{n}$ появи ознаки X у вибірковій сукупності не є випадковою величиною, оскільки її числове значення вже відоме після отримання результатів реалізації вибіркової сукупності, хоча до моменту закінчення незалежних випробувань (закінчення формування вибіркової сукупності) відносна частота $W = \frac{m}{n}$ появи ознаки X у вибірковій сукупності вважається випадковою величиною, оскільки може набувати різних випадкових значень.

Розглянемо випадкову величину $W = \frac{m}{n}$, яка має біноміальний розподіл. Для великих $n(n > 30)$ можна наближено вважати нормальним закон розподілу величини W з математичним сподіванням $\mathrm{M}(W) = p_0$, дисперсією $D(W) = \frac{p_0 q_0}{n}$ і середнім квадратичним відхиленням $\sigma(W) = \sqrt{\frac{p_0 q_0}{n}}$, де $q_0 = 1 - p_0$.

За статистичний критерій приймемо випадкову величину Z, яка за умови достовірності гіпотези H_0 має нормований нормальний закон розподілу:

$$Z = \frac{W - M(W)}{\sigma(W)} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}. \qquad (5.27)$$

Для обчислення спостережуваного значення критерію z^* у формулу (5.27) підставляємо замість $W = \frac{m}{n}$ числове значення цієї випадкової величини $W = \frac{m^*}{n}$, обчислене за результатами обробки вибірки:

$$z^* = \frac{m/n - p_0}{\sqrt{\frac{p_0 q_0}{n}}}. \qquad (5.28)$$

Критична область будується залежно від змісту альтернативної гіпотези H_1 аналогічно випадку перевірки гіпотези про значення генеральної середньої випадкової величини, що має нормальний закон розподілу.

Приклад 5.10. При виробництві продукції харчування проводиться контроль якості для забезпечення рівня частки неякісної продукції менше ніж 1 %. Серед відібраних 400 зразків продукції 5 виявились неякісними. Чи є підстави вважати, що виробничий процес вийшов з-під контролю і виробляється багато неякісної продукції?

Р о з в ’ я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : p = 0{,}01$ — частка неякісної продукції складає 1 %,

$H_1 : p > 0{,}01$ — частка неякісної продукції становить більше ніж 1 %, виробничий процес вийшов з-під контролю.

Крок другий. За статистичний критерій обираємо випадкову величину Z, яка є нормально розподіленою, нормованою у випадку достовірності нульової гіпотези, тобто $Z \sim N(0;1)$:

$$Z = \frac{W - M(W)}{\sigma(W)} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Крок третій. Обчислюємо спостережуване значення критерію z^*. Для цього знайдемо:

$$W = \frac{m^*}{n} = \frac{5}{400} = 0{,}0125,\ p_0 = 0{,}01,\ q_0 = 0{,}99,\ n = 400,$$

тоді:

$$z^* = \frac{m^*/n - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0{,}0125 - 0{,}01}{\sqrt{\frac{0{,}01 \cdot 0.99}{400}}} \approx 0{,}5025.$$

Крок четвертий. За змістом альтернативної гіпотези будуємо критичну правосторонню область. Рівень значущості за умовою задачі не задано, тож обираємо його самостійно, наприклад 5 %. Межу правосторонньої критичної області $z_{кр}$ знаходимо за формулою $\Phi(z_{кр}) = \frac{1-2\alpha}{2}$, $\Phi(z_{кр}) = \frac{1 - 2\cdot 0{,}05}{2} = \frac{0{,}9}{2} = 0{,}45$, звідки $z_{кр} = 1{,}645$. Отримали область відхилення H_0: $Z > 1{,}645$ (рис. 5.15).

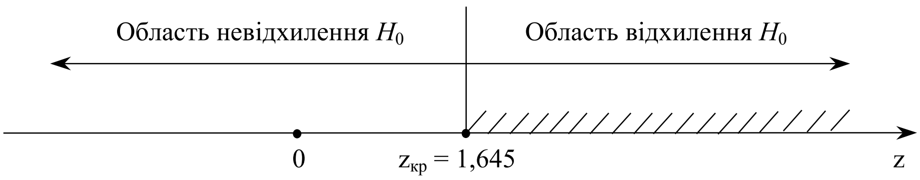


Рис. 5.15

Крок п'ятий. Оскільки спостережуване значення критерію не потрапляє до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних доказів з 5-відсотковим рівнем значущості на користь альтернативної гіпотези про те, що виробничий процес вийшов з-під контролю, а саме, що частка неякісної продукції становить більше ніж 1 %.

Зауваження 13. Обчислимо для попередньої задачі *p*-значення як найменший можливий рівень значущості, для якого нульова гіпотеза відхиляється для даного спостережуваного значення z^* обраного критерію:

$$P(Z > 0{,}5025) = 0{,}5 - P(0 \le Z \le 0{,}5025) = 0{,}5 - \Phi(0{,}5025) = 0{,}5 - 0{,}1915 = 0{,}3085.$$

Одержане значення 0,3085 перевищує 0,10, отже, не існує доказів для визнання альтернативної гіпотези істинною.

Приклад 5.11. Директор великої страхової компанії отримав дані відділу кадрів про те, що 25 % страхових агентів звільняються, не відпрацювавши й одного року. Для 150 страхових агентів на початку першого робочого року був проведений інтенсивний тренінг. Наприкінці року з цих 150 працівників звільнилося 29 осіб.

а) Чи можна вважати, що відсоток звільнених робітників, які пройшли тренінг, становить менше 25 %? Рівень значущості прийняти рівним $\alpha = 0{,}01$.

б) З яким найменшим рівнем значущості можна зробити висновок про можливість відхилення гіпотези про те, що 25 % страхових агентів звільняються, не відпрацювавши й одного року, навіть після того, як вони пройшли інтенсивний тренінг?

Р о з ' я з а н н я.

а) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

H_0: $p = 0{,}25$ – 25 % страхових агентів звільняються, не відпрацювавши й одного року, навіть після того, як вони пройшли інтенсивний тренінг;

H_1: $p < 0{,}25$ — відсоток звільнених робітників, які пройшли тренінг, становить менше 25 % (оскільки $\frac{29}{150} \approx 0{,}19 < 0{,}25$).

Крок другий. За статистичний критерій обираємо випадкову величину *Z*, яка є нормально розподіленою, нормованою у разі достовірності нульової гіпотези, тобто $Z \sim N(0;1)$:

$$Z = \frac{W - M(W)}{\sigma(W)} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim N(0;\ 1) \text{ у разі достовірності нульової гіпотези.}$$

Крок третій. Обчислюємо спостережуване значення критерію z^*. Для цього знайдемо:

$$w = \frac{m^*}{n} = \frac{29}{150} \approx 0{,}19,\ p_0 = 0{,}25,\ q_0 = 0{,}75,\ n = 150,$$

тоді:

$$z^* = \frac{0{,}19 - 0{,}25}{\sqrt{\frac{0{,}25 \cdot 0{,}75}{150}}} = \frac{-0{,}06}{0{,}03536} = -1{,}6968 \approx -1{,}7\,.$$

Крок четвертий. За змістом альтернативної гіпотези будуємо критичну лівосторонню область. За заданим рівнем значущості $\alpha = 0{,}01$ знайдемо межу лівосторонньої критичної області $z_{кр}$, яке знаходимо за формулою $\Phi(z_{кр}) = -\frac{1-2\alpha}{2}$, $\Phi(z_{кр}) = -\frac{1-2\cdot 0{,}01}{2} = -\frac{0{,}98}{2} = -0{,}49$, звідки $z_{кр} = -2{,}325$.

Дістали область відхилення H_0: $\{Z|Z < -2{,}325\}$.

Крок п'ятий. Оскільки спостережуване значення критерію не потрапляє до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези про те, що відсоток звільнених робітників, які пройшли тренінг, зменшився та становить менше 25 %.

б) Відповідь на запитання про найменший рівень значущості для висновку про можливість відхилення нульової гіпотези надає обчислення p-значення.

p-значення: $P(Z < -1{,}7) = P(Z > 1{,}7) = 0{,}5 - P(0 \le Z \le= 1{,}7) = 0{,}5 - \Phi(1{,}7) = 0{,}5 - 0{,}4554 = 0{,}0446$.

Відхилення нульової гіпотези для даного спостережуваного значення z^* обраного критерію на користь альтернативної гіпотези, можливо лише починаючи з рівня значущості 4,46 %.

Приклад 5.12. У гарантійному свідоцтві виробником лазерних принтерів певної марки стверджується, що не більше 10 % лазерних принтерів потребуватимуть ремонту протягом перших двох років експлуатації (виходячи з досвіду минулих років). З метою перевірки достовірності цього твердження товариством захисту прав споживачів була організована випадкова вибірка 100 покупців лазерних принтерів, яка виявила, що 14 із них ремонтували куплені лазерні принтери даної марки протягом перших двох років експлуатації.

а) Перевірте, чи є твердження виробника правильним з 1 %-м рівнем значущості. Чи підтверджує отриманий висновок знайдене p-значення?

б) Перевірте, чи є правильним твердження, що саме 10 % лазерних принтерів потребуватимуть ремонту протягом перших двох років експлуатації з 1 %-м рівнем значущості. Чи підтверджує отриманий висновок знайдене р-значення?

а) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : p \le 0{,}1$ — не більше 10 % лазерних принтерів потребують ремонту протягом перших двох років експлуатації (твердження виробника є правильним).

$H_1 : p > 0{,}1$ (оскільки $\frac{14}{100} = 0{,}14 > 0{,}1$) понад 10 % лазерних принтерів потребує ремонту протягом перших двох років експлуатації (твердження виробника є хибним).

Крок другий. За статистичний критерій обираємо випадкову величину Z, яка є нормально розподіленою, нормованою у разі достовірності нульової гіпотези, тобто $Z \sim N(0;1)$:

$$Z = \frac{W - M(W)}{\sigma(W)} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Крок третій. Обчислюємо спостережуване значення критерію:

$$w = \frac{m^*}{n} = \frac{14}{100} = 0{,}14,\ p_0 = 0{,}1,\ q_0 = 0{,}9,\ n = 100,$$

$$Z^* = \frac{0{,}14 - 0{,}1}{\sqrt{\frac{0{,}1 \cdot 0{,}9}{100}}} = \frac{0{,}04}{0{,}03} \approx 1{,}33\ .$$

Крок четвертий. За змістом альтернативної гіпотези будуємо критичну правосторонню область. Для заданого рівня значущості 1 % будуємо межу правосторонньої критичної області $z_{кр}$, яку знаходимо за формулою:

$$\Phi(z_{кр}) = \frac{1 - 2\alpha}{2},\quad \Phi(z_{кр}) = \frac{1 - 2 \cdot 0{,}01}{2} = \frac{0{,}98}{2} = 0{,}49,\ \text{звідки } z_{кр} = 2{,}325.$$

Отримали область відхилення гіпотези $H_0 : \{Z|Z > 2{,}325\}$.

Крок п'ятий. Оскільки спостережуване значення критерію не потрапляє до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези про те, що твердження виробника є хибним, а саме, що більше ніж 10 % лазерних принтерів потребує ремонту протягом перших двох років експлуатації.

p-значення: $P(Z > 1{,}33) = 0{,}5 - \Phi(1{,}33) = 0{,}5 - 0{,}4082 = 0{,}0918$.

Відхилення нульової гіпотези для даного спостережуваного значення z^* обраного критерію на користь альтернативної гіпотези можливе лише починаючи з рівня значущості 9,18 %.

б) ***Крок перший***. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : p = 0{,}1$ — 10 % лазерних принтерів потребує ремонту протягом перших двох років експлуатації.

$H_1 : p \neq 0{,}1$ — кількість лазерних принтерів, що потребує ремонту протягом перших двох років експлуатації, не дорівнює 10 %.

Крок другий. За статистичний критерій обираємо випадкову величину Z, яка є нормально розподіленою, нормованою у разі достовірності нульової гіпотези, тобто $Z \sim N(0;1)$:

$$Z = \frac{W - M(W)}{\alpha(W)} = \frac{\frac{m^*}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Крок третій. Обчислюємо спостережуване значення критерію z^*:

$$w = \frac{m^*}{n} = \frac{14}{100} = 0{,}14,\ p_0 = 0{,}1,\ q_0 = 0{,}9,\ n = 100,$$

$$\text{тоді } z^* = \frac{0{,}14 - 0{,}1}{\sqrt{\frac{0{,}1 \cdot 0{,}9}{100}}} = \frac{0{,}04}{0{,}03} \approx 1{,}33.$$

Крок четвертий. За змістом альтернативної гіпотези будуємо критичну двосторонню область. Для заданого рівня значущості 1 % будуємо межі правосторонньої та лівосторонньої критичних областей, які знаходимо за формулами:

$$\Phi(z_{\text{кр}}^{\text{пр}}) = \frac{1-\alpha}{2},\ \Phi(z_{\text{кр}}^{\text{пр}}) = \frac{1-0{,}01}{2} = \frac{0{,}99}{2} = 0{,}495\text{, звідки } z_{\text{кр}}^{\text{пр}} = 2{,}575,$$

$$\text{та } \Phi(z_{\text{кр}}^{\text{лів}}) = -\frac{1-\alpha}{2},\ \Phi(z_{\text{кр}}^{\text{лів}}) = -0{,}495\text{, звідки } z_{\text{кр}}^{\text{лів}} = -2{,}575.$$

Отримали область невідхилення гіпотези $H_0 : Z \in (-2{,}575,\ 2{,}575)$.

Крок п'ятий. Оскільки спостережуване значення критерію не потрапляє до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези.

p-значення:

$$P(Z < -1{,}33) + P(Z > 1{,}33) = 2P(Z > 1{,}33) = 2 \cdot (0{,}5 - \Phi(1{,}33)) = 2 \cdot (0{,}5 - 0{,}4082) = 2 \cdot 0{,}0918 = 0{,}1836.$$

Отримане p-значення більше, ніж 0,1, отже робимо висновок, що не існує доказів для визнання альтернативної гіпотези істинною.

Приклад 13. Аналітик з інвестицій промислової компанії стверджує, що 70 % акцій, які він радив придбати, зросли в ціні. Припустимо, що випадкова вибірка 125 акцій виявила, що 85 з них зросли в ціні.

а) Перевірити справедливість твердження аналітика з рівнем значущості α = 0,05.

б) Обчислити р-значення та дати його інтерпретацію.

Р о з в ' я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : p = 0{,}7$ – 70% акцій, які радив придбати аналітик, зросли в ціні,

$H_1 : p \neq 0{,}7$ — твердження аналітика не правильне.

Крок другий. За статистичний критерій обираємо випадкову величину Z, яка є нормально розподіленою, нормованою у разі достовірності нульової гіпотези, тобто $Z \sim N(0;1)$:

$$Z = \frac{W - M(W)}{\sigma(W)} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Крок третій. Обчислюємо спостережуване значення критерію z^*.

Маємо:

$$w = \frac{m}{n} = \frac{85}{125} = 0{,}68,\ p_0 = 0{,}7,\ q_0 = 0{,}3,\ n = 125,$$

тоді $z^* = \dfrac{0{,}6 - 0{,}7}{\sqrt{\dfrac{0{,}7 \cdot 0{,}3}{125}}} \approx -0{,}49$.

Крок четвертий. За змістом альтернативної гіпотези будуємо критичну двосторонню область. Для заданого рівня значущості 5 % межі правосторонньої та лівосторонньої критичних областей знаходимо за формулами:

$$\Phi(z_{\text{кр}}^{\text{пр}}) = \frac{1-\alpha}{2},\ \Phi(z_{\text{кр}}^{\text{пр}}) = \frac{1-0{,}05}{2} = 0{,}5 - 0{,}025 = 0{,}475,\ \text{звідки}\ z_{\text{кр}}^{\text{пр}} = 1{,}96;$$

$$\Phi(z_{\text{кр}}^{\text{лів}}) = -\frac{1-\alpha}{2},\ \Phi(z_{\text{кр}}^{\text{лів}}) = -0{,}475,\ \text{звідки}\ z_{\text{кр}}^{\text{лів}} = -1{,}96.$$

Отримали область невідхилення гіпотези $H_0 : Z \in (-1{,}96,\ 1{,}96)$.

Крок п'ятий. Оскільки спостережуване значення критерію не потрапляє до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з 5 %-м рівнем значущості на користь альтернативної гіпотези про те, що твердження аналітика неправильне.

б) Обчислимо p-значення:

$$2 \cdot P(Z < -0{,}49) = 2 \cdot P(Z > 0{,}49) = 2 \cdot (0{,}5 - \Phi(0{,}49)) = 2 \cdot (0{,}5 - 0{,}1879) = 2 \cdot 0{,}3121 = 0{,}6242.$$

Отримане p-значення більше ніж 0,1, отже, робимо висновок, що не існує доказів для визнання альтернативної гіпотези істинною.

5.5. ПОРІВНЯННЯ ДИСПЕРСІЙ ДВОХ НОРМАЛЬНО РОЗПОДІЛЕНИХ СУКУПНОСТЕЙ

Необхідність перевірки гіпотези про порівняння двох дисперсій виникає досить часто, оскільки дисперсія характеризує такі важливі показники, як точність приладів технологічних вимірювань, точність методів досліджень, ризик інвестицій тощо. Зрозуміло, що найбільшу точність технологічних вимірювань має той прилад, який забезпечує найменше розсіювання результатів вимірювань, тобто має найменшу дисперсію.

Маємо дві нормально розподілені генеральні сукупності з ознаками X_1 і X_2, дисперсії яких невідомі. Необхідно перевірити нульову гіпотезу про рівність дисперсій ознак цих генеральних сукупностей $H_0 : D(X_1) = D(X_2)$. Для відповіді на поставлене запитання обираємо незалежні вибірки обсягів n_1 та n_2 з кожної генеральної сукупності та знаходимо відповідні виправлені вибіркові дисперсії s_1^2 і s_2^2. Відомо, що виправлені вибіркові дисперсії є незміщеними оцінками генеральних дисперсій, тобто, $\text{M}(s_1^2) = D(X_1)$, $\text{M}(s_2^2) = D(X_2)$, Отже, необхідно перевірити, що математичні сподівання виправлених вибіркових дисперсій рівні між собою. Зрозуміло, що майже завжди знайдені числові значення виправлених вибіркових дисперсій є різними. Це можна пояснити однією із двох причин: або дисперсії генеральних сукупностей дійсно різні, або випадковий відбір елементів вибірки був невдалим (інакше це називають помилкою вибірки). Тобто, якщо в результаті перевірки нульової гіпотези $H_0 : D(X_1) = D(X_2)$ зроблено висновок про недостатність статистичних фактів для її відхилення, то різні числові значення виправлених вибіркових дисперсій пояснюються випадковими причинами, а саме випадково невдалим відбором елементів вибіркових сукупностей тощо. Наприклад, якщо $s_1^2 \neq s_2^2$ — різні виправлені вибіркові дисперсії результатів технологічних вимірювань двох приладів, але статистичних доказів для відхилення нульової гіпотези недостатньо, робимо висновок про недостатність статистичних доказів на користь альтернативної гіпотези про різну точність приладів.

Якщо в результаті перевірки нульової гіпотези $H_0 : D(X_1) = D(X_2)$ зроблено висновок про недостатність статистичних фактів на її підтримку, то різні числові значення виправлених вибіркових дисперсій пояснюються тим, що генеральні дисперсії різні. Наприклад, якщо $s_1^2 \neq s_2^2$ — різні виправлені вибіркові дисперсії результатів технологічних вимірювань двох приладів і є достатньо доказів для відхилення нульової гіпотези $H_0 : D(X_1) = D(X_2)$, тоді робимо висновок про достатність статистичних доказів на користь альтернативної гіпотези про різну точність приладів.

Для перевірки гіпотези про рівність дисперсій двох нормально розподілених генеральних сукупностей необхідно обрати статистичний критерій.

Порівняння генеральних дисперсій σ_1^2 і σ_2^2 здійснюється зіставленням виправлених дисперсій s_1^2, s_2^2, які відповідно мають закон розподілу χ^2 із $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$ ступенями вільності, де n_1 і n_2 — обсяги першої і другої незалежних вибірок, зроблених із генеральних сукупностей з ознаками X_1 і X_2 відповідно.

За статистичний критерій приймемо відношення більшої $s_б^2$ з двох виправлених вибіркових дисперсій s_1^2 та s_2^2 до меншої виправленої вибіркової дисперсії $s_м^2$.

Отже, за статистичний критерій обирається випадкова величина:

$$F = \frac{s_б^2}{s_м^2}. \tag{5.29}$$

Випадкова величина F, за умови справедливості нульової гіпотези, має розподіл Фішера-Снедекора із $v_1 = n_б - 1$, $v_2 = n_м - 1$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія s_σ^2, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2$.

Нагадаємо, що розподіл Фішера-Снедекора залежить тільки від кількості ступенів вільності ти не є симетричним.

Щільність імовірностей розподілу Фішера-Снедекора:

$$g(f) = \frac{r\left(\frac{v_1 + v_2}{2}\right)}{r\left(\frac{v_1}{2}\right) r\left(\frac{v_2}{2}\right)} \cdot \left(\frac{v_2}{v_1}\right)^{\frac{v_2}{2}} (f)^{\frac{v_2}{2}-1} \left(1 + \frac{v_2}{v_1} f\right)^{\frac{v_1+v_2}{2}},\ f \ge 0 \tag{5.30}$$

визначена лише на додатній півосі, тобто $0 \le f < \infty$ (рис. 5.16).

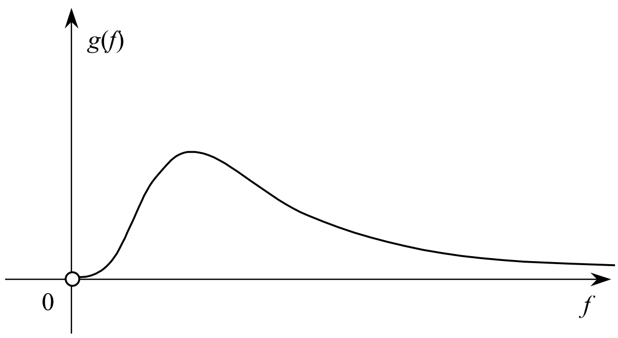


Рис. 5.16

Залежно від вигляду альтернативної гіпотези H_1 будується критична область.

Правило. Для перевірки із заданим рівнем значущості α нульової гіпотези $H_0 : D(X_1) = D(X_2)$ (або $H_0 : \sigma_1^2 = \sigma_2^2$) про рівність дисперсій двох нормально розподілених генеральних сукупностей за альтернативної гіпотези $H_1 : D(X_1) > D(X_2)$ знаходиться спостережуване значення f^* критерію $F = \frac{s_б^2}{s_м^2} = \frac{s_1^2}{s_2^2}$, як відношення більшої вибіркової дисперсії до меншої:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{s_1^2}{s_2^2}. \tag{5.31}$$

Будується правостороння критична область. Критична точка — межа правосторонньої критичної області — визначається із таблиці розподілу Фішера-Снедекора (таблиця додатка 9) за відповідними значеннями рівня значущості α та кількості ступенів вільності $v_1 = n_б - 1$ і $v_2 = n_м - 1$, а саме $f_{кр} = F(\alpha, v_1 = n_б - 1\ v_2 = n_м - 1)$, Якщо спостережуване значення критерію знаходиться в критичній області $f^* > f_{кр}$, то робимо висновок про достатність статистичних фактів для відхилення нульової гіпотези $H_0 : D(X_1) = D(X_2)$ на користь альтернативної H_1: $D(X_1) > D(X_2)$.

Зауважимо, що для перевірки із заданим рівнем значущості α нульової гіпотези H_0: $D(X_1) = D(X_2)$ за альтернативної гіпотези H_1: $D(X_1) \ne D(X_2)$ будується двостороння критична область так, щоб за умови справедливості нульової гіпотези ймовірність потрапляння спостережуваного значення критерію до цієї області дорівнювала би α. Позначимо критичні точки — межі двосторонньої критичної області через f_1 (ліва межа) та f_2 (права межа) (рис. 5.17).

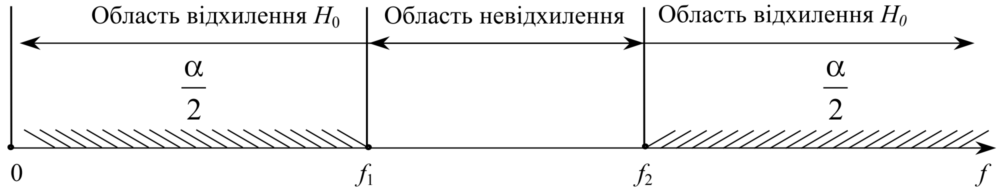


Рис. 5.17

Найбільша потужність критерію досягається, якщо ймовірність потрапляння спостережуваного значення критерію до кожного з двох інтервалів двосторонньої критичної області дорівнює $\frac{\alpha}{2}$, тобто $P(f < f_1) = \frac{\alpha}{2}$, і $P(f > f_2) = \frac{\alpha}{2}$. Права критична точка f_2 визначається із таблиці розподілу Фішера-Снедекора (таблиця додатка 9) за відповідними значеннями рівня значущості $\frac{\alpha}{2}$ та кількістю ступенів вільності $v_1 = n_б - 1$ і $v_2 = n_м - 1$, а саме $f_2 = F(\frac{\alpha}{2}, v_1 = n_б - 1, v_2 = n_м - 1)$.

Зауважимо, що таблиця містить тільки критичні точки — межі правосторонніх критичних областей. Отже, ліву межу двосторонньої критичної області — критичну точку f_1 за таблицею знайти не можна. Формула знаходження критичної точки f_1 відома, але важливіше зрозуміти, чому достатньо знайти лише критичну точку f_2 — праву межу двосторонньої критичної області.

Важливим є той факт, що очікуване спостережуване значення $f^* = \frac{s_б^2}{s_м^2}$ статистичного критерію F дорівнює одиниці, оскільки припускаємо, що дисперсії однакові (тобто їх відношення дорівнює одиниці). Як тільки в чисельнику, обраного для перевірки нульової гіпотези статистичного критерію $F = \frac{s_б^2}{s_м^2}$, ставимо більшу з виправлених дисперсій $s_б^2$, одразу отримаємо спостережуване значення $f^* = \frac{s_б^2}{s_м^2}$ статистичного критерію F більше одиниці. Тобто спостережуване значення f^* критерію F завжди буде знаходитися праворуч від очікуваного значення — одиниці (правіше центру розподілу — математичного сподівання). Ось чому достатньо знайти лише критичну точку f^* — праву межу двосторонньої критичної області.

Правило. Для перевірки із заданим рівнем значущості α нульової гіпотези $H_0 : D(X_1) = D(X_2)$ (або $H_0 : \sigma_1^2 = \sigma_2^2$) про рівність дисперсій двох генеральних сукупностей за альтернативної гіпотези $H_1 : D(X_1) \neq D(X_2)$ знаходиться спостережуване значення f^* критерію $F = \frac{s_б^2}{s_м^2}$ як відношення більшої вибіркової дисперсії до меншої. Критична точка — межа правосторонньої критичної області — визначається із таблиці розподілу Фішера-Снедекора (таблиця додатка 9) за відповідними значеннями рівня значущості $\frac{\alpha}{2}$ (у два рази меншим, ніж заданий) та ступенів вільності:

$$v_1 = n_б - 1, \text{ і } v_2 = n_м - 1, \text{ а саме } f_{кр} = F(\frac{\alpha}{2}, v_1 = n_б - 1, v_2 = n_м - 1).$$

Якщо спостережуване значення критерію знаходиться в критичній області $f^* > f_{кр}$, то робимо висновок про достатність статистичних фактів для відхилення нульової гіпотези $H_0 : D(X_1) = D(X_2)$ на користь альтернативної $H_1 : D(X_1) \neq D(X_2)$.

Зауваження 14. Для перевірки на заданому рівні значущості α нульової гіпотези H_0: $D(X_1) = D(X_2)$ (або $H_0 : \sigma_{X_1}^2 = \sigma_{X_1}^2$) про рівність дисперсій двох генеральних сукупностей за альтернативної гіпотези H_1: $D(X_1) < D(X_2)$ не потрібно будувати лівосторонню критичну область. Слід альтернативну гіпотезу H_1: $D(X_1) < D(X_2)$ записати як H_1: $D(X_2) > D(X_1)$ та застосувати правило знаходження критичної точки для правосторонньої критичної області до побудови області відхилення та невідхилення нульової гіпотези.

Приклад 5.14. Однакова продукція виробляється на двох технологічних лініях. На другій лінії зроблено вдосконалення, котре привело до скорочення варіації часу виготовлення продукції. Вибіркова перевірка варіації часу виготовлення на першій і другій лініях дала значення $s_1^2 = 1{,}05$ для 13 спостережень *і* $s_2^2 = 0{,}6$ для 12 спостережень. Чи можна вважати суттєвою розбіжність між варіаціями тривалості процесу виготовлення продукції на першій і другій лініях, якщо час виготовлення продукції змінюється за нормальним законом розподілу?

Р о з в ’ я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0: \sigma_1^2 = \sigma_2^2$ — суттєвої розбіжності між варіаціями тривалості процесу виготовлення продукції на першій і другій лініях немає;

$H_0: \sigma_1^2 > \sigma_2^2$ — є суттєва розбіжність, а саме варіація часу на першій лінії більша, ніж на другій.

Крок другий. За статистичний критерій обирається випадкова величина $F = \dfrac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $v_1 = n_б - 1$, і $v_2 = n_м - 1$, ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_М^2$.

Крок третій. Знаходимо спостережуване значення критерію:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{1{,}05}{0{,}6} = 1{,}75\,.$$

Число ступенів вільності для більшої виправленої дисперсії $s_б^2 = s_1^2 = 1{,}05$ становить $v_1 = n_б - 1 = 13 - 1 = 12$, для меншої виправленої дисперсії $s_М^2 = s_2^2 = 0{,}6$ – буде становить $v_2 = n_м - 1 = 12 - 1 = 11$

Крок четвертий. За змістом альтернативної гіпотези $H_1: \sigma_1^2 > \sigma_2^2$ будуємо правосторонню критичну область. Для заданого рівня значущості 1 %, ступенів вільності v_1 і v_2 за таблицею додатка 9 знаходимо межу правосторонньої критичної області — критичну точку:

$$f_{кр} = F(\alpha = 0{,}01;\ v_1 = 12;\ v_2 = 11) = 44.$$

Отримали критичну область — область відхилення нульової гіпотези H_0 : $\{F \mid F > 4{,}4\}$ (рис. 5.18).

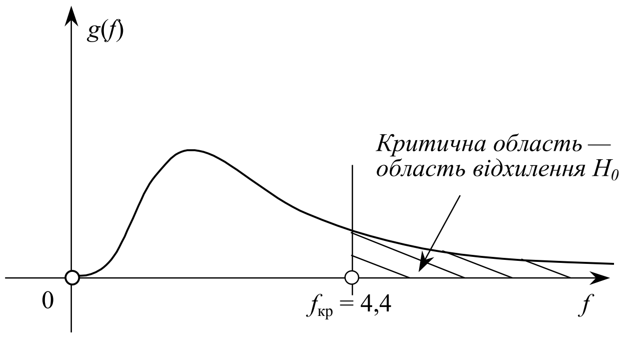


Рис. 5.18

Крок п’ятий. Оскільки спостережуване значення критерію $f^* = 1{,}75$ не потрапляє до області відхилення нульової гіпотези $\{f \mid f > 4{,}4\}$ робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези для висновку про суттєву розбіжність між варіаціями тривалості процесу виготовлення продукції на першій і другій лініях.

Приклад 5.15. Під час дослідження стабільності температури в термостаті дістали такі результати: 21,2; 21,8; 21,3; 21,0; 21,4; 21,3. З метою стабілізації температури було використано удосконалений пристрій, після цього заміри температури показали такі результати: 37,7; 37,6; 37,6; 37,4. Чи можна з рівнем значущості $\alpha = 0{,}01$ вважати використання удосконаленого пристрою для стабілізації температури ефективним, якщо закон розподілення обох випадкових величин — нормальний?

Р о з в ’ я з а н н я.

Очевидно, що ефективність стабілізаторів без удосконаленого пристрою і з ним залежить від дисперсій вимірюваних ними температур. Отже, задача звелась до порівняння двох дисперсій.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0: \sigma_X^2 = \sigma_Y^2$ — використання удосконаленого пристрою для стабілізації температури не ефективне;

$H_0: \sigma_X^2 > \sigma_Y^2$ — використання удосконаленого пристрою для стабілізації температури є ефективним.

Крок другий. За статистичний критерій обирається випадкова величина $F = \frac{s_б^2}{s_М^2}$, яка має розподіл Фішера-Снедекора із $v_1 = n_б - 1$, і $v_2 = n_м - 1$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_М^2$.

Крок третій. Знаходимо спостережуване значення критерію, для чого обчислюємо виправлені вибіркові дисперсії. Маємо:

$$\bar{x}_в = \frac{\sum x_i n_i}{n_X} = \frac{21{,}2 + 21{,}4 + 21{,}0 + 21{,}3 \cdot 2 + 21{,}8}{6} \approx 21{,}333;$$

$$\frac{\sum x_i^2 n_{x_i}}{n_X} = \frac{21{,}2^2 \cdot 1 + 21{,}4^2 \cdot 1 + 21{,}0^2 \cdot 1 + 21{,}3^2 \cdot 2 + 21{,}8^2 \cdot 1}{6} = \frac{2731{,}02}{6} =$$
$$= 455{,}17;$$

$$D_в(X) = \frac{\sum x_i^2 n_{x_i}}{n_X} - (\bar{x}_в)^2 = 455{,}17 - (21{,}333)^2 \approx 455{,}17 - 455{,}097 = 0{,}073;$$

$$s_x^2 = \frac{n_X}{n_X - 1} = D\text{в}(\text{X}) = \frac{6}{6-1} \cdot 0{,}073 = 0{,}0876;$$

$$\bar{y}_в = \frac{\sum y_j n_{yj}}{n_Y} = \frac{37{,}7 + 37{,}6 \cdot 2 + 37{,}4}{4} = \frac{37{,}7 + 75{,}2 + 37{,}4}{4} = \frac{150{,}3}{4} = 37{,}575;$$

$$\frac{\sum y_j^2 n_{yj}}{n_Y} = \frac{37{,}7^2 \cdot 1 + 37{,}6^2 \cdot 2 + 37{,}4^2 \cdot 1}{4} = \frac{5647{,}57}{4} = 1411{,}8925;$$

$$D_в(Y) = \frac{\sum y_j^2 n_{yj}}{n_Y} - (\bar{y}_в)^2 = 1411{,}8925 - (37{,}575)^2 =$$
$$= 1411{,}8925 - 1411{,}880625 = 0{,}011875;$$

$$s_Y^2 = \frac{n_Y}{n_Y - 1} D_в(Y) = \frac{4}{4-1} \cdot 0{,}011875 \approx 0{,}01583.$$

Обчислимо спостережуване значення критерію:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{0{,}0876}{0{,}01583} \approx 5{,}534.$$

Число ступенів вільності для більшої виправленої дисперсії $s_б^2 = s_X^2 = 0{,}0876$, $v_1 = n_X - 1 = 6 - 1 = 5$, для меншої виправленої дисперсії $s_м^2 = s_Y^2 = 0{,}01583, v_2 = n_Y - 1 = 4 - 1 = 3$.

Крок четвертий. За змістом альтернативної гіпотези H$_1$: $\sigma_X^2 > \sigma_Y^2$ будуємо критичну правосторонню область. Для заданого рівня значущості 1 %, ступенів вільності v_1 і v_2 за таблицею додатка 9 знаходимо межу правосторонньої критичної області — критичну точку:

$$f_{кр} F(\alpha = 0{,}01; v_1 = 5; v_2 = 3) = 28{,}2$$

Отримали область відхилення нульової гіпотези H_0: $\{f \mid f > 28{,}2\}$.

Схематично правобічну критичну область зображено на рис. 5.19.

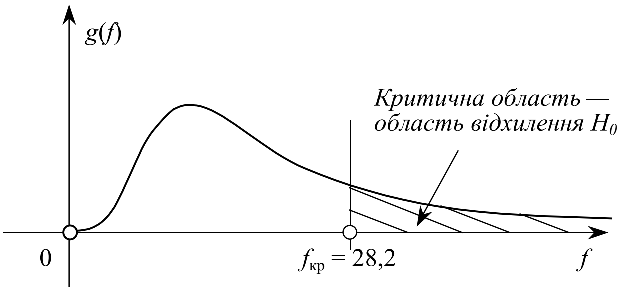


Рис. 5.19

Крок п'ятий. Оскільки спостережуване значення критерію $f^* = 5{,}534$ не належить до області відхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези про використання удосконаленого пристрою для стабілізації температури.

Приклад 5.16. Компанія *K* досліджує можливості отримання доходу від двох інвестицій *A* і *B*. Інвестиція *A* передбачається на строк 10 років з очікуваним прибутком 17,8 % річних. Інвестиція В розрахована на термін 8 років з таким самим очікуваним прибутком. Вибіркові дисперсії щорічних прибутків від інвестицій становлять відповідно для А — $(7{,}14\ \%)^2$, для В — $(3{,}21\ \%)^2$. Чи є підстава стверджувати, що ризики інвестицій А і В різні для рівня значущості 5 %? Передбачається, що випадкова величина щорічного прибутку на інвестиції має нормальний закон розподілу.

Р о з в ' я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези.

$H_0: \sigma_A^2 > \sigma_B^2$ — суттєвої розбіжності між ризиками для інвестицій *A* і *B* немає;

$H_1: \sigma_A^2 \neq \sigma_B^2$ — є суттєва розбіжність між ризиками для інвестицій *A* і *B*.

Крок другий. За статистичний критерій обирається випадкова величина $F = \frac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $\nu_1 = n_б - 1$, і $\nu_2 = n_м - 1$, ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2$.

Крок третій. Для знаходження спостережуваного значення f^* критерію *F*, знайдемо $s_б^2 = \frac{10}{9} \times$ $\times (7{,}14)^2\ (\%) = 56{,}644\ (\%)$, $s_м^2 = \frac{8}{7} \cdot (3{,}21)^2(\%) \approx 11{,}7761(\%)$,

$$f^* = \frac{s_б^2}{s_м^2} = \frac{56{,}644}{11{,}7761} \approx 4{,}81.$$

Число ступенів вільності для більшої виправленої дисперсії $s_б^2$ становить $\nu_1 = 8 - 10 = 9$, для меншої виправленої дисперсії $s_м^2$ становить $\nu_2 = 8 - 1 = 7$

Крок четвертий. За змістом альтернативної гіпотези H_1: $\sigma_X^2 \neq \sigma_Y^2$ для рівня значущості $\frac{5\ \%}{2} = 2{,}5(\%)$, ступенів вільності ν_1 і ν_2 за таблицею (додаток 9) знаходимо межу правосторонньої критичної області — критичну точку (рис. 5.20).

$$f_{кр} = F(\frac{\alpha}{2} = 0{,}025;\ \nu_1 = 10 - 1 = 9,\ \nu_2 = 8 - 1 = 7) = 4{,}82.$$

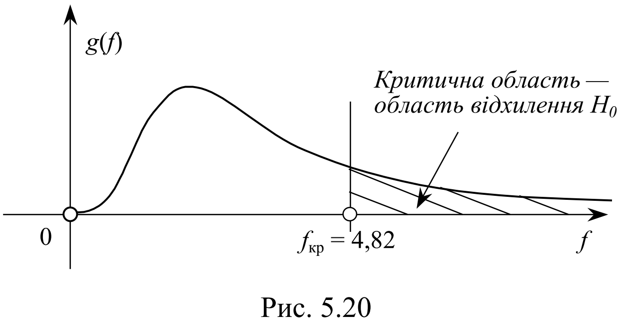


Рис. 5.20

Крок п'ятий. Оскільки спостережуване значення критерію $f^* = 4{,}81$ не належить області відхилення нульової гіпотези $\{f \mid f > 28{,}2\}$, робимо висновок про недостатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези про суттєву розбіжність між ризиками для інвестицій A і B.

Приклад 5.17. Норму витрат на технічне обслуговування і ремонт нових марок автомобілів досліджували на двох станціях технічного обслуговування. Результати спостережень наведено в таблицях двох статистичних розподілів:

x_i	0,56	0,6	0,64	0,7	0,74
n_{x_i}	4	6	3	2	1

y_j	0,58	0,6	0,62	0,64	0,66
n_{y_j}	2	3	10	4	1

Ознаки X і Y (норми витрат, які обчислені в гривнях на 1 км) є незалежними випадковими величинами, які мають нормальний закон розподілу. З рівнем значущості перевірити правильність нульової гіпотези $H_0: \sigma_X^2 \neq \sigma_Y^2$, якщо альтернативна гіпотеза: а) $H_1: \sigma_Y^2 < \sigma_X^2$; б) $H_1: \sigma_Y^2 \neq \sigma_X^2$.

Р о з в ' я з а н н я.

Проведемо необхідні обчислення за даними таблиць вибіркових розподілів.

За вибірковим розподілом X: $n_X = 16$, $\bar{x}_B = 0{,}61875$,

$$D_B(X) = 0{,}385675 - (0{,}61875)^2 \approx 0{,}0028234,\ s_X^2 = \frac{16}{15} \cdot 0{,}0028234 \approx 0{,}0030116.$$

За вибірковим розподілом Y: $n_Y = 20$, $\bar{y}_B = 0{,}619$,

$$D_B(Y) = 0{,}38354 - (0{,}619)^2 \approx 0{,}000379,\ s_Y^2 = \frac{20}{19} 0{,}000379 \approx 0{,}0004.$$

Оскільки $s_X^2 > s_Y^2$, за $s_б^2$ обираємо $s_X^2 = 0{,}0030116$, $n_б = n_X = 16$, а за $s_м^2$ відповідно $s_Y^2 = 0{,}0004$, $n_м = n_Y = 20$.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0: s_X^2 = s_Y^2$ — суттєвої розбіжності між дисперсіями норм витрат немає.

H_1: $s_Y^2 < s_X^2$, — є суттєва розбіжність між дисперсіями норм витрат на технічне обслуговування і ремонт нових марок автомобілів на двох станціях технічного обслуговування.

Крок другий. За статистичний критерій обирається випадкова величина $F = \frac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $v_1 = n_б - 1 = 16 - 1 = 15$, $v_2 = n_м = 20 - 1 = 19$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2$.

Крок третій. Обчислюємо спостережуване значення f^* критерію F:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{0{,}0030116}{0{,}0004} = 7{,}529.$$

Крок четвертий. Альтернативну гіпотезу H_1: $\sigma_Y^2 < \sigma_X^2$, для зручності, переписуємо у вигляді $H_1: \sigma_X^2 \neq \sigma_Y^2$ для побудови правосторонньої критичної області.

За таблицею (додаток 9) знаходимо границю правосторонньої критичної області — критичну точку $f_{кр} = F(\alpha = 0{,}01;\ v_1 = 15,\ v_2 = 19) = 3{,}15$, що відповідає рівню значущості 1 % (рис. 5.21).

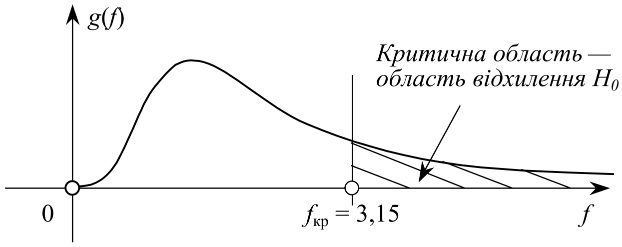


Рис. 5.21

Крок п'ятий.

Оскільки спостережуване значення критерію $f^* = 7{,}529$ знаходиться в області відхилення нульової гіпотези $\{f \mid f > 3{,}15\}$, робимо висновок про достатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези про суттєву розбіжність між дисперсіями норм витрат на технічне обслуговування і ремонт нових марок автомобілів на двох станціях технічного обслуговування.

б) ***Крок перший***. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0: \sigma_X^2 - \sigma_Y^2$ — немає суттєвої розбіжності між дисперсіями норм витрат на технічне обслуговування і ремонт нових марок автомобілів на двох станціях технічного обслуговування.

$H_1: \sigma_Y^2 \neq \sigma_X^2$, — є суттєва розбіжність між дисперсіями норм витрат на технічне обслуговування і ремонт нових марок автомобілів на двох станціях технічного обслуговування.

Крок другий. За статистичний критерій обирається випадкова величина $F = \frac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $\nu_1 = n_б - 1$, $\nu_2 = n_м - 1$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2$.

Крок третій. Обчислюємо спостережуване значення f^* критерію F:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{0{,}0030116}{0{,}0004} = 7{,}529.$$

Крок четвертий. Для альтернативної гіпотези $H_1: \sigma_Y^2 \neq \sigma_X^2$ критичну точку (рис. 5.22) знаходимо, як $f_{кр} = F(\frac{\alpha}{2} = 0{,}005;\ \nu_1 = 15,\ \nu_2 = 19) = 3{,}59$.

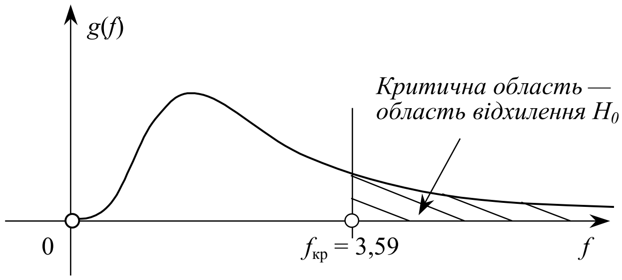


Рис. 5.22

Крок п'ятий. Оскільки спостережуване значення критерію $f^* = 7{,}529$ знаходиться в області відхилення нульової гіпотези $\{f \mid f > 3{,}59\}$, робимо висновок про достатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези про суттєву розбіжність між дисперсіями норм витрат на технічне обслуговування і ремонт нових марок автомобілів на двох станціях технічного обслуговування.

Послідовність розрахунків при перевірці гіпотез про рівність (чи розбіжність) дисперсій розглянемо ще на такому прикладі.

Приклад 5.18. Визначення вмісту (у грамах) харчових добавок у двох партіях продукції подано двома розподілами:

x_i (г)	4,45	5,05	5,55	6,05	6,55
n_{x_i}	1	2	4	4	5

y_j (г)	4,54	5,14	5,70	6,2	6,74
n_{x_i}	5	4	3	2	1

Ознаки X і Y (вміст добавок) є незалежними нормально розподіленими випадковими величинами.

а) З рівнем значущості α перевірити гіпотезу $H_0: D(X) = D(Y)$ за альтернативної гіпотези $H_1: D(X) < D(Y)$.

б) Розв'язати цю задачу для рівня значущості $\alpha = 0{,}1$ *та за альтернативної гіпотези* $H_1 : D(X) \neq$ $\neq D(Y)$.

Р о з в ' я з а н н я. **Проведемо необхідні обчислення за даними таблиць вибіркових розподілів. Обчислимо виправлені вибіркові дисперсії s_x^2 та s_y^2. Для цього знайдемо:**

$$n_x = \sum_{i=1}^{5} n_{x_i} = 1 + 2 + 4 + 4 + 5 = 16,$$

$$n_y = \sum_{i=1}^{5} n_{y_i} = 5 + 4 + 3 + 2 + 1 = 15,$$

$$\bar{x}_B = \frac{1}{n_X}\sum_{i=1}^{5} x_i n_{x_i} = \frac{1}{16}(4{,}45 + 5{,}05 \cdot 2 + 5{,}55 \cdot 4 + 6{,}05 \cdot 4 + 6б55 \cdot 5) =$$

$$= \frac{1}{16}(4{,}45 + 10{,}1 + 22{,}2 + 32{,}75) \approx 5{,}856\,;$$

$$\bar{y}_B = \frac{1}{n_Y}\sum_{i=1}^{5} y_j n_{y_j} = \frac{1}{15}(4{,}54 \cdot 5 + 5{,}14 \cdot 4 + 5{,}7 \cdot 3 + 6{,}2 \cdot + 6{,}74) = \frac{1}{15}(22{,}7 + 20{,}56 + 17{,}1 + 12{,}4 + 6{,}74) = 5{,}3\,;$$

$$\overline{x^2} = \frac{1}{n_X}\sum_{i=1}^{5} x_i^2 n_{x_i} = \frac{1}{16}((4{,}45)^2 \cdot 1 + (5{,}05)^2 \cdot 2 + (5{,}55)^2 \cdot 4 + (6{,}05)^2 \cdot 4 + (6{,}55)^2 \cdot 5) = \frac{1}{16}(19{,}8025 +$$

$$+ 51{,}005 + 123{,}21 + 146{,}41 + 214{,}5126) \approx 34{,}684;$$

$$\overline{y^2} = \frac{1}{n_Y}\sum_{j=1}^{5} y_j^2 n_{y_j} = \frac{1}{15}((4{,}54)^2 \cdot 5 + (5{,}14)^2 \cdot 4 + (5{,}7)^2 \cdot 3 + (6{,}2)^2 \cdot 2 + (6{,}74)^2 \cdot 1) \approx 28{,}5676\,.$$

Тоді

$$D_{\text{в}}(X) = \bar{x}^2 - (\bar{x}_B)^2 = 34{,}684 - (5{,}856)^2 \approx 0{,}3913\,,$$

$$D_{\text{в}}(Y) = \bar{y}^2 - (\bar{y}_{\text{B}})^2 = 28{,}5676 - (5{,}3)^2 \approx 0{,}4776\,.$$

$$s_X^2 = \frac{n_X}{n_x - 1} \cdot D_{\text{B}}(X) = \frac{16}{15} \cdot 0{,}3913 \approx 0.4174\,,$$

$$s_Y^2 = \frac{n_Y}{n_Y - 1} \cdot D_B(Y) = \frac{15}{14} \cdot 0{,}4776 \approx 0{,}5117\,.$$

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : D(X) = D(Y)$ — суттєвої розбіжності у варіації вмісту (у грамах) харчових добавок у двох партіях продукції немає.

$H_1 : D(X) = D(Y)$ — є суттєва розбіжність у варіації вмісту (у грамах) харчових добавок у двох партіях продукції, а саме випадкова величина X має меншу дисперсію, ніж випадкова величина Y.

$H_1 : D(X) \neq D(Y)$ — існує суттєва розбіжність у варіації вмісту (у грамах) харчових добавок у двох партіях продукції.

Кроки другий та третій. Оскільки $s_Y^2 > s_X^2$, то за статистичний критерій візьмемо випадкову величину:

$$F = \frac{s_Y^2}{s_X^2}$$

і обчислимо спостережуване значення критерію:

$$f^* = \frac{s_Y^2}{s_X^2} = \frac{0{,}5117}{0{,}4174} \approx 1{,}226\,.$$

Кроки четвертий та п'ятий.

а) Для альтернативної гіпотези $H_1 : D(X) < D(Y)$ будуємо правобічну критичну область. Критичну точку знаходимо за таблицею додатка 9, ураховуючи ступені вільності $v_1 = n_Y - 1 = 15 - 1 = 14$ (для більшої виправленої дисперсії $s_б^2 = s_Y^2$), $v_2 = n_X - 1 = 16 - 1 = 15$ (для меншої виправленої дисперсії $s_{\text{м}}^2 = s_X^2$) та рівень значущості $\alpha = 0{,}01$:

$$f_{\text{кр}} = (0{,}01;\ 14;\ 15) = 3{,}57.$$

Будуємо критичну область (рис. 5.23).

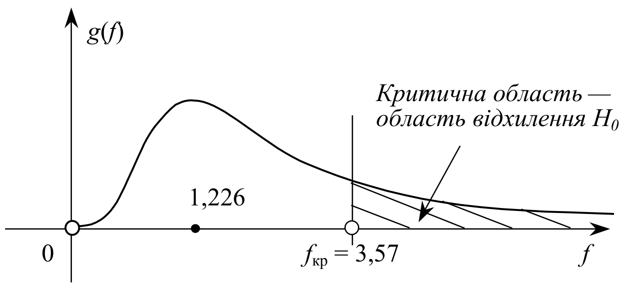


Рис.5.23

Оскільки спостережуване значення f^* не належить критичній області, $\{f \mid f > 3{,}57\}$, то нульова гіпотеза $H_0 : D(X) = D(Y)$ не відхиляється.

б) Згідно з альтернативною гіпотезою $H_1 : D(X) = D(Y)$, критична область — двобічна. Отже, для знаходження критичної точки, беремо рівень значущості вдвічі менший від заданого:

$$\frac{\alpha}{2} = \frac{0{,}1}{2} = 0{,}05 .$$

За таблицею додатка 9 знаходимо $f_{кр} = F(0{,}05;\ 14;\ 15) = 2{,}42$ (рис. 5.24).

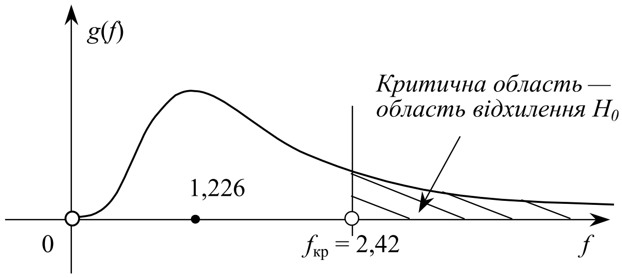


Рис.5.24

Оскільки $f^* = 1{,}226 < f_{кр}$, то основна гіпотеза про рівність генеральних дисперсій не відхиляється.

Висновок. Статистичних даних про суттєву розбіжність варіації вмісту (у грамах) харчових добавок у двох досліджених партіях продукції недостатньо і в п. а) і в п. б).

5.6. ПЕРЕВІРКА СТАТИСТИЧНОЇ ГІПОТЕЗИ ПРО РІВНІСТЬ ВИПРАВЛЕНОЇ ВИБІРКОВОЇ ДИСПЕРСІЇ З ГІПОТЕТИЧНИМ ЗНАЧЕННЯМ ДИСПЕРСІЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ, РОЗПОДІЛЕНОЇ ЗА НОРМАЛЬНИМ ЗАКОНОМ

Гіпотетичне (передбачуване) значення σ_0^2 дисперсії генеральної сукупності σ^2 частіше встановлюється теоретично, або на підставі попередніх досліджень. Для перевірки цього значення із генеральної сукупності обирається вибірка обсягом n, обчислюється її виправлена дисперсія s^2 і, використовуючи значення виправленої дисперсії s^2, з заданим рівнем значущості α перевіряється нульова гіпотеза про рівність дисперсії генеральної сукупності σ^2 встановленому гіпотетичному значенню σ_0^2. Відомо, що виправлена дисперсія s^2 є незміщеною статистичною оцінкою генеральної дисперсії ($M(s^2) = D(X)$), тому нульову гіпотезу можна записати так: H_0: $M(s^2) = \sigma_0^2$. Потрібно перевірити, чи дорівнює математичне сподівання виправленої вибіркової дисперсії гіпотетичному значенню генеральної дисперсії. Перевірка показує, значущою чи ні є різниця між виправленою вибірковою дисперсією та гіпотетичним значенням генеральної дисперсії. Перевірка гіпотез даного типу на практиці необхідна для перевірки точності приладу, методу дослідження тощо. Наприклад, якщо відомо, що σ_0^2 — допустима характеристика відхилення контролюючого розміру деталі, виготовленої верстат-автоматом, а за результатом дослідження вибірки ми отримали значення характеристики розсіювання значно більше від допустимого, то робимо висновок про необхідність налагодження верстат-автомата.

За критерій перевірки нульової гіпотези обираємо випадкову величину $(n-1)\,\frac{s^2}{\sigma_0^2}$. Вона є випадковою, оскільки s^2 набуває випадкових для різних вибірок наперед невідомих значень, і за умови достовірності нульової гіпотези має розподіл χ^2 з $(n-1)$ ступенями вільності, отже:

$$\chi^2 = (n-1)\,\frac{s^2}{\sigma_0^2}. \tag{5.32}$$

Критична область будується залежно від вигляду альтернативної гіпотези.

Випадок перший.

Для нульової гіпотези $H_0 : \sigma^2 = \sigma_0^2$ та альтернативної гіпотези $H_1 : \sigma^2 = \sigma_0^2$ будується правостороння критична область. Для побудови цієї області враховується умова $P(\chi^2 > \chi^2_{\text{кр}}(\alpha, \nu))$ якщо вважати, що нульова гіпотеза справедлива.

Критичну точку $\chi^2_{\text{пр.кр.}}(\alpha, \nu)$ знаходять із таблиці додатка 7 за заданим рівнем значущості α та числом ступенів вільності $\nu = n - 1$. Якщо позначити через $\chi^2_{\text{спост.}}$ значення статистичного критерію, визначеного за результатами вибірки, то висновок за результатом перевірки нульової гіпотези буде таким:

1. Якщо $\chi^2_{\text{спост.}}$ належить критичній області, тобто $\chi^2_{\text{спост.}} > \chi^2_{\text{кр.}}(\alpha, \nu)$, робиться висновок про відхилення нульової гіпотези на користь альтернативної або про достатню кількість статистичних фактів для підтримки альтернативної гіпотези.
2. Якщо $\chi^2_{\text{спост.}}$ належить області прийняття нульової гіпотези, тобто $\chi^2_{\text{спост.}} < \chi^2_{\text{кр.}}(\alpha, \nu)$ робиться висновок про невідхилення нульової гіпотези на користь альтернативної або про недостатність статистичних фактів для підтримки альтернативної гіпотези (рис. 5.25).

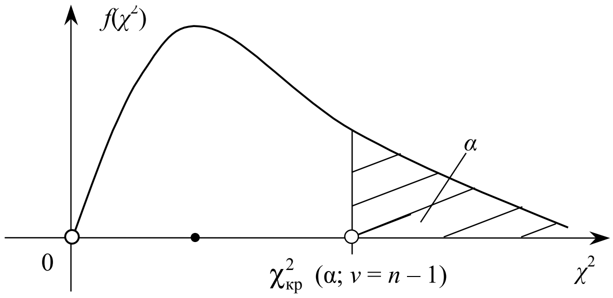


Рис. 5.25

Випадок другий.

Для нульової гіпотези $H_0 : \sigma^2 = \sigma_0^2$ та альтернативної гіпотези $H_1 : \sigma^2 \neq \sigma_0^2$ будується двостороння критична область. Для побудови області враховується те, що ймовірність потрапляння спостережуваного значення критерію до критичної області дорівнює α за умови справедливості нульової гіпотези. Зауважимо, що найбільша потужність критерію досягається тоді, коли ймовірність потрапляння спостережуваного значення критерію до лівої та правої частин двосторонньої критичної області дорівнює $\frac{\alpha}{2}$. Критичні α точки знаходять із таблиці додатку 7 за заданим рівнем значущості α та числом ступенів вільності $\nu = n - 1$ за таким правилом:

$$P\left(\chi^2 < \chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, \nu)\right) = \frac{\alpha}{2},$$

$$P\left(\chi^2 > \chi^2_{\text{прав.кр.}}(\frac{\alpha}{2}, \nu)\right) = \frac{\alpha}{2}.$$

У таблиці додатка 7 указані лише правосторонні межі критичної області, назвемо їх «праві» критичні точки. Для знаходження лівосторонньої межі двосторонньої критичної області, або «лівої» критичної точки, зауважимо, що події $\chi^2 < \chi^2_{\text{лів.кр.}}$ та $\chi^2 > \chi^2_{\text{лів.кр.}}$ є несумісними протилежними випадковими подіями, отже, сума ймовірностей цих подій дорівнює одиниці:

$$P\left(\chi^2 < \chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, \nu)\right) + P\left(\chi^2 > \chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, \nu)\right) = 1,$$

звідки $\mathrm{P}\left(\chi^2 > \chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, \nu)\right) = 1 - P\left(\chi^2 < \chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, \nu)\right) = 1 - \frac{\alpha}{2}$.

Отже, лівосторонню межу двосторонньої критичної області можна знайти за таблицею додатка 7, як правосторонню за умови, що ймовірність потрапляння спостережуваного значення критерію до інтервалу, який знаходиться правіше «лівої» критичної точки, дорівнює $1 - \frac{\alpha}{2}$, тобто:

$$\chi^2_{\text{лів.кр.}}(\frac{\alpha}{2}, v) = \chi^2_{\text{прав.кр.}}(1 - \frac{\alpha}{2}, v).$$

Висновок за результатом перевірки нульової гіпотези буде таким:

якщо $\chi^2_{\text{спост}}$ належить до критичної області, тобто: $\chi^2_{\text{спост}} < \chi^2_{\text{лів.кр.}}$, або $\chi^2_{\text{спост}} > \chi^2_{\text{прав.кр.}}$, робиться висновок про відхилення нульової гіпотези на користь альтернативної, або про достатню кількість статистичних фактів для підтримки альтернативної гіпотези.

Якщо $\chi^2_{\text{спост}}$ належить області прийняття нульової гіпотези (рис. 5.26), тобто: $\chi^2_{\text{лів.кр.}} < \chi^2_{\text{спост}} < \chi^2_{\text{прав.кр.}}$, робиться висновок про невідхилення нульової гіпотези на користь альтернативної, або про недостатність статистичних фактів для підтримки альтернативної гіпотези.

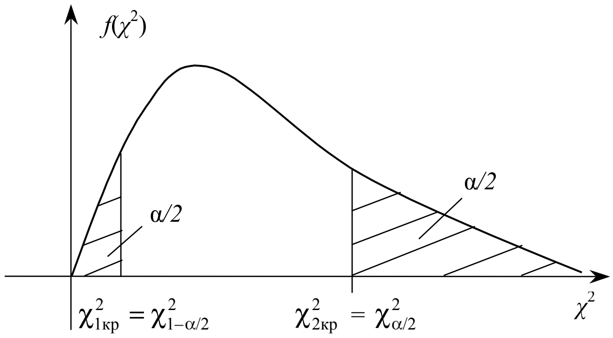


Рис. 5.26

Випадок третій.

Для нульової гіпотези $H_0 : \sigma^2 = \sigma_0^2$ та альтернативної гіпотези $H_1 : \sigma^2 < \sigma_0^2$ будується лівостороння критична область. Для її побудови, знаходиться межа критичної області — «ліва» критична точка. Щоб знайти цю точку, ураховуємо умову $P(\chi^2 < \chi^2_{\text{лів.кр.}}(\alpha, \nu)) = \alpha$, якщо вважати, що нульова гіпотеза справедлива.

Критичну точка $\chi^2_{\text{лів.кр.}}$ знаходять із таблиці додатка 7 за заданим рівнем значущості α та числом ступенів вільності $v = n - 1$, як «праву» критичну точку (див. пояснення до випадку другого):

$$\chi^2_{\text{лів.кр.}} = \chi^2_{\text{прав.кр.}}(1 - \alpha, v),$$

Якщо позначити через $\chi^2_{\text{спост}}$ значення статистичного критерію, визначеного за результатами вибірки, то висновок за результатом перевірки нульової гіпотези буде таким:

1. Якщо $\chi^2_{\text{спост}}$ належить критичній області, тобто $\chi^2_{\text{спост}} < \chi^2_{\text{лів.кр.}}$, робиться висновок про відхилення нульової гіпотези на користь альтернативної, або про достатню кількість статистичних фактів для підтримки альтернативної гіпотези.

2. Якщо $\chi^2_{\text{спост}}$ належить області прийняття нульової гіпотези, тобто $\chi^2_{\text{спост}} > \chi^2_{\text{лів.кр.}}$ (рис. 5.27) робиться висновок про невідхилення нульової гіпотези на користь альтернативної, або про недостатність статистичних фактів для підтримки альтернативної гіпотези.

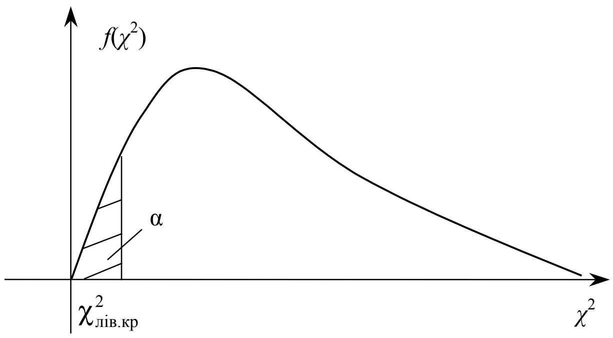


Рис. 5.27

Зауваження. Якщо за умовою задачі відома вибіркова дисперсія, то за статистичний критерій перевірки нульової гіпотези обираємо випадкову величину:

$$\chi^2 = \frac{nD_{\text{в}}}{\sigma_0^2}. \qquad (5.33)$$

Зауваження. Якщо кількість ступенів вільності більша за тридцять, то критичну точку можна знайти наближено, використовуючи нерівність Уілсона-Гілферті.

Приклад 5.19. Експерти з дорожнього руху впевнені, що головною причиною ДТП на трасі є різна швидкість автомобілів. Тимчасом, коли деякі машини рухаються повільно, швидкість інших набуває максимально дозволеної. Припустимо, що експерт вважає, що коли відхилення швидкостей перевищує 20 км/год, число ДТП стає неприпустимо великим. На ділянці траси високої аварійності була зроблена випадкова вибірка швидкостей 30 автомобілів та визначено $s^2 = 22{,}5^2$. Чи можна з рівнем значущості 10 % стверджувати, що відхилення у швидкостях машин перевищує 20 км/год?

Перевірити гіпотезу $H_0 : \sigma^2 = 20^2$ за альтернативної гіпотези:

а) $H_0 : \sigma^2 > 20^2$,
б) $H_1 : \sigma^2 \neq 20^2$.

Швидкість руху машин вважається розподіленою за нормальним законом.

Р о з в ' я з а н н я.

а) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \sigma^2 = 20^2$ — середне квадратичне відхилення у швидкостях машин дорівнює 20 км/год.

$H_1 : \sigma^2 > 20^2$ — середне квадратичне відхилення у швидкостях машин перевищує 20 км/год.

Крок другий. За статистичний критерій обирається випадкова величина $\chi^2 = (n-1)\dfrac{s^2}{\sigma_0^2}$, яка за умови справедливості нульової гіпотези має розподіл χ^2 із $v = n - 1 = 29$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення $\chi^2_{\text{спост}}$ критерію χ^2:

$$\chi^2_{\text{спост}} \ \frac{29 \cdot 22{,}5^2}{20^2} \approx 36{,}7.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \sigma^2 \neq 20^2$, для рівня значущості 10 % і кількості ступенів вільності 29 за таблицею додатока 7 знаходимо межу правосторонньої критичної області— критичну точку:

$$\chi^2_{\text{спост.кр.}}(\alpha = 0{,}1,\ v = 29) = 39{,}087.$$

Крок п'ятий. Оскільки спостережуване значення критерію $\chi^2_{\text{спост}} = 36{,}7$ не належить області відхилення нульової гіпотези критерію $\{\chi^2 | \chi^2 > 39{,}087\}$, робимо висновок про недостатність статистичних фактів з рівнем значущості 10 % на користь альтернативної гіпотези для висновку про те, що відхилення у швидкостях машин більше, ніж 20 км/год.

Розглянемо розв'язок цієї самої задачі, якщо альтернативна гіпотеза $H_1 : \sigma^2 \neq 20^2$.

Р о з в ' я з а н н я.

б) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \sigma^2 = 20^2$ — середнє квадратичне відхилення у швидкостях машин дорівнює 20 км/год.

$H_1 : \sigma^2 \neq 20^2$ — середнє квадратичне відхилення у швидкостях машин не дорівнює 20 км/год.

Крок другий. За статистичний критерій обирається випадкова величина $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}(n-1)$, яка за умови справедливості гіпотези H_0 має розподіл χ^2 із $v = n - 1 = 30 - 1 = 29$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення $\chi^2_{\text{спост}}$ критерію χ^2:

$$\chi^2_{\text{спост.}} = \frac{29 \cdot 22{,}5^2}{20^2} = 36{,}7.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1:\ \sigma^2 \neq 20^2$ для рівня значущості $\frac{10\ \%}{2} =$ $= 5$ % та кількості ступенів вільності 29 за таблицею додатка 7 знаходимо межі двосторонньої критичної області — критичні точки:

$$\chi^2_{\text{пр.кр.}}\ (\alpha, v) = \chi^2\ (\frac{\alpha}{2} = 0{,}05;\ v = 29) = 42{,}557,$$

$$\chi^2_{\text{лів.кр.}}\ (\frac{\alpha}{2}, v) = \chi^2_{\text{прав.кр.}}\ (1 - \frac{\alpha}{2}, v) = \chi^2_{\text{прав.кр.}}\ (1 - \frac{\alpha}{2} = 0{,}95;\ v = 29) = 17{,}708.$$

Крок п'ятий. Оскільки спостережуване значення критерію $\chi^2_{\text{спост}} = 36{,}7$ не належить області відхилення нульової гіпотези критерію

$\{\chi^2\,|\,\chi^2 < 17{,}708$ або $\chi^2 > 42{,}557\}$, робимо висновок про недостатність статистичних фактів з рівнем значущості 10 % на користь альтернативної гіпотези для висновку про те, що відхилення у швидкостях машин більше, ніж 20 км/год.

Приклад 5.20. На першій технологічній лінії зроблено вдосконалення, котре сприяло скороченню варіації часу виготовлення продукції. Вибіркова перевірка варіації часу виготовлення продукції на першій лінії показала значення $s^2 = 0{,}8$ на основі 13 спостережень, за умови існуючої допустимої варіації $\sigma_0^2 = 1{,}05$.

Чи можна вважати суттєвою отриману розбіжність між варіацією тривалості процесу виготовлення продукції на першій лінії з існуючою допустимою варіацією для подальшого вдосконалення інших технологічних ліній? Час виготовлення продукції розподілений за нормальним законом.

Р о з в ' я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0:\ \sigma^2 < 1{,}05$ — немає необхідності починати реконструкцію на інших лініях;

$\text{H}_1:\ \sigma^2 < 1{,}05$ — варто починати реконструкцію на інших лініях.

Крок другий. За статистичний критерій обирається випадкова величина $\chi^2 = (n - 1)\frac{s^2}{\sigma_0^2}$, яка за умови справедливості нульової гіпотези має розподіл χ^2 із $v = n - 1 = 13 - 1 = 12$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення $\chi^2_{\text{спост}}$ критерію χ^2:

$$\chi^2_{\text{спост}} = \frac{12 \cdot 0{,}8}{1{,}05} \approx 9{,}14.$$

Крок четвертий. За умовою задачі не задано рівень значущості, тож обираємо його самостійно, наприклад 5 %. За змістом альтернативної гіпотези $H_1:\ \sigma^2 < 1{,}05$, для рівня значущості 5 % та кількості ступенів вільності 12 за таблицею додатку 7 знаходимо межу лівосторонньої критичної області — критичну точку:

$$\chi^2_{\text{лів.кр.}} \left(\frac{\alpha}{2}, v\right) = \chi^2_{\text{прав.кр.}}\ (1 - \alpha, v) = \chi^2_{\text{прав.кр.}}\ (1 - \alpha = 0{,}95;\ v = 12) = 5{,}226.$$

Крок п'ятий. Оскільки спостережуване значення критерію $\chi^2_{\text{спост}} = 9{,}14$ належить області відхилення нульової гіпотези критерію $\{\chi^2\,|\,\chi^2 < 5{,}226\}$, робимо висновок про недостатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези для висновку щодо необхідності початку реконструкції на інших лініях.

Приклад 5.21. Компанія К досліджує можливості отримання доходу від інвестиції А. Інвестиція А передбачається на строк 10 років з очікуваним прибутком 17,8 % річних. Вибіркова дисперсія щорічних прибутків від інвестиції А становить — $(7.14\ \%)^2$. Але інформація, одержана фінансовим аналітиком, дає підставу стверджувати, що це помилка вибірки і характеристика ризику не перевищує $(5{,}3\ \%)^2$. Чи варто довіряти одержаній інформації? Зробіть перевірку одержаної інформації з рівнем значущості 5 % двома способами:

а) $H_0:\ \sigma_x^2 \le 5{,}3^2$, $H_1:\ \sigma_x^2 > 5{,}3^2$.

б) $H_0:\ \sigma_x^2 = 5{,}3^2$, $H_1:\ \sigma_x^2 \neq 5{,}3^2$.

Передбачається, що випадкова величина щорічного прибутку на інвестицію А має нормальний закон розподілу.

Р о з в ' я з а н н я.

а) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \sigma_x^2 = 5{,}3^2$ — дисперсія ризику не перевищує (5,3 %)^2 для інвестиції A.

$H_1 : \sigma_x^2 > 5{,}3^2$ — дисперсія ризику перевищує (5,3 %)^2 для інвестиції A.

Крок другий. За статистичний критерій обирається випадкова величина $\chi^2 = \frac{nD_в}{\sigma_0^2}$, яка за умови справедливості гіпотези H_0 має розподіл χ^2 із $v = n - 1 = 10 - 1 = 9$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення $\chi^2_{спост}$ критерію χ^2:

$$\chi^2_{спост} = \frac{10 \cdot 7{,}14^2}{5{,}3^2} = 18{,}149.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \sigma_x^2 > 5{,}3^2$ для рівня значущості 5 % і кількості ступенів вільності 9 за таблицею додатка 7 знаходимо межу правосторонньої критичної області — критичну точку:

$$\chi^2_{пр.кр.}(\alpha, v) = \chi^2(\alpha = 0{,}05,\ v = 9) = 16{,}919.$$

Крок п'ятий. Оскільки спостережуване значення критерію $\chi^2_{спост} = 18{,}149$ належить області відхилення нульової гіпотези $\{\chi^2 | \chi^2 > 16{,}919\}$, робимо висновок про достатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези для висновку про те, що характеристика ризику інвестиції А перевищує (5,3 %)^2 .

Розглянемо розв'язок цієї самої задачі, але з альтернативною гіпотезою:

$$H_1 : \sigma_x^2 \neq 5{,}3^2.$$

б) Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \sigma_x^2 = 5{,}3^2$ — характеристика ризику для інвестиції A дорівнює (5,3 %)^2.

H_1: $\sigma_x^2 \neq 5{,}3^2$ — характеристика ризику для інвестиції A відрізняється від (5,3 %)^2.

Крок другий. За статистичний критерій обирається випадкова величина $\chi^2 = \frac{nD_в}{\sigma_0^2}$,

яка у випадку справедливості гіпотези H_0 має розподіл χ^2 із $v = n - 1 = 10 - 1 = 9$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення $\chi^2_{спост.}$ критерію χ^2:

$$\chi^2_{спост} = \frac{10 \cdot 7{,}14^2}{5{,}3^2} = 18{,}149.$$

Крок четвертий. За змістом альтернативної гіпотези H_1: $\sigma_x^2 \neq 5{,}3^2$, для рівня значущості $\frac{5\,\%}{2} = 2{,}5$ (%), та кількості ступенів вільності $v = 9$ за таблицею додатка 7 знаходимо межі двосторонньої критичної області — критичні точки:

$$\chi^2_{пр.кр.}(\alpha, v) = \chi^2\left(\frac{\alpha}{2} = 0{,}025;\ v = 9\right) = 19{,}023.$$

$$\chi^2_{лів.кр.}\left(\frac{\alpha}{2}, v\right) = \chi^2_{прав.кр.}\left(1 - \frac{\alpha}{2}, v\right) = \chi^2_{прав.кр.}\left(1 - \frac{\alpha}{2} = 0{,}957;\ v = 9\right) = 2{,}7.$$

Крок п'ятий. Оскільки спостережуване значення критерію $\chi^2_{спост} = 18{,}149$ не належить області відхилення нульової гіпотези критерію $\{\chi^2 | \chi^2 < 2{,}7$ або $\chi^2 > 19{,}023\}$, робимо висновок про недостатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези для висновку про те, що характеристика ризику для інвестиції A не дорівнює (5,3 %)^2.

5.7. ПЕРЕВІРКА ГІПОТЕЗИ ПРО РІВНІСТЬ СЕРЕДНІХ ДВОХ ГЕНЕРАЛЬНИХ СУКУПНОСТЕЙ (НЕЗАЛЕЖНІ ВИБІРКИ)

Порівняння середніх двох сукупностей має важливе практичне значення. Часто середні результати однієї серії експериментів відрізняються від іншої. Виникає питання, чи можливо пояснити різницю середніх випадковими помилками експерименту, чи ця різниця викликана деякими закономірностями? У промисловості завдання порівняння середніх частот виникає у разі вибіркового контролю якості однакових виробів, що виготовлені на різних підприємствах, або на одному підприємстві за різних технологічних режимів, у фінансовому аналізі — за порівняння рівня дохідності різних активів тощо.

Випадок перший. Дисперсії генеральних сукупностей відомі.

Нехай генеральні сукупності, ознаки яких X_1 і X_2 розподілені нормально, є незалежними одна від одної, а саме $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$. Середні значення генеральних сукупностей (їх математичні сподівання) дорівнюють $M(X_1) = \mu_1$, $M(X_2) = \mu_2$ і нам невідомі. Дисперсії генеральних сукупностей $D(X_1) = \sigma_1^2$ та $D(X_2) = \sigma_2^2$ є відомими (наприклад, із попереднього досвіду або знайдені теоретично). За результатами незалежних вибірок обсягами n_1 та n_2 відповідно, здійснених із цих генеральних сукупностей, знайдено вибіркові середні арифметичні $\bar{x}_1$ та $\bar{x}_2$. Вибіркові середні $\overline{X}_1$ та $\overline{X}_2$ є незалежними випадковими величинами, оскільки для різних вибірок з генеральних сукупностей вони набувають незалежних різних значень. Нагадаємо, що закон розподілу середніх вибіркових величин $\overline{X}_1$ та $\overline{X}_2$ — нормальний, а саме: $X_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $X_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$.

Потрібно за обчисленими вибірковими середніми $\bar{x}_1$ та , $\bar{x}_2$ із заданим рівнем значущості α перевірити достовірність нульової гіпотези про рівність генеральних середніх цих сукупностей, тобто H_0: $M(X_1) = M(X_2)$.

Оскільки вибіркові середні є незміщеними оцінками генеральних середніх, тобто $M(\overline{X}_1) = M(X_1)$, $M(\overline{X}_2) = M(X_2)$, нульову гіпотезу можна записати як рівність математичних сподівань вибіркових середніх, які, за зауваженим вище, для різних вибірок приймають різні значення. Як зрозуміти, значущою чи ні є різниця між ними?

Якщо ми робимо висновок про невідхилення нульової гіпотези з заданим рівнем значущості, то різниця між вибірковими середніми вважається несуттєвою і пояснюється випадковими причинами, наприклад випадковим відбором об’єктів вибірки.

Якщо ж нульова гіпотеза відхиляється на користь альтернативної, робимо висновок про значущість різниці між вибірковими середніми (її не можна пояснити випадковими причинами), тобто про дійсно існуючі різні значення генеральних середніх. Якщо нульова гіпотеза H_0 справедлива, то різниця середніх значень $\overline{X}_1 - \overline{X}_2$ також має нормальний закон розподілу, причому:

$$M(\overline{X}_1 - \overline{X}_2) = M(\overline{X}_1) - M(\overline{X}_2) = \mu_1 - \mu_2 = 0, \text{ а } D(\overline{X}_1 - \overline{X}_2) = D(\overline{X}_1) + D(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

За статистичний критерій обираємо випадкову величину Z, яка за умови справедливості гіпотези H_0 має нормований нормальний розподіл:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - M(\overline{X}_1 - \overline{X}_2)}{\sigma(\overline{X}_1 - \overline{X}_2)} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{5.34}$$

$M(Z) = 0$, $\sigma(Z) = 1$. Залежно від формулювання альтернативної гіпотези H_1 будуються правостороння, двостороння чи лівостороння критичні області. Критичні точки знаходять за таблицею значень функції Лапласа при заданого рівня значущості.

Зауваження 1. Якщо закон розподілу генеральних сукупностей X_1 та X_2 невідомий, але обсяги вибірок, здійснених із цих генеральних сукупностей, великі ($n_1 > 30$, $n_2 > 30$), то вважаємо, що вибіркові середні $\overline{X}_1$ та $\overline{X}_2$ мають наближено нормальний закон розподілу, відповідно $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$. Тому для перевірки H_0 про рівність середніх генеральних сукупностей застосовують також критерій (5.34).

Зауваження 2. Критерій (5.34) застосовується для перевірки H_0 про рівність середніх генеральних сукупностей у разі здійснення незалежних вибірок малого розміру (обсяг вибірки менше 30) лише при виконанні умов:

1. Генеральні сукупності X_1 та X_2 незалежні та розподілені за нормальним законом;
2. Генеральні дисперсії $D(X_1) = \sigma_1^2$ та $D(X_2) = \sigma_2^2$ — відомі.

Приклад 5.22. Для перевірки ефективності нової технології відібрано дві групи робітників: у першій групі кількістю $n_1 = 50$ осіб, де було застосовано нову технологію, вибіркова середня склала $\bar{x}_1 = 85$ одиниць продукції за годину; у другій групі кількістю $n_2 = 60$ осіб вибіркова середня склала $\bar{x}_2 = 79$ одиниць продукції за годину. Із попереднього досвіду відомі дисперсії продуктивності робітників обох груп, що становлять відповідно $\sigma_1^2 = 100$, $\sigma_2^2 = 81$. З рівнем значущості 5 % визначити, чи впливає нова технологія на середню продуктивність праці.

Р о з в ' я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_1 = \mu_2$ — впливу нової технології на середню продуктивність праці немає;

$H_1 : \mu_1 \neq \mu_2$ — існує вплив нової технології на середню продуктивність праці.

Крок другий. За статистичний критерій обирається випадкова величина:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}.$$

Крок третій. Знаходимо спостережуване значення z^* критерію Z:

$$z^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{85 - 79}{\sqrt{\dfrac{100}{50} + \dfrac{81}{60}}} \approx 3{,}28.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \mu_1 \neq \mu_2$, для рівня значущості 5 % знаходимо межі двосторонньої критичної області (рис. 5.28) — критичні точки: $\Phi(z_{кр}) = \dfrac{1-\alpha}{2} =$ $= 0{,}475$; звідки $z_{кр} = 1{,}96$. Область невідхилення нульової гіпотези $z \in (-1{,}96;\ 1{,}96)$.

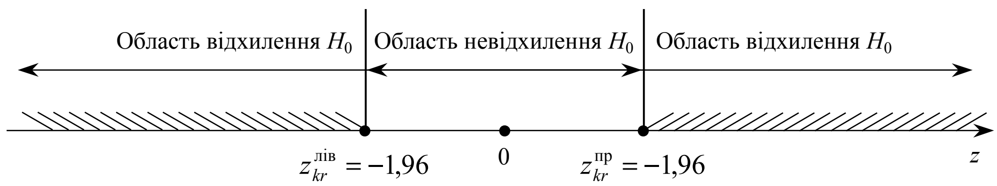


Рис. 5.28

Крок п'ятий. Оскільки спостережуване значення критерію $z^* \approx 3{,}28$ знаходиться в області відхилення нульової гіпотези критерію Z, робимо висновок про достатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези для висновку щодо існування впливу нової технології на середню продуктивність праці.

Приклад 5.23. Для перевірки ефективності нової технології виробництва відібрали дві групи робітників: перша працювала за новою технологією, друга — за старою. Результати виробітку продукції на одного робітника подано в наступних таблицях розподілів:

1-ша група:	x_i	72	77	81	84	88
	n_{x_i}	5	7	10	13	5

2-га група:	y_j	64	66	69	72	75
	n_{y_j}	2	6	13	5	3

Ознаки і (виробіток продукції на одного робітника) — незалежні, нормально розподілені випадкові величини відповідно в 1-й та 2-ій групах із значеннями $D(X) = \sigma_x^2 = 81$ та $D(Y) = \sigma_y^2 = 49$. З рівнем значущості $\alpha = 0{,}01$ перевірити:

а) гіпотезу H_0 $M(X) = M(Y)$,
якщо альтернативна гіпотеза $H_1 : M(X) > M(Y)$;

б) $H_0 : M(X) = M(Y)$, якщо $H_1 : M(X) \neq M(Y)$.

Р о з в ' я з а н н я.
Обчислюємо за таблицями розподілів:

$$n_x = 5 + 7 + 10 + 13 + 5 = 40, \qquad n_y = 2 + 6 + 13 + 5 + 3 = 29,$$

$$\bar{x}_в = \frac{1}{40}(72\cdot5 + 77\cdot7 + 81\cdot10 + 84\cdot13 + 88\cdot5) = \frac{1}{40}(360 + 539 + 810 + 1092 + 440) = 81{,}025;$$

$$\bar{y}_в = \frac{1}{29}(64\cdot2 + 66\cdot6 + 69\cdot13 + 72\cdot5 + 75\cdot3) = \frac{1}{29}(128+396+897+360+225) \approx 69{,}17\,.$$

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:
а) H_0 : $M(X) = M(Y)$, якщо альтернативна гіпотеза H_1 : $M(X) > M(Y)$;
б) H_0 : $M(X) = M(Y)$, якщо H_1 : $M(X) \neq M(Y)$;

Крок другий. Використовуємо критерій (5.34), оскільки X та Y мають нормальний розподіл і їхні дисперсії відомі:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\,.$$

Крок третій. Знаходимо спостережуване значення z^* критерію Z:

$$z^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_в - \bar{y}_в}{\sqrt{\frac{D(X)}{n_1} + \frac{D(Y)}{n_2}}} = \frac{81{,}025 - 69{,}17}{\sqrt{\frac{81}{40} + \frac{49}{29}}} \approx \frac{11{,}855}{1{,}92734} \approx 6{,}15\,.$$

Крок четвертий. Для рівня значущості 1 % знаходимо критичні точки — межі критичних областей — та будуємо критичні області.

а) За змістом альтернативної гіпотези H_1 : $M(X) > M(Y)$ будуємо правосторонню критичну область. Критичні точки (рис. 5.29) знаходимо з рівняння $\Phi(Z_{кр}) = \frac{1}{2} - \alpha = \frac{1}{2} - 0{,}01 = 0{,}49$. Із таблиці значень функції Лапласа знаходимо критичну точку $z_{кр} = 2{,}33$.

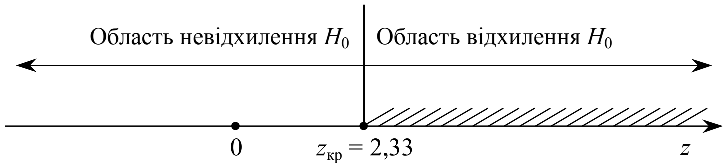


Рис. 5.29

б) За змістом альтернативної гіпотези H_1 : $M(X) \neq M(Y)$ будуємо двосторонню критичну область. Критичні точки (рис. 5.30) знаходимо з рівняння:

$$\Phi(Z_{кр}) = \frac{1-\alpha}{2} = \frac{0{,}99}{2} = 0{,}495, \text{ звідки } z_{кр} = 2{,}58.$$

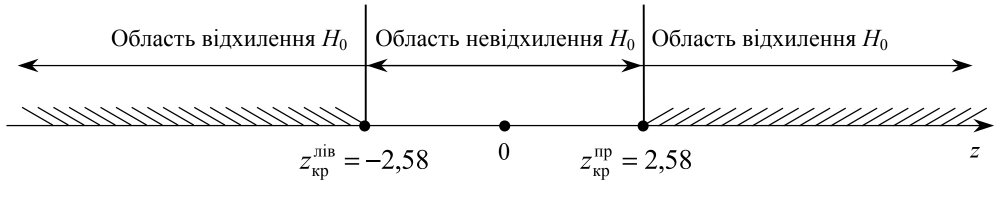


Рис. 5.30

Крок п'ятий. Оскільки спостережуване значення критерію $z^* \approx 6{,}15$ для обох випадків знаходиться в області відхилення основної гіпотези H_0, то робимо висновок про достатність статистичних доказів на користь альтернативної гіпотези H_1, отже, нова технологія привела до підвищення середнього виробітку робітників.

Випадок другий. Дисперсії генеральних сукупностей невідомі, але однакові.

Нехай ознаки двох генеральних сукупностей X_1 і X_2 є незалежними та розподілені нормально, а саме $X_1 \sim N(\mu_1,\ \sigma_1^2)$, $X_2 \sim N(\mu_2,\ \sigma_2^2)$.

Середні значення генеральних сукупностей (їх математичні сподівання) дорівнюють $M(X_1) = \mu_1$, $M(X_2) = \mu_2$, і нам невідомі.

Дисперсії генеральних сукупностей $D(X_1) = \sigma_1^2$ та $D(X_2) = \sigma_2^2$ також невідомі. Наприклад, за вибірками малого обсягу не можна отримати якісні оцінки генеральних дисперсій. Тому метод порівняння середніх, запропонований вище, застосовувати не можна. Але якщо додатково, за невідомих дисперсій, можна зробити припущення про рівність невідомих дисперсій $\sigma_1^2 = \sigma_2^2$, то можна побудувати критерій Стьюдента порівняння середніх. Наприклад, коли порівнюються середні розміри двох партій деталей, виготовлених на одній технологічній лінії, тоді можна припустити, що дисперсії контрольованих розмірів однакові.

Якщо за умовою задачі припущення про рівність дисперсій генеральних сукупностей зробити не можна, слід обов'язково перевірити можливість цього факту, використовуючи критерій Фішера-Снедекора (п. 5.5) про рівність генеральних дисперсій.

Припустимо, що невідомі дисперсії генеральних сукупностей однакові, тобто $\sigma_1^2 = \sigma_2^2 = \sigma^2$. За невідому величину σ^2 можна вибрати її оцінку — виправлену дисперсію першої або другої вибірки ($s_1^2 = \dfrac{n_1}{n_1 - 1}\cdot\sigma^2_{1виб}$ або $s_2^2 = \dfrac{n_2}{n_2 - 1}\cdot\sigma^2_{2виб}$). Але кращою оцінкою для σ^2 буде дисперсія «змішаної» сукупності обсягу $n_1 + n_2$, яку позначимо через s^2 та будемо знаходити за формулою:

$$s^2 = \frac{n_1\cdot\sigma^2_{1виб} + n_2\cdot\sigma^2_{2виб}}{n_1 + n_2 - 2} = \frac{(n_1 - 10)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}.$$

Звертаємо увагу на те, що кількість ступенів вільності $v = n_1 + n_2 - 2$ на два менша від загальної кількості спостережень $n_1 + n_2$, оскільки два ступеня вільності зникають при визначенні середніх $\bar{x}_1$ та $\bar{x}_2$ за вибірковими даними.

Доведено, що у разі справедливості нульової гіпотези про рівність генеральних середніх статистичний критерій

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - M(\bar{X}_1 - \bar{X}_2)}{\sigma(\bar{X}_1 - \bar{X}_2)} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}, \tag{5.35}$$

$$\text{де } s^2 = \frac{n_1\cdot\sigma^2_{1виб} + n_2\cdot\sigma^2_{2виб}}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}. \tag{5.36}$$

є випадковою величиною, що має розподіл Стьюдента з кількістю ступенів вільності $v = n_1 + n_2 - 2$.

Далі перевірка проводиться за відомою схемою з використанням таблиць розподілу Стьюдента.

Зауваження. Нагадуємо, що за великих обсягів вибірки (більших ніж 30) розподіл Стьюдента асимптотично наближається до нормального, тому критичні точки визначаються за таблицею значень функції Лапласа.

Приклад 5.24. Методист денної форми навчання одержав завдання перевірити, чи є суттєва різниця між балами, отриманими студентами першого курсу попереднього навчального року факультету A на іспитах з навчальних дисциплін «Соціологія» та «Політекономія». Для дослідження було зроблено незалежні вибірки отриманих студентами балів з $n_1 = 42$ екзаменаційних робіт із соціології та $n_2 = 58$ екзаменаційних робіт з політекономії За результатами обробки вибіркових даних отримано середні бали: $\bar{x}_1 = 74$, $\bar{x}_2 = 68$ та відповідні виправлені середні квадратичні відхилення: $s_1 = 8$ і $s_2 = 12$. Вважаючи, що невідомі дисперсії генеральних сукупностей є однакові ($\sigma_1^2 = \sigma_2^2$) зробіть висновок з 10-відсотковим рівнем значущості.

Р о з в ' я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_1 = \mu_2$ — суттєвої різниці між балами немає.

$H_1 : \mu_1 \neq \mu_2$ — існує суттєва різниця між балами.

Крок другий. За статистичний критерій обирається випадкова величина:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}},$$

де $s^2 = \dfrac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}$, що за умови справедливості гіпотези H_0 має розподіл Стьюдента з кількістю ступенів вільності $v = n_1 + n_2 - 2$.

Крок третій. Знаходимо спостережуване значення t^* критерію T, обчислюючи спочатку s^2:

$$s^2 = \frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}\ \frac{41\cdot 8^2+57\cdot 12^2}{41+57} \approx 110{,}53,$$

$$\text{тоді } t^* = \frac{\overline{X}_1-\overline{X}_2}{\sqrt{\dfrac{s^2}{n_1}+\dfrac{s^2}{n_2}}} = \frac{74-68}{\sqrt{\dfrac{110{,}53}{42}+\dfrac{110{,}53}{58}}} \approx 2{,}82.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \mu_1 \neq \mu_2$, для рівня значущості 10 % і кількості ступенів вільності $v = 42 + 58 - 2 = 98$, знаходимо межі двосторонньої критичної області — критичні точки розподілу Стьюдента: $t_{кр} = t(0{,}05; \nu = 98) = 1{,}66$. Отже, область невідхилення нульової гіпотези $t \in (-1{,}66; 1{,}66)$.

Зауважимо, що за великих обсягів вибірки (більших ніж 30), розподіл Стьюдента асимптотично наближається до нормального, тому критичні точки визначаються за таблицею значень функції Лапласа. Отже, критичні точки можна було б знаходити так:

$$\Phi(t_{кр}) = \frac{1-\alpha}{2} = 0{,}45;\ t_{кр} = 1{,}65.$$

Тоді отримаємо область невідхилення нульової гіпотези $t \in (-1{,}65; 1{,}65)$.

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = 2{,}82$ знаходиться в області відхилення нульової гіпотези критерію T, робимо висновок про достатність статистичних фактів з рівнем значущості 10 % на користь альтернативної гіпотези для висновку про суттєвість різниці між балами, отриманими студентами першого курсу попереднього навчального року факультету A на іспитах з навчальних дисциплін «Соціологія» та «Політекономія».

Приклад 5.25. За умовою задачі попереднього прикладу дослідити можливість надання статистичних доказів для таких тверджень:

а) бали, отримані студентами на іспиті із соціології, суттєво вищі за бали, отримані на іспиті з політекономії;

б) середній бал студентів іспиту із соціології вищий за середній бал іспиту з політекономії на 10;

в) середній бал студентів іспиту із соціології вищий за середній бал іспиту з політекономії на 5.

Розв'язання.

Крок перший.

Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

а) $H_0 : \mu_1 - \mu_2 \leq 0$; $\quad H_1 : \mu_1 - \mu_2 > 0$;

б) $H_0 : \mu_1 - \mu_2 = 10$; $\quad H_1 : \mu_1 - \mu_2 \neq 10$;

в) $H_0 : \mu_1 - \mu_2 = 5$; $\quad H_1 : \mu_1 - \mu_2 \neq 5$;

Крок другий повністю повторює попередній приклад для всіх пунктів: а), б), в).

Отже, за статистичний критерій обирається випадкова величина:

$$\mathrm{T} = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{s^2}{n_1}+\dfrac{s^2}{n_2}}}.$$

Крок третій. Знайдемо спостережуване значення t^* критерію T.

Спочатку обчислюємо дисперсію «змішаної» сукупності:

$s^2 = \dfrac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2} = \dfrac{41\cdot 8^2+57\cdot 12^2}{41+57} \approx 110{,}53$ — для всіх пунктів: а), б), в).

Отже,

а) спостережуване значення критерію $t^* = \dfrac{\overline{X}_1-\overline{X}_2}{\sqrt{\dfrac{s^2}{n_1}+\dfrac{s^2}{n_2}}} = \dfrac{44-68}{\sqrt{\dfrac{110\cdot 53}{42}+\dfrac{110\cdot 53}{58}}} \approx 2{,}82;$

б) $t^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \dfrac{(74-68)-10}{\sqrt{\dfrac{110{,}53}{42} + \dfrac{110{,}53}{58}}} \approx -1{,}88;$

в) $t^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \dfrac{(74-68)-5}{\sqrt{\dfrac{110{,}53}{42} + \dfrac{110{,}53}{58}}} \approx 0{,}469.$

Крок четвертий і п'ятий.

а) За змістом альтернативної гіпотези $H_1 : \mu_1 - \mu_2 > 0$, для рівня значущості 10 % знаходимо межу правосторонньої критичної області. Оскільки $\Phi(t_{кр}) = 0{,}4$, тоді за таблицею значень функції Лапласа $t_{кр} = 1{,}285$. Отже, область відхилення нульової гіпотези $\{T \mid T > 1{,}285\}$.

Спостережуване значення критерію $t^* = 2{,}82$ знаходиться в області відхилення нульової гіпотези $t > 1{,}285$. З рівнем значущості 10 % достатньо статистичних доказів для висновку про те, що бали, отримані студентами на іспиті із соціології, суттєво вищі за бали, отримані на іспиті із політекономії.

б) За змістом альтернативної гіпотези $H_1 : \mu_1 - \mu_2 \neq 0$ з рівнем значущості 10 % знаходимо межі двосторонньої критичної області $t_{кр} = t(0{,}05;\ k \approx 100) = 1{,}66$. Отже, область невідхилення нульової гіпотези $t \in (-1{,}66;\ 1{,}66)$. Спостережуване значення критерію $t^* = -1{,}88$ — знаходиться в області відхилення нульової гіпотези $H_0 : \mu_1 - \mu_2 = 10$ на користь альтернативної $H_1 : \mu_1 - \mu_2 \neq 10$. З рівнем значущості 10 % недостатньо статистичних доказів для висновку про те, що середній бал студентів іспиту із соціології вищий за середній бал іспиту з політекономії на 10.

в) За змістом альтернативної гіпотези; $H_1 : \mu_1 - \mu_2 \neq 5$ з рівнем значущості 10 % знаходимо межі двосторонньої критичної області $t_{кр} = t(0{,}05;\ k \approx 100) \approx 1{,}66$. Отже, область невідхилення нульової гіпотези $t \in (-1{,}66;\ 1{,}66)$. Спостережуване значення критерію $t^* = 0{,}469$ — знаходиться в області невідхилення нульової гіпотези $H_0 : \mu_1 - \mu_2 = 5$. З рівнем значущості 10 % достатньо статистичних доказів для висновку про те, що середній бал студентів іспиту із соціології вищий за середній бал іспиту з політекономії на 5.

Приклад 5.26. Менеджер з реклами компанії, що виробляє чіпси, одержав завдання з'ясувати, чи суттєво вплинула нова форма упаковки на збут продукції компанії. Для дослідження ним була організована випадкова вибірка 30 однотипних магазинів, 18 з яких продавали чіпси в новій упаковці, а решта — у старій. Через тиждень він одержав такі результати: $\bar{x}_1 = 130$ $s_1 = 12$ коробок за $n_1 = 18$, та $s_2 = 16$ коробок за $n_2 = 12$. Використовуючи рівень значущості 5 % потрібно зробити висновок про суттєвість впливу нової форми упаковки на збут продукції. Вважаємо, що обидві генеральні сукупності розподілені за нормальним законом.

Р о з в ' я з а н н я.

Розв'язання даної задачі містить дві частини. Перша частина — це перевірка гіпотези про рівність дисперсій двох генеральних сукупностей, друга — перевірка гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \sigma_1^2 = \sigma_2^2$ — дисперсії генеральних сукупностей рівні;
$H_1 : \sigma_1^2 \neq \sigma_2^2$ — дисперсії генеральних сукупностей різні.

Крок другий. За статистичний критерій обирається випадкова величина $F = \dfrac{s_б^2}{s_м^2}$, яка за умови справедливості гіпотези H_0 має розподіл Фішера-Снедекора із $v_1 = n_б - 1 = 12 - 1 = 11$ та $v_2 = n_м - 1 = 18 - 1 = 17$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2 = s_2 = 16^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2 = s_1 = 12^2$.

Крок третій. Знаходимо спостережуване значення критерію:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{16^2}{12^2} \approx 1{,}78\ .$$

Крок четвертий. За змістом альтернативної гіпотези H_1: $\sigma_1^2 \neq \sigma_2^2$ знаходимо тільки праву межу двосторонньої критичної області. Із заданим рівнем значущості 5 % та ступенями вільності $v_1 = 11$

$\nu_2 = 17$ за таблицею додатка 9 знаходимо праву межу двосторонньої критичної області — критичну точку $f_{\text{кр}} = F(\alpha = 0{,}025;\ \nu_1 = 11;\ \nu_2 = 17) = 2{,}87$.

Крок п'ятий. Оскільки спостережуване значення критерію $f^{\,*} = 1{,}78$ не знаходиться у правосторонній області відхилення нульової гіпотези $F > 2{,}87$, робимо висновок про недостатність статистичних доказів з рівнем значущості 5 % для відхилення нульової гіпотези про рівність дисперсій генеральних сукупностей $H_0:\ \sigma_1^2 - \sigma_2^2$.

Переходимо до другої частини розв'язання задачі — до перевірки гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_1 = \mu_2$ — нова форма упаковки не впливає на збут продукції.

$H_1 : \mu_1 = \mu_2$ — нова форма упаковки впливає на збут продукції.

Крок другий. За статистичний критерій обирається випадкова величина $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}$, де $s^2 = \dfrac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}$, що за умови справедливості гіпотези H_0 має розподіл Стьюдента з кількістю ступенів вільності $\nu = n_1 + n_2 .- 2$.

Крок третій. Знаходимо спостережуване значення t^* критерію T.

Для обчислення s^2 використовуємо дані умови: $\bar{x}_1 = 130$, $s_1 = 12$, $n_1 = 18$, та $\bar{x}_2 = 117$, $s_2 = 16$, $n_2 = 12$. Тоді:

$$s^2 = \frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2} = \frac{17\cdot 12^2 + 11\cdot 16^2}{18 + 12 - 2} = 188;$$

$$t^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \frac{130 - 117}{\sqrt{\dfrac{188}{18} + \dfrac{188}{12}}} = 2{,}54.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \mu_1 > \mu_2$, з рівнем значущості 5 % і кількістю ступенів вільності $k = 18 + 12 - 2 = 28$, знаходимо межу правосторонньої критичної області — критичну точку розподілу Стьюдента за таблицею додатка 8 : $t_{\text{кр}} = t(0{,}05;\ k = 28) = 1{,}701$. Отже, область відхилення нульової гіпотези $T > 1{,}701$.

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = 2{,}54$ знаходиться в області відхилення нульової гіпотези критерію T, робимо висновок про достатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези для висновку про значущість впливу нової форми упаковки на збут продукції.

Випадок третій. Дисперсії генеральних сукупностей невідомі та різні

Якщо дисперсії генеральних сукупностей невідомі та не можна припустити, що вони однакові, за критерій обирається випадкова величина, яка за умови справедливості гіпотези H_0 має розподіл Стьюдента:

$$\mathrm{T} = \frac{(\overline{X}_1 - \overline{X}_2) - M(\overline{X}_1 - \overline{X}_2)}{\sigma(\overline{X}_1 - \overline{X}_2)} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2}{n_2}}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2}{n_2}}}, \tag{5.37}$$

де $s_1^2 = \dfrac{n_1}{n_1 - 1}\cdot\sigma_1^2$, $s_2^2\ \dfrac{n_2}{n_2 - 1}\cdot\sigma_2^2$ — виправлені вибіркові дисперсії, але відповідна кількість ступенів вільності k обчислюється наближено, за формулою, яку приводимо без доведення:

$$\nu = \frac{\left[\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right]^2}{\dfrac{\left[\dfrac{s_1^2}{n_1}\right]^2}{n_1} + \dfrac{\left[\dfrac{s_2^2}{n_2}\right]^2}{n_2}}. \tag{5.38}$$

Зауваження (стосовно формули 5.38). Не поспішайте обчислювати кількість ступенів вільності за формулою (5.38). Спочатку слід обчислити спостережуване значення критерію за формулою (5.37). Якщо це значення є досить великим числом (наприклад 3 і більше), то нульову гіпотезу можна сміливо відхилити без розрахунків кількості ступенів вільності за формулою (5.38), тобто без знаходження критичних точок. Іншими словами, якщо спостережуване значення критерію знаходиться далеко від центру розподілу, воно завжди буде належати критичній області — області відхилення нульової гіпотези.

Зауваження (до випадків 2 та 3). Нагадуємо, що за великих обсягів вибірки (більших ніж 30) розподіл Стьюдента асимптотично наближається до нормального, тому критичні точки визначаються за таблицею значень функції Лапласа.

Зауваження. Для всіх випадків перевірки гіпотез на порівняння середніх генеральних сукупностей ми можемо побудувати нульову гіпотезу так, що різниця між середніми величинами двох генеральних сукупностей буде деяким числом, відмінним від нуля. Наприклад, $H_0 : \mu_1 - \mu_2 = 10$, $H_0 : \mu_1 - \mu_2 = 10$.

Приклад 5.27. Методист денної форми навчання одержав завдання перевірити, чи є суттєвою різниця між балами, отриманими студентами першого курсу попереднього навчального року факультету А на іспитах з навчальних дисциплін «Соціологія» та «Політекономія». Для дослідження було зроблено незалежні вибірки отриманих студентами балів з $n_1 = 42$ екзаменаційних робіт із соціології та $n_2 = 58$ екзаменаційних робіт з політекономії. За результатами обробки вибіркових даних отримано середні бали: $\dot{x}_1 = 74$, $\dot{x}_2 = 68$ та відповідні виправлені середні квадратичні відхилення: $s_1 = 8$ та $s_2 = 12$. Зробіть висновок стосовно суттєвості різниці між балами, отриманими студентами за результатами вибіркових даних, з рівнем значущості 10 %. Вважається, що обидві генеральні сукупності розподілені за нормальним законом.

Р о з в ’ я з а н н я.

Зауважимо, що даних стосовно рівності дисперсій генеральних сукупностей немає, тому розв’язання цієї задачі складається з двох частин. Перша частина — це перевірка гіпотези про рівність дисперсій двох генеральних сукупностей, друга — перевірка гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

H_0: $\sigma_1^2 = \sigma_2^2$ — дисперсії генеральних сукупностей рівні;

H_1: $\sigma_1^2 \neq \sigma_2^2$ — дисперсії генеральних сукупностей різні.

Крок другий. За статистичний критерій обирається випадкова величина $F = \dfrac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $v_1 = n_б - 1 = 58 - 1 = 57$ та $v_2 = n_м - 1 = 42 - 1 = 41$ ступенями вільності, де $n_б$ — обсяг вибірки, за якою була обчислена більша вибіркова дисперсія $s_б^2 = s_2^2 = 12^2$, а $n_м$ — обсяг вибірки, за якою була обчислена менша вибіркова дисперсія $s_м^2 = s_1^2 = 8^2$.

Крок третій. Знаходимо спостережуване значення критерію:

$$f = \frac{s_б^2}{s_м^2} = \frac{12^2}{8^2} = 2{,}25.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \sigma_1^2 \neq \sigma_2^2$ знаходимо тільки праву межу двосторонньої критичної області. Для заданого рівня значущості 10 % та ступенів вільності $v_1 = 57$, $v_1 = 41$ за таблицею (додатка 9) знаходимо праву межу двосторонньої критичної області — критичну точку:

$$f_{кр} = F(\frac{\alpha}{2} = 0{,}05; v_1 = 57; v_2 = 41) \approx F(\frac{\alpha}{2} = 0{,}05; v_1 = 60; v_2 = 40) \approx 1{,}64.$$

Крок п’ятий. Оскільки спостережуване значення критерію $f^* = 2{,}25$ потрапляє до правосторонньої області відхилення нульової гіпотези $\{F \mid F > 1{,}64\}$, робимо висновок про достатність статистичних доказів з рівнем значущості 5 % для висновків на користь альтернативної гіпотези про різні значення дисперсій генеральних сукупностей.

Переходимо до другої частини розв’язання задачі: до перевірки гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_1 = \mu_2$ — суттєвої різниці між балами немає;

$H_1 : \mu_1 \neq \mu_2$ — існує суттєва різниця між балами.

Крок другий. За критерій обирається випадкова величина, яка за умови справедливості гіпотези H_0 має розподіл Стьюдента:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Відповідна кількість ступенів вільності ν обчислюється наближено, за формулою:

$$\nu \approx \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2}}.$$

Крок третій. Знаходимо спостережуване значення t^* критерію T:

$$t^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{74 - 68}{\sqrt{\frac{8^2}{42} + \frac{12^2}{58}}} = 2{,}9975.$$

Крок четвертий. Знаходимо кількість ступенів вільності:

$$\nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2}} = \frac{\left(\frac{8^2}{42} + \frac{12^2}{58}\right)^2}{\frac{\left(\frac{8^2}{42}\right)^2}{42} + \frac{\left(\frac{12^2}{58}\right)^2}{58}} \approx 99{,}4.$$

За змістом альтернативної гіпотези $H_1 : \mu_1 \neq \mu_2$, з рівнем значущості 10 % і кількістю ступенів вільності $\nu = 100$, знаходимо межі двосторонньої критичної області — критичні точки розподілу Стьюдента за таблицею додатка 8:

$$t_{\text{кр}} = t(0{,}05;\ \nu \approx 100) = 1{,}66.$$

Отже, область невідхилення нульової гіпотези $t \in (-1{,}66;\ 1{,}66)$.

Зауважимо, що за великих обсягів вибірки (більших ніж 30) розподіл Стьюдента асимптотично наближається до нормального, тому критичні точки можна визначити за таблицею значень функції Лапласа, тобто критичні точки можна було б знаходити із рівності $\Phi(t_{\text{кр}}) = \frac{1-\alpha}{2}$; $\Phi(t_{\text{кр}}) = 0{,}45$; звідки $t_{\text{кр}} = 1{,}65$. Дістали область невідхилення нульової гіпотези $t \in (-1{,}65;\ 1{,}65)$.

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = 2{,}9975$ знаходиться в області відхилення нульової гіпотези критерію T, робимо висновок про достатність статистичних фактів з рівнем значущості 10 % на користь альтернативної гіпотези для висновку про істотінсть різниці між балами, отриманими студентами першого курсу попереднього навчального року факультету A на іспитах з навчальних дисциплін «Соціологія» та «Політекономія».

Порада. Завжди доцільніше спочатку обчислити спостережуване значення статистичного критерію. Якщо воно виявилося достатньо великим числом (наприклад, не менше ніж 3), тоді з'являється можливість відхилення нульової гіпотези навіть без обчислення кількості ступенів вільності та критичних значень. Іншими словами, якщо спостережуване значення критерію знаходиться на відстані більше трьох стандартних відхилень від центру розподілу (від математичного сподівання), тоді незалежно від того, де саме буде знаходитися критична точка, спостережуване значення критерію завжди знаходитиметься в критичній області для будь-якого рівня значущості (правило трьох σ).

Приклад 5.28. На підприємствах однієї й тієї самої галузі у двох регіонах України визначали обсяг валової продукції (результати розрахунків подано двома статистичними розподілами):

x_i млн грн	350	400	450	500	530	560
n_{x_i}	15	20	25	30	15	5

y_j млн грн	400	430	460	490	520
n_{y_j}	10	20	35	25	20

Ураховуючи, що ознаки X *і* Y — незалежні й мають нормальний закон розподілу, з рівнем значущості α перевірити гіпотезу $H_0 : M(X) = M(Y)$, якщо альтернативна гіпотеза $H_1 : M(X) \neq M(Y)$.

Р о з в ' я з а н н я.

Оскільки генеральні дисперсії невідомі, спочатку перевіримо гіпотезу про рівність дисперсій: $H_0 : D(X) = D(Y)$, якщо альтернативна гіпотеза $H_1 : D(X) \neq D(Y)$.

Для цього обчислимо n_1, n_2, $\sigma_в(X)$, $\sigma_в(Y)$, $s^2(X)$, $s^2(Y)$:

$$n_X = \sum n_{x_i} = 15 + 20 + 25 + 30 + 15 + 5 = 110;$$

$$n_Y = \sum n_{y_j} = 10 + 20 + 35 + 25 + 20 = 110;$$

$$\bar{x}_в = \frac{1}{110}(350\cdot 15 + 400\cdot 20 + 450\cdot 25 + 500\cdot 30 + 530\cdot 15 + 560\cdot 5) =$$
$$= \frac{1}{110}(5250 + 8000 + 11250 + 15000 + 7950 + 2800) \approx 456{,}82;$$

$$\bar{y}_в = \frac{1}{110}(400\cdot 10 + 430\cdot 20 + 460\cdot 35 + 490\cdot 25 + 520\cdot 20) = \frac{1}{110}(4000 + 8600 + 16\,100 + 12\,250 + 10\,400) \approx 466{,}82;$$

$$\overline{x^2} = \frac{1}{110}(350^2\cdot 15 + 400^2\cdot 20 + 450^2\cdot 25 + 500^2\cdot 30 + 530^2\cdot 15 + 5602\cdot 5 = 11\,101\,837\,500 + 3\,200\,000 + 5\,062\,500 +$$
$$+ 7\,500\,000 + 4\,213\,500 + 1\,568\,000 = 212\,559{,}09;$$

$$\overline{y^2} = \frac{1}{110}(400^2\cdot 10 + 430^2\cdot 20 + 460^2\cdot 35 + 490^2\cdot 25 + 520^2\cdot 20) =$$
$$= \frac{1}{110}(1\,600\,000 + 3\,698\,000 + 7\,406\,000 + 6\,002\,500 + 5\,408\,000) \approx 219\,222{,}72;$$

$$Д_в(X) = \bar{x}^2 - (\bar{x}_В)^2 = 212\,559{,}09 - 208\,684{,}51 = 3\,874{,}58;$$

$$Д_в(Y) = \bar{y}^2 - (\bar{y}_В)^2 = 219\,222{,}799 - 217\,9204{,}91 = 1\,301{,}81;$$

$$\sigma_в(X) \approx 62{,}240;$$

$$\sigma_в(Y) \approx 36{,}081;$$

$$s^2(X) = \frac{n_x}{n_x - 1}\cdot D_в(X) = \frac{110}{109}\cdot 3874{,}58 \approx 3910{,}13;$$

$$s^2(У) = \frac{n_x}{n_x - 1}\cdot D_В(У) = \frac{110}{109}\cdot 3801{,}81 \approx 1313{,}75.$$

Оскільки: $s^2(X) > s^2(Y)$, то $\mathrm{v}_1 = n_X - 1 = 109, v_2 = n_Y - 1 = 109$.

Зауважимо, що даних стосовно рівності дисперсій генеральних сукупностей немає, тому розв'язання даної задачі складається із двох частин. Перша частина — це перевірка гіпотези про рівність дисперсій двох генеральних сукупностей, друга — перевірка гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : D(X) = D(Y)$ — дисперсії генеральних сукупностей рівні;

H_1: $D(X) \neq D(Y)$ — дисперсії генеральних сукупностей різні.

Крок другий. За статистичний критерій обирається випадкова величина $F = \frac{s_б^2}{s_м^2}$, яка має розподіл Фішера-Снедекора із $v_1 = 109$ та $v_2 = 109$ ступенями вільності.

Крок третій. Знаходимо спостережуване значення критерію:

$$f^* = \frac{s_б^2}{s_м^2} = \frac{s^2(Y)}{s^2(X)} = \frac{3910{,}13}{1313{,}75} \approx 2{,}976.$$

Крок четвертий. За змістом альтернативної гіпотези H_1: $D(X) \neq D(Y)$ знаходимо тільки праву межу двосторонньої критичної області. Для рівня значущості 5 % та ступенів вільності $v_1 = 109$, $v_2 = 109$ знаходимо праву межу двосторонньої критичної області — критичну точку

Крок п'ятий. Оскільки спостережуване значення критерію $f^* = 2{,}976$ належить правостороній області відхилення нульової гіпотези $F > 1{,}46$, робимо висновок про достатність статистичних доказів з рівнем значущості 5 % для висновків на користь альтернативної гіпотези про різні значення дисперсій генеральних сукупностей.

Переходимо до другої частини розв'язання задачі — до перевірки гіпотези про рівність середніх значень генеральних сукупностей.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : M(X) = M(Y)$, (або $\mu_1 = \mu_2$) — суттєвої різниці між обсягами валової продукції немає.

$H_1 : M(X) \neq M(Y)$, (або $\mu_1 \neq \mu_2$) — існує суттєва різниця між обсягами валової продукції.

Крок другий. За критерій обирається випадкова величина, яка має розподіл Стьюдента:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Відповідна кількість ступенів вільності k обчислюється наближено, за формулою:

$$\nu = \approx \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2}}.$$

Крок третій. Знаходимо спостережуване значення t^* критерію T:

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_в - \bar{y}_в}{\sqrt{\frac{s^2(X)}{n_1} + \frac{s^2(Y)}{n_2}}} = \frac{456{,}82 - 466{,}82}{\sqrt{\frac{3910{,}13}{110} + \frac{1313{,}75}{110}}} \approx -1{,}45..$$

Крок четвертий

За змістом альтернативної гіпотези $H_1 : M(X) \neq M(Y)$, для рівня значущості 1 %, знаходимо межі двосторонньої критичної області — критичні точки розподілу Стьюдента. Оскільки за великих обсягів вибірки (більших ніж 30) розподіл Стьюдента асимптотично наближається до нормального, то критичні точки можна визначити за таблицею значень функції Лапласа за формулою $\Phi(t_{кр}) = \frac{1-\alpha}{2} = \frac{1-0{,}01}{2} = 0{,}495$, звідки $t_{кр} = 2{,}58$.

Будуємо двосторонню критичну область (рис. 5.32):

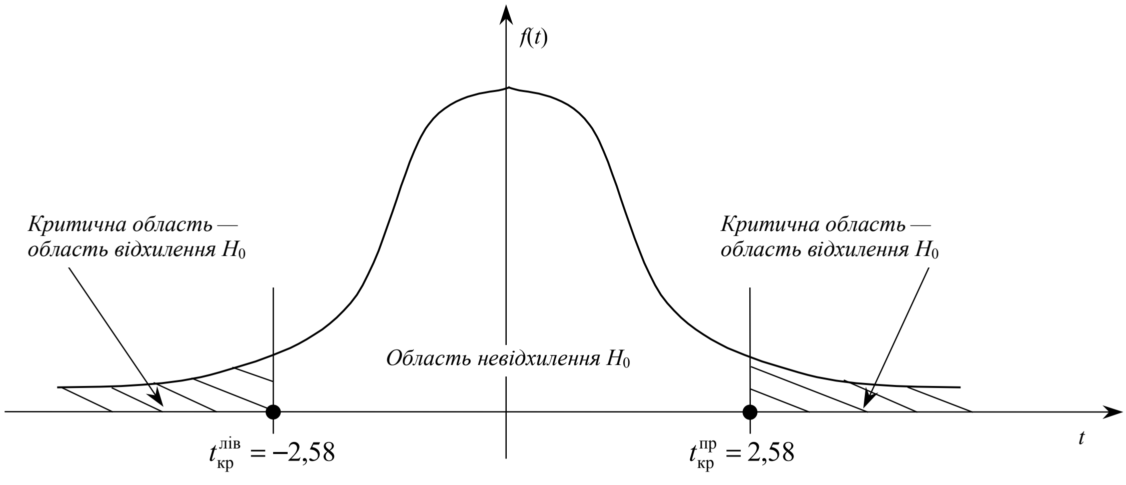


Рис. 5.32

Отже, область невідхилення нульової гіпотези $t \in (-2{,}58;\ 2{,}58)$.

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = -1{,}45$ належить області невідхилення нульової гіпотези критерію T, робимо висновок про недостатність статистичних фактів з рівнем значущості 1 % на користь альтернативної гіпотези для висновку про суттєвість різниці між обсягами валової продукції на підприємствах однієї й тієї самої галузі у двох регіонах України.

Зауваження 9. У разі, коли обсяги вибірок невеликі ($n_1 < 30$ і $n_2 < 30$), вибираючи необхідний (належний) статистичний критерій для перевірки нульової гіпотези H_0 за умови, що генеральні дисперсії невідомі, потрібно спочатку встановити, чи є однаковими генеральні дисперсії.

5.8. ПЕРЕВІРКА ГІПОТЕЗИ ПРО РІВНІСТЬ СЕРЕДНІХ ДВОХ НОРМАЛЬНО РОЗПОДІЛЕНИХ ГЕНЕРАЛЬНИХ СУКУПНОСТЕЙ (ЗАЛЕЖНІ ВИБІРКИ)

У цьому пункті розглядаються задачі порівняння середніх значень генеральних сукупностей, які є залежними одна від одної. Наприклад, ми бажаємо визначити, чи отримали студенти першого курсу попереднього навчального року на заключному іспиті з вищої математики під час літньої екзаменаційної сесії таку саму кількість балів, як і на проміжному іспиті з вищої математики під час зимової екзаменаційної сесії.

Якби в умові отриманого завдання зазначалося, що дослідження були проведені на основі незалежних вибірок, то для перевірки гіпотези про рівність середніх двох сукупностей потрібно було б застосувати один із критеріїв, запропонованих у попередньому пункті 5.7 (формули 5.35—5.38).

Розглянемо інший варіант, коли немає зауваження щодо незалежності вибірок за умовою задачі. У цьому разі можемо припустити, що вибірки не є незалежними, оскільки в середньому бали, отримані студентами на проміжному іспиті з вищої математики під час зимової екзаменаційної сесії, певним чином визначають результати заключного іспиту з вищої математики під час літньої екзаменаційної сесії. Дослідження для випадку залежних вибірок проводиться інакше. Позначимо x_{1i} ($i = 1, 2, 3,\ldots, n$) бали, отримані студентами на проміжному іспиті під час зимової екзаменаційної сесії, а x_{2i} ($i = 1, 2, 3,\ldots, n$) — бали, отримані студентами на заключному іспиті під час літньої екзаменаційної сесії, записані у тому самому порядку, причому обсяги вибірок однакові. Можна передбачити, що для багатьох пар значень можливий результат, коли $x_{1i} \neq x_{2i}$. Виникає потреба встановити, наскільки значущою є різниця значень відповідних пар вибіркових варіант. Дослідження, засноване на даних, наявних у парних спостереженнях, часто називають тестом парних порівнянь.

Ще один приклад застосування цього методу: тестування різниці прибутків двох інвестиційних стратегій у досліджуваному періоді. Спостереження у цьому випадку розглядаємо як залежні з погляду на те, що існує одне спостереження для кожної стратегії в кожному місяці, але кожна пара спостережень залежить від деяких загальних факторів ризику, наприклад таких, як ринкова прибутковість тощо. Отже, вибірки у цьому випадку вже не можна розглядати як незалежні.

Наведемо інший приклад застосування тесту парних порівнянь. Наприклад, маємо справу з дивідендною політикою компанії «до» і «після» змін у податковому законодавстві, що впливають на оподаткування дивідендів. Для перевірки гіпотези про середнє відмінностей, які спостерігаємо по всій компанії з питання оподаткування дивідендів, доцільно використовувати саме тест парних порівнянь значень вибірок, побудованих за даними «до» і « після» змін у податковому законодавстві.

Нехай X_1 та X_2 — дві генеральні сукупності, розподілені за нормальним законом (дисперсії їх на практиці найчастіше невідомі), які характеризують дивідендну політику компанії «до» і «після» змін у податковому законодавстві, що впливають на оподаткування дивідендів.

Потрібно за рівнем значущості α перевірити нульову гіпотезу: $H_0 : M(\overline{X}_1) = M(\overline{X}_2)$ про рівність генеральних середніх з невідомими дисперсіями за альтернативної гіпотези $H_1 : M(\overline{X}_1) \neq M(\overline{X}_2)$ за двома залежними вибірками однакового розміру. Зведемо цю задачу порівняння двох середніх до задачі порівняння однієї вибіркової середньої $\overline{d}$ з гіпотетично відомим значенням генеральної середньої різниць парних спостережень μ_{d_0}, розв'язаної в пункті 5.3 із застосуванням як статистичного критерію — випадкової величини $T = \frac{x_{в} - \mu_0}{s}\sqrt{n}$, що за умови справедливості гіпотези H_0 має розподіл Стьюдента з $v = n - 1$ ступенями вільності, де $s = \sqrt{\frac{n}{n-1}D_{В}}$ — виправлене середньоквадратичне відхилення вибірки.

Для цього розглянемо випадкові величини: різниці $D_i = X_{1i} - X_{2i}$, і $= 1, \overline{n}$ та їх середню $\overline{D} = \frac{\sum D_i}{n} = \frac{\sum(x_{1i} - x_{2i})}{n} = \frac{\sum x_{1i}}{n} - \frac{\sum x_{2i}}{n} = \overline{X}_1 - \overline{X}_2$.

Якщо справедливою є нульова гіпотеза $H_0 : M(\overline{X}_1) = M(\overline{X}_2)$, або $M(\overline{X}_1) = M(\overline{X}_2) = 0$, то $M(\overline{D}) = M(\overline{X}_1 - \overline{X}_2) = M(\overline{X}_1) - M(\overline{X}_2) = 0$. Отже, нульову гіпотезу можна записати, як $H_0 : M(\overline{D}) = 0$ за альтернативної гіпотези $H_1 : M(\overline{D}) \neq 0$.

Відповідні вибірки із генеральних сукупностей X_1 та X_2: x_{1i} ($i = 1,2,3,\ldots, n$) та x_{2i} ($i = 1,2,3,\ldots, n$) — залежні, як сказано вище. Позначимо через $d_i = x_{1i} - x_{2i}$ різниці двох спостережень для кожної із пар. Зауважимо, що потрібно розрізняти випадкові величини $D_i = X_{1i} - X_{2i}$ та $d_i = x_{1i} - x_{2i}$ — невипадкові різниці, які спостерігаємо за результатами вибіркових даних.

Позначимо середню вибіркову:

$$\overline{d} = \sum_{i=1}^{n} \frac{d_i}{n}, \tag{5.39}$$

де $\overline{d}$ — середнє значення різниць спостережень для кожної з пар вибірки,

n — кількість таких пар.

Виправлену вибіркову дисперсію позначимо s_d^2 і знайдемо за формулою:

$$s_d^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1} = \frac{1}{n-1} \cdot \left[\sum_{i=1}^{n} d_i^2 - \frac{\left[\sum_{i=1}^{n} d_i\right]^2}{n}\right]. \tag{5.40}$$

Тоді $\sqrt{s_d^2} = s_d$ — виправлене стандартне відхилення вибірки. Воно дозволяє обчислити стандартне відхилення (середню помилку) середнього значення різниць парних спостережень за формулою:

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}}. \tag{5.41}$$

Нехай μ_d — генеральна середня різниць парних спостережень. Сформулюємо можливі гіпотези стосовно значення μ_d, якщо μ_{d_0} — гіпотетичне значення генеральної середньої різниць парних спостережень:

$H_0 : \mu_d = \mu_{d_0}$, за альтернативної гіпотези $H_1 : \mu_d \neq \mu_{d_0}$,

$H_0 : \mu_d \leq \mu_{d_0}$, за альтернативної гіпотези $H_1 : \mu_d > \mu_{d_0}$,

$H_0 : \mu_d \geq \mu_{d_0}$, за альтернативної гіпотези $H_1 : \mu_d < \mu_{d_0}$.

У розв'язанні практичних задач частіше перевіряється гіпотеза:

$H_0 : \mu_d = 0$, а альтернативною є гіпотеза $H_1 : \mu_d \neq 0$.

Отже, для побудови критерію, підставляємо $\overline{X}_{в} = \overline{d}$, $\mu_0 = \mu_{d_0}$, $s = s_d$ у формулу $T = \frac{X_{в} - \mu_0}{s}\sqrt{n}$. Дістанемо випадкову величину, яка за умови справедливості гіпотези H_0 має розподіл Стьюдента:

$$T = \frac{\overline{d} - \mu_{d_0}}{s_d} \cdot \sqrt{n}. \tag{5.42}$$

Далі перевірка нульової гіпотези проводиться за відомою схемою з використанням таблиць розподілу Стьюдента для даного рівня значущості α та відповідної кількості ступенів вільності $v = n - 1$, де n — кількість пар вибіркових спостережень.

Ще раз зауважимо, що перевірка гіпотези про порівняння середніх двох нормально розподілених генеральних сукупностей у разі залежних вибірок зводиться до дослідження середнього значення різниць $\overline{d}$ між значеннями у парах спостережень, оскільки спостережувані значення двох вибірок попарно взаємозв'язані.

Приклад 5.29. Викладачі кафедри вищої математики бажають перевірити, чи є суттєвою різниця між балами, отриманими студентами першого курсу попереднього навчального року на іспиті з вищої математики під час зимової та літньої екзаменаційних сесій.

Р о з в ' я з а н н я.

За умовою задачі немає зауваження стосовно використання незалежних вибірок під час дослідження. У цьому випадку можемо припустити, що вибірки не є незалежними, оскільки в середньому бали, отримані студентами на проміжному іспиті з вищої математики, будуть певним чином визначати результати заключного іспиту. Оскільки умова задачі не передбачає нормального розподілу генеральних сукупностей — балів, отриманих студентами під час зимової та літньої екзаменаційних сесій відповідно, та інформація про значення генеральних дисперсій невідома, то для можливості застосування T-критерію Стьюдента для перевірки гіпотези про рівність середніх генеральних значень у разі залежних вибірок, треба здійснити вибірки обсягом, більшим ніж 30. Отже, для перевірки гіпотези про відмінність між балами, отриманими студентами на іспитах з вищої математики під час зимової та літньої екзаменаційних сесій, було здійснено вибірку із $n = 100$ пар спостережень кількості балів, отриманих під час зимової та літньої екзаменаційних сесій відповідно. Середня різниця оцінок студентів, отриманих на проміжному і заключному іспитах, склала 6 балів ($\overline{d} = 6$), а виправлене стандартне відхилення цієї різниці, обчислене за результатами вибіркових спостережень, склало 4 бали ($s_d = 4$). Оскільки рівень значущості не задано, обираємо його самостійно, наприклад 10 %.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_d = 0$ — суттєвої розбіжності між балами студентів, отриманими на проміжному та заключному іспитах, немає ($\mu_d = \mu_{d_0} = 0$ — це означає, що гіпотетична різниця між середніми значеннями генеральних сукупностей дорівнює нулю).

$H_1 : \mu_d \neq 0$ — суттєва розбіжність між балами студентів, отриманими на проміжному та заключному іспитах, існує.

Крок другий. Середня вибіркова різниця у разі істинності гіпотези H_0 розподілена за законом Стьюдента, оскільки дисперсія генеральної сукупності невідома.

Крок третій. Знаходимо спостережуване значення критерію:

$$t^* = \frac{\overline{d} - \mu_{d_0}}{s_d} \cdot \sqrt{n} = \frac{6 - 0}{4} \cdot \sqrt{100} = 15.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \mu_d \neq 0$ будуємо двосторонню критичну область. Оскільки взято вибірковий обсяг спостережень 100, критичні точки знаходимо з таблиці значень функції Лапласа: $\Phi(t_{кр}) = \frac{1-\alpha}{2}$, $\Phi(t_{кр}) = 0{,}45$. Отже, $t_{кр} = 1{,}645$ — правостороння межа критичної області. Область невідхилення нульової гіпотези:

$$H_0 : t \in (-1{,}645;\ 1{,}645).$$

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = 15$ знаходиться в області відхилення нульової гіпотези, робимо висновок про достатність статистичних фактів з рівнем значущості 10 % на користь альтернативної гіпотези про розбіжність між балами студентів, отриманими на іспитах з вищої математики під час зимової та літньої екзаменаційних сесій.

Вправа. За умовою попередньої задачі перевірити нульову гіпотезу $H_0 : \mu_d = 0$, за альтернативної гіпотези $H_1 : \mu_d \neq 0$, якщо $n = 30$, $\alpha = 10$ %, $\overline{d} = s_d = 4$ (відповідь: $t^* = 8{,}22$; $t_{кр} = t$ ($v = 29$; $\frac{\alpha}{2} = 0{,}05$) $= 1{,}699$; висновок на користь невідхилення H_1: $\mu_d \neq 0$).

Приклад 5.30. У таблиці наведено дані щоквартальних звітів дохідності з 1997 по 2002 рік для двох портфелів, спеціалізованих на дорогоцінних металах. Відомо, що вони тісно пов'язані однаковими факторами ризику та майже ідентичними формами затрат. Але на початку 2003 року, за результатами звітів щоквартальної дохідності, портфель В був поставлений інвестиційною компанією набагато вище за рейтингом, ніж портфель А. Чи можна отримати статистичний доказ того, що середня різниця прибутків портфеля А і портфеля В дійсно була значущою протягом звітного періоду? Вважаємо, що випадкові величини щоквартальної дохідності обох портфелів були розподілені нормально протягом звітного періоду.

Квартал	Дохідність портфеля А(%)	Дохідність портфеля В(%)	Різниця
IV кв. 2002	11,40	14,64	–3,24
III кв. 2002	–2,17	0,44	–2,61
II кв. 2002	10,72	19,51	–8,79
I кв. 2002	38,91	50,40	–11,49
IV кв. 2001	4,36	1,01	3,35
III кв. 2001	5,13	10,18	–5,05
II кв. 2001	26,36	17,77	8,59
I кв. 2001	–5,53	4,76	–10,29
IV кв. 2000	5,27	–5,36	10,63
III кв. 2000	–7,82	–1,54	–6,28
II кв. 2000	2,34	0,19	2,15
I кв. 2001	–14,38	–12,07	–2,31
IV кв. 1999	–9,80	–9,98	0,18
III кв. 1999	19,03	26,18	–7,15
II кв. 1999	4,11	–2,39	6,50
I кв. 1999	–4,12	–2,51	–1,61
IV кв. 1998	–0,53	–11,32	10,79
III кв. 1998	5,06	0,46	4,60
II кв. 1998	–14,01	–11,56	–2,45
I кв. 1998	12,50	3,52	8,98
IV кв. 1997	–29,05	–22,45	–6,60
III кв. 1997	3,60	0,10	3,50
II кв. 1997	–7,97	–8,96	0,99
I кв. 1997	–8,62	–0,66	–7,96
Середнє	1,87	2,52	–0,65

Стандартне відхилення різниць парних спостережень від середньої вибіркової дорівнює 6,71 (перевірити самостійно).

Оскільки відомо, що два портфелі тісно пов'язані однаковими факторами ризику та майже ідентичними формами затрат, їх прибутковості не можна розглядати як незалежні. Отже, для відповіді на поставлене запитання необхідно застосовувати критерій парного порівняння. Нехай μ_d середнє за сукупністю значення різниць між прибутками двох портфелів протягом звітного періоду.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : \mu_d = 0$ — суттєвої розбіжності між прибутковістю портфелів A та B немає.

$H_1 : \mu_d \neq 0$ — існує суттєва розбіжність між прибутковістю портфелів A та B (або інакше $H_0 : \mu_1 - \mu_2 = 0$ за альтернативної гіпотези $H_1 : \mu_1 - \mu_2 \neq 0$).

Крок другий. Середня різниця вибірки розподілена за законом Стьюдента, оскільки дисперсія генеральної сукупності невідома.

Крок третій. Знаходимо спостережуване значення критерію:

$$t^* = \frac{\bar{d} - \mu_{d_0}}{s_d} \cdot \sqrt{n} = \frac{-0{,}65 - 0}{6{,}71} \cdot \sqrt{24} = -\,0{,}475\,.$$

Крок четвертий. За змістом альтернативної гіпотези $H_1 : \mu_d \neq 0$ будуємо двосторонню критичну область. Рівень значущості обираємо 5 %. За таблицею критичних значень розподілу Стьюдента знаходимо:

$$t_{кр} = t(v = 23;\ \frac{\alpha}{2} = 0,\,025) = 2{,}069.$$

Область невідхилення нульової гіпотези $H_0 : t \in (-2{,}069;\ 2{,}069)$.

Крок п'ятий. Оскільки спостережуване значення критерію $t^* = -0{,}475$ знаходиться в області невідхилення нульової гіпотези, робимо висновок про недостатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези про розбіжність між прибутковістю портфелів A та B.

Зауважимо, що навіть для рівня значущості 10 % $t_{кр} = t\ (\nu = 23;\ \frac{\alpha}{2} = 0{,}05) = 1{,}714$, і знову спостережуване значення критерію потрапляє до області невідхилення нульової гіпотези $t \in (-1{,}714;\ 1{,}714)$. Висновок не змінюється.

Зауваження 10. Для великих обсягів вибірок (n_1, $n_2 > 30$) розподіл випадкової величини U асимптотично наближається до нормального $N(0;\ 1)$. У цьому випадку критичні точки визначають за таблицею значень функції Лапласа.

Зауваження 11. Перевірку статистичних гіпотез з використанням критеріїв значущості можна здійснити, застосовуючи довірчі інтервали. При цьому односторонньому критерію значущості відповідає односторонній довірчий інтервал, а двосторонньому критерію значущості — двосторонній довірчий інтервал. Гіпотеза H_0 приймається, якщо спостережуване значення критерію покривається відповідним довірчим інтервалом.

Якщо перевіряється гіпотеза $H_0:\ \theta_1 = \theta_2$ про рівність параметрів генеральних сукупностей, то розглядається довірчий інтервал для різниці θ_1 та θ_2. Гіпотеза H_0 приймається, якщо довірчий інтервал для різниці параметрів θ_1 та θ_2 покриває нульове значення. Винятком є перевірка гіпотези про рівність дисперсій $H_0:\ \sigma_X^2 = \sigma_Y^2$, оскільки довірчий інтервал будується для відношення дисперсій, тому гіпотеза H_0 у цьому випадку приймається, якщо довірчий інтервал покриває значення, що дорівнює 1.

5.9. ПЕРЕВІРКА СТАТИСТИЧНОЇ ГІПОТЕЗИ ПРО РІВНІСТЬ ЧАСТОК ОЗНАКИ У ДВОХ СУКУПНОСТЯХ

Задача про порівняння часток ознаки у двох генеральних сукупностях досить часто трапляється на практиці. Наприклад, вибіркова частка ознаки однієї сукупності відмінна від вибіркової частки цієї самої ознаки в іншій сукупності. Чи свідчить це про те, що частки ознаки двох генеральних сукупностей є дійсно різними, чи отримана розбіжність часток ознаки двох генеральних сукупностей є випадковою?

Нехай незалежні випробування проводяться у двох генеральних сукупностях. У результаті кожного випробування досліджувана ознака може з'явитися з невідомими ймовірностями $p_1 = \text{const}$ і $p_2 = \text{const}$ та не з'явитися з імовірностями $q_1 = 1 - p_1$ та $q_2 = 1 - p_2$ у першій і другій генеральних сукупностях відповідно. Невідомі ймовірності p_1 та p_2 інакше називають ***генеральними частками ознаки*** першої та другої сукупностей. Необхідно із заданим рівнем значущості α перевірити нульову гіпотезу про рівність значень імовірностей p_1 та p_2, тобто перевірити гіпотезу H_0: $p_1 = p_2 = p$. Для перевірки гіпотези H_0 із генеральних сукупностей реалізують дві незалежні вибірки достатньо великих розмірів n_1 і n_2. Припустимо, що спостережуване число появ досліджуваної ознаки для першої вибірки дорівнює m_1, для другої вибірки дорівнює m_2 (m_1 і m_2 — відповідні кількості елементів першої та другої вибіркових сукупностей, яким притаманна ознака, що досліджується).

Позначимо через $w_1 = \frac{m_1}{n_1}$ та $w_2 = \frac{m_2}{n_2}$ відносні частоти появи події A в першій і другій вибіркових сукупностях (інакше їх називають частками ознаки вибіркових сукупностей). Приймемо спостережувані відносні частоти w_1 та w_2 за оцінки невідомих імовірностей p_1 та p_2 появи події A під час проведення незалежних випробувань у генеральних сукупностях. Оскільки ймовірності оцінюються відносними частотами, задачу можна переформулювати так: установити, чи є значущою різниця між значеннями відносних частот W_1 та W_2.

Зауважимо, що до проведення випробувань значення відносних частот W_1 та W_2 нам невідомі, тому розглядаємо їх як випадкові величини $W_1 = \frac{M_1}{n_1}$ та $W_2 = \frac{M_2}{n_2}$ тимчасом як не є випадковими величинами, оскільки їх значення отримані в результаті обробки вибіркових даних (після проведення незалежних випробувань). Тобто потрібно розуміти, що в результаті проведення незалежних випробувань значення w_1 та w_2, які спостерігаємо, є значеннями випадкових величин W_1 та W_2.

Випадкові величини $W_1 = \frac{M_1}{n_1}$ та $W_2 = \frac{M_2}{n_1}$ мають біноміальний закон розподілу, але для достатньо великих обсягів вибірок (n_1 та n_2 більше тридцяти) їх можна вважати розподіленими майже нормально з математичними сподіваннями:

$M(W_1) = p_1$; $M(W_2) = p_2$ та дисперсіями $D(W_1) = \frac{p_1 \cdot q_1}{n_1}$; $D(W_2) = \frac{p_2 \cdot q_2}{n_2}$.

Якщо справджується гіпотеза $H_0 : p_1 = p_2, = p$, то різниця випадкових величин $(W_1 - W_2)$ також розподілена майже нормально з математичним сподіванням $M(W_1 - W_2) = M(W_1) - M(W_2) = p_1 - p_2 =$ $= p - p = 0$ та дисперсією $D(W_1 - W_2) = D(W_1) + D(W_2) = \frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2} = \frac{p \cdot q}{n_1} + \frac{p \cdot q}{n_2} =$ $= p \cdot q \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = p\,(1 - p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Отже, за статистичний критерій перевірки нульової гіпотези обираємо випадкову величину Z, яка за умови достовірності гіпотези H_0 має наближено нормальний закон розподілу, причому $M(Z) = 0$; $D(Z) = 1$.

$$Z = \frac{(W_1 - W_2) - (p_1 - p_2)}{\sigma(W_1 - W_2)} = \frac{(W_1 - W_2) - (p_1 - p_2)}{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{5.43}$$

За невідоме значення ймовірності p у формулі (5.43) оберемо її найкращу оцінку, що дорівнює вибірковій частці ознаки, якщо дві вибірки об’єднати в одну, тобто:

$$p = \frac{m_1 + m_2}{n_1 + n_2}. \tag{5.44}$$

Доведемо, що $M(Z) = 0$; $D(Z) = 1$.

Дійсно, оскільки $M(W_1 - W_2) = p - p = 0$, то дістанемо:

$$M(Z) = M\left(\frac{W_1 - W_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right) = \frac{M(W_1) - M(W_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0;$$

$$D(Z) = D\left(\frac{W_1 - W_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right) = \frac{D(W_1) - D(W_2)}{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{\frac{pq}{n_1} + \frac{pq}{n_2}}{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 1.$$

Спостережуване значення критерію знаходимо за формулою (5.43), підставляючи результати, одержані з вибіркових сукупностей:

$$\mathrm{Z} = \frac{(w_1 - w_2) - (p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ де } w_1 = \frac{m_1}{n_1} \text{ та } w_2 = \frac{m_2}{n_2}. \tag{5.45}$$

Вибір критичної області проводиться згідно з методикою, розглянутою раніше для випадкової величини, розподіленої нормально. Критичні точки — межі критичної області — знаходять за таблицями значень функції Лапласа.

Приклад 5.31. Податкова інспекція перевіряє фінансову документацію підприємства. Детально вивчено випадкову вибірку $n_1 = 500$ закритих рахунків. У сорока із них виявили неправильне нарахування розміру податку. Інспектори запропонували вдосконалити процедуру оподаткування. Після закінчення відведеного терміну знову перевірили випадкову вибірку $n_2 = 600$ закритих рахунків; у тридцяти з них виявили порушення під час нарахування податку. Чи є підстави вважати, що нова процедура привела до зменшення порушень?

Р о з в ’ я з а н н я.

Крок перший. Обираємо основну нульову H_0 та альтернативну H_1 гіпотези:

$H_0 : p_1 = p_2$ — удосконалення процедури оподаткування не привело до зменшення порушень у нарахуваннях розміру податку.

$H_1 : p_1 > p_2$ — удосконалення процедури оподаткування привело до зменшення порушень у нарахуваннях розміру податку.

Крок другий. За статистичний критерій перевірки нульової гіпотези обираємо випадкову величину (5.43):

$$Z = \frac{(W_1 - W_2) - (p_1 - p_2)}{\sigma(W_1 - W_2)} = \frac{(W_1 - W_2) - (p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Крок третій. Знаходимо спостережуване значення критерію, але спочатку обчислимо значення p за формулою (5.44):

$$p = \frac{m_1 + m_2}{n_1 + n_2} = \frac{40 + 30}{500 + 600} \approx 0{,}0636,$$

$$\text{тоді } z^* = \frac{\frac{40}{500} - \frac{30}{600}}{\sqrt{0{,}0636 \cdot 0{,}9364\left(\frac{1}{500} + \frac{1}{600}\right)}} \approx 2{,}03.$$

***Крок четвертий*.**

За змістом альтернативної гіпотези $H_1 : p_1 > p_2$ будуємо правосторонню критичну область. Обираємо рівень значущості 5 %. За таблицею значень функції Лапласа з умови $\Phi(z_{\text{кр}}) = \frac{1-2\alpha}{2}$, знайдемо $\Phi(z_{\text{кр}}) = 0{,}45$, звідки $z_{\text{кр}} = 1{,}645$.

Область відхилення нульової гіпотези $H_0 : \{Z/z > 1{,}645\}$.

Крок п'ятий.

Оскільки спостережуване значення критерію $z^* \approx 2{,}03$ знаходиться в області відхилення нульової гіпотези, робимо висновок про достатність статистичних фактів з рівнем значущості 5 % на користь альтернативної гіпотези про те, що вдосконалення процедури оподаткування привело до зменшення порушень у нарахуваннях розміру податку.

5.10. ПЕРЕВІРКА ПРАВИЛЬНОСТІ НЕПАРАМЕТРИЧНИХ ГІПОТЕЗ

Важливим завданням математичної статистики є встановлення теоретичного закону розподілу: перевірка гіпотези щодо припустимого закону невідомого розподілу в генеральній сукупності.

Для розв'язання цієї задачі потрібно визначити вид і параметри закону розподілу.

Припущення про вид закону розподілу (нормальний, Пуассона, біноміальний тощо) висувається на підставі деяких формальних властивостей статистичного розподілу:

- графічного зображення емпіричного розподілу (гістограми частот або полігону вибірки);
- виконання умов центральної граничної теореми (це свідчить про нормальний закон розподілу випадкової величини);
- досвід аналогічних попередніх досліджень;
- якщо $As = 0$ і $Es = 0$, то це свідчить про нормальний закон розподілу;
- якщо $\bar{x}_{\text{в}} = \sigma_{\text{в}}$ — це ознака показникового розподілу.

Нехай гістограма має вигляд, зображений на рис. 5.32. Якщо з'єднати кривою середини сторін прямокутників, то одержимо лінію, схожу на графік щільності нормального розподілу. Тому можемо припустити, що ознака генеральної сукупності розподілена за нормальним законом.

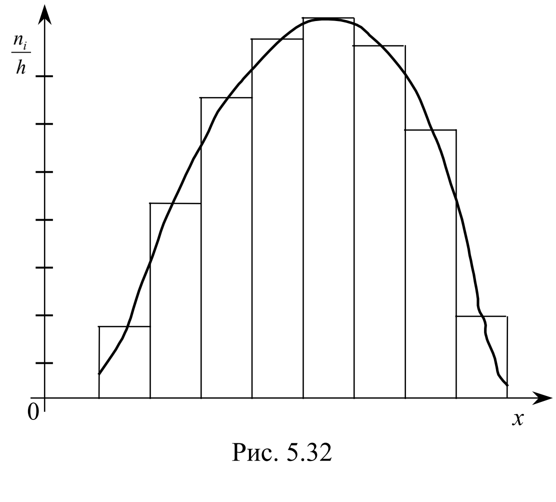


Рис. 5.32

Якщо гістограма має вигляд, зображений на рис. 5.33, маємо підставу вважати, що ознака розподілена за показниковим законом, оскільки лінія, що з'єднує середини сторін прямокутників, подібна до графіка щільності показникового закону.

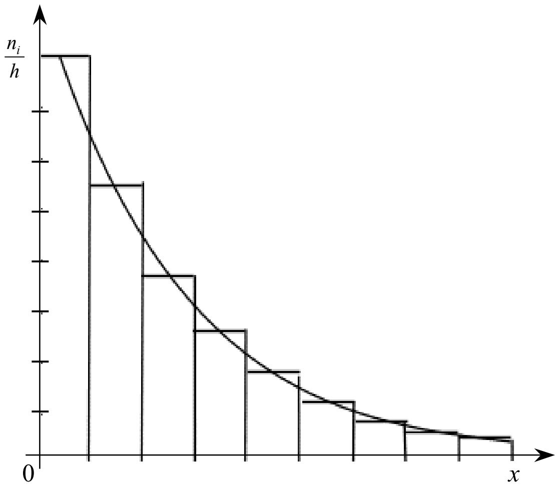


Рис. 5.33

Якщо гістограма має вигляд, зображений на рис. 5.34, можемо висунути гіпотезу про рівномірний закон розподілу сукупності.

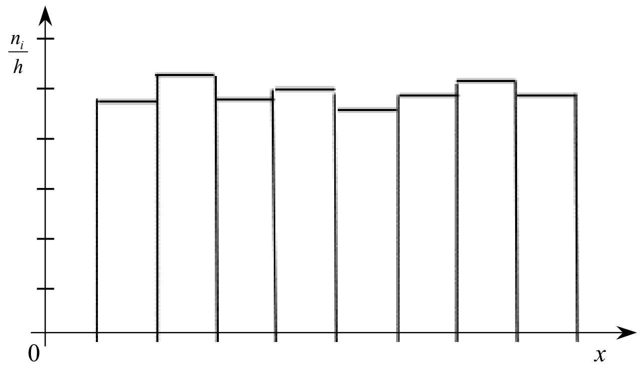


Рис. 5.34

Параметри розподілу, як правило, невідомі, тому їх замінюють найкращими їхніми оцінками за вибіркою. Як би добре не був підібраний теоретичний закон розподілу, між теоретичним і емпіричним законами можливі розбіжності. Закономірно, що виникає запитання: ці розбіжності пояснюються тільки випадковими обставинами, що пов'язані з обмеженою кількістю спостережень, чи вони є суттєвими і пов'язані з тим, що невдало підібрано теоретичний закон. Щоб розв'язати задачу про перевірку правильності вибору виду розподілу, погодженості дійсного теоретичного розподілу з емпіричним, застосовують спеціально підібрану випадкову величину — так званий ***критерій згоди*** (критерій узгодження).

Нагадаємо, що статистичні критерії бувають двох видів: параметричні та непараметричні. ***Параметричними*** називають критерії, які ґрунтуються на припущенні, що розподіл випадкової величини в сукупності підпорядковується деякому відомому закону розподілу (нормальному, Пуассона, біноміальному тощо).

Багато із розглянутих вище критеріїв перевірки параметричних статистичних гіпотез ґрунтувалися на припущенні, що ознака генеральної сукупності розподілена за нормальним законом. Тому використовувати їх можна у випадку, коли є достатня впевненість у тому, що спостережувана ознака генеральної сукупності має нормальний закон розподілу або близький до нього і що за іншого розподілу висновки стосовно статистичних гіпотез можуть бути хибними.

Існують різні критерії, які не потребують значень параметрів генеральної сукупності. Вони називаються ***непараметричними критеріями***.

Основна перевага непараметричних критеріїв полягає в тому, що вони не вимагають виконання припущення про те, що вибірка взята з нормально розподіленої сукупності. Крім того, вони простіші в розрахунках, легше інтерпретуються і підходять для широкого кола розподілів генеральної сукупності. Недолік їх полягає в тому, що вони не використовують усю інформацію *і менш ефективні, ніж параметричні, тобто непараметричні критерії порівняно з параметричними мають трохи меншу потужність, але цей недолік компенсується простішою побудовою вибіркових статистик.*

Отже, вдруге акцентуємо увагу на тому, що параметричні критерії більш ефективні порівняно з непараметричними, однак їх можна використовувати для сукупностей, які розподілені за нормальним або близьким до нормального законом. Натомість непараметричні критерії використовують за будь-якої форми розподілу. Такими критеріями згоди є: критерій К. Пірсона, О. М. Колмогорова, Б. С. Ястремського, М. В. Смирнова, критерій серій, критерій знаків та ін.

Оскільки нормальний розподіл трапляється найчастіше, то, здебільшого, перевіряють гіпотези про відповідність вибіркового розподілу нормальному, хоча непоодинокі випадки перевірки гіпотез щодо відповідності вибраних законів розподілу (біноміальний, Пуассона, тощо) розподілу в генеральній сукупності.

Найчастіше для розрахунку критеріїв згоди використовують відхилення емпіричних частот від теоретичних. Що меншим є вказане вище відхилення, Тим точніше теоретичний розподіл відтворює вибірковий (і навпаки).

Із усієї множини критеріїв згоди, що використовуються під час перевірки гіпотез стосовно розподілів, частіше від інших (є найпоширенішим) застосовується критерій χ^2 Пірсона, який вважається найпотужнішим.

5.10.1. Критерій χ^2 Пірсона

Критерій Пірсона забезпечує мінімальну ймовірність похибки другого роду порівняно з іншими критеріями згоди.

Критерій ґрунтується на порівнянні теоретичних та емпіричних частот.

Нехай результати вибірки подано у вигляді інтервального статистичного ряду, що містить k інтервалів. Скористаємось вибірковою функцією:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i} = \sum_{i=1}^{k} \frac{\left(n_i - n_i'\right)^2}{n_i'}. \tag{5.46}$$

Тут n — обсяг вибірки, n_i — емпіричні частоти, одержані за результатами вибірки, $P\left(x_{i-1} < X < x_i\right)$, $i = \overline{1, \ell}$, p_i — імовірності потрапляння випадкової величини X на інтервал $\left(x_{i-1}; x_i\right)$, які обчислені для гіпотетичного закону розподілу, що перевіряється; $n_i' = np_i$ — теоретичні частоти.

Для нормального закону розподілу теоретичні частоти обчислюють за такими формулами:

$$n_i' = np_i = \frac{nh}{\sigma_B} \varphi\left(\frac{x_i - \overline{x}_B}{\sigma_B}\right), \; h = x_i - x_{i\text{-}1}, \tag{5.47}$$

де $\varphi(x)$ — функція Гаусса, h — довжина інтервалу;

$$n_i' = np_i = n\left(\Phi\left(\frac{x_i - \overline{x}_B}{\sigma_B}\right) - \Phi\left(\frac{x_{i-1} - \overline{x}_B}{\sigma_B}\right)\right), \tag{5.48}$$

де $\Phi(x)$ — функція Лапласа.

Формула (5.47) використовується, якщо результати вибірки подано у вигляді точкового статистичного ряду з рівновіддаленими варіантами, формула (5.48) використовується, якщо результати вибірки подано у вигляді інтервального статистичного ряду.

Для експоненційного закону розподілу теоретичні частоти обчислюють за формулою:

$$n_i' = np_i = n\left(e^{\frac{-x_i}{\overline{x}_B}} - e^{\frac{-x_{i-1}}{\overline{x}_B}}\right). \tag{5.49}$$

Якщо потрібно перевірити гіпотезу H_0 про те, що сукупність розподілена за законом Пуассона, то теоретичні частоти обчислюють за формулою:

$$n_m' = np_m = nP(X = m) = \frac{n\overline{x}_{\text{в}}^{\,m}}{m!} e^{-\overline{x}_{\text{в}}} \tag{5.50}$$

Для знаходження ймовірностей $P(X=m)$ може бути використана таблиця ймовірностей розподілу Пуассона. Якщо для обчисленого значення $\bar{x}_B$ ймовірності $P(X=m)$ нетабульовані, то для полегшення обчислень можна застосувати рекурентну формулу:

$$P(X=m+1)=\frac{\bar{x}_{\text{в}}}{m+1}P(X=m); \qquad P(X=0)=e^{-\bar{x}_{\text{в}}}. \tag{5.51}$$

Якщо $n\to\infty$, то незалежно від закону розподілу ознаки генеральної сукупності вибіркова функція має розподіл χ^2 з $\nu=l-r-1$ ступенями вільності. Тут l кількість інтервалів, r — кількість параметрів закону розподілу, оцінки яких знайдено за вибірковими даними.

Критерій згоди χ^2 Пірсона можна використовувати, якщо обсяг вибірки досить великий (близько кількох сотень). Іноді його використовують при $n\ge 50$. Окрім того, частоти у кожному інтервалі мають бути не менші ніж п'ять. Якщо у деяких інтервалах частоти менші за п'ять, слід об'єднати кілька сусідніх інтервалів.

Зауваження. Якщо усі емпіричні частоти збігаються з теоретичними, то за формулою (5.46) маємо $\chi^2=0$, у протилежному випадку — завжди $\chi^2>0$. Тому критична область є правобічною.

За результатами вибірки обчислюють значення χ^2_{cn}. За заданим рівнем значущості α і числом ступенів вільності $\nu=l-r-1$ із таблиці критичних точок розподілу χ^2 знаходимо $\chi^2_{kp}=\chi^2(\alpha;k)$. Якщо $\chi^2_{cn}<\chi^2_{kp}$, то гіпотезу про закон розподілу приймаємо; якщо $\chi^2_{cn}>\chi^2_{kp}$, то гіпотезу про закон розподілу відхиляємо.

Перевірку гіпотези H_0 про те, що генеральна сукупність має нормальний розподіл ознаки X, виконують за такою схемою:

1. Обчислюють $\bar{x}_{\text{в}};\sigma_{\text{в}};u_i=\dfrac{x_i-\bar{x}_{\text{в}}}{\sigma_{\text{в}}}, i=\overline{0,k}$.
2. Обчислюють теоретичні частоти за формулою (5.48), причому $u_0=-\infty,\ u_k=+\infty$.
3. Обчислюють спостережуване значення критерію $\chi^2_{\text{спост.}}$ за формулою (5.46).
4. Визначається критична точка $\chi^2_{\text{кр}}$ правобічної критичної області за заданого рівня значущості α і числом ступенів вільності ν з таблиці 7 додатка.
5. Робиться висновок щодо гіпотези H_0.

Використання критерію χ^2 як критерію згоди для перевірки гіпотези про відповідність емпіричного розподілу нормальному.

Приклад 5.32. Для аналізу результатів екзамену з теорії ймовірностей і математичної статистики було вибрано чотири академічних групи. Дані про кількість балів, одержаних за екзаменаційну роботу, подано у вигляді інтервального статистичного ряду:

Інтервали	6—12	12—18	18—24	24—30	30—36	36—42	42—48	48—54	54—60
Частоти	5	6	8	7	17	24	16	11	6

Побудувати гістограму частот. Висунути гіпотезу про закон розподілу оцінок за екзаменаційну роботу та перевірити її за рівня значущості $\alpha=0{,}05$.

Р о з в ' я з а н н я. Будуємо гістограму частот (рис. 5.35):

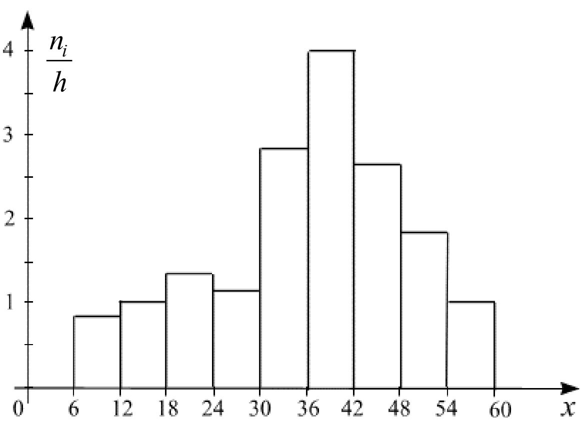


Рис. 5.35

Аналізуючи вигляд гістограми частот, висуваємо нульову гіпотезу H_0: оцінки екзаменаційної роботи розподілені за нормальним законом.

Перевіримо правильність нульової гіпотези H_0 за критерієм Пірсона. Обчислимо $\overline{x}_в$ та $\sigma_в$. Для цього від інтервального статистичного ряду перейдемо до статистичного ряду:

x_i	9	15	21	27	33	39	45	51	57
n_i	5	6	8	7	17	24	16	11	6

Обсяг вибірки $n=\sum_{i=1}^{9} n_i = 100.$ Обчислимо:

$$\overline{x}_в = \frac{1}{100}\left(9\cdot 5 + 15\cdot 6 + 21\cdot 8 + 27\cdot 7 + 33\cdot 17 + 39\cdot 24 + 45\cdot 16 + 51\cdot 11 + 57\cdot 6\right) = 36{,}12\ ;$$

$$\overline{x_в^{\,2}} = \frac{1}{n}\sum_{i=1}^{9} x_i^2\cdot n_i = \frac{1}{100}(9^2\cdot 5 + 15^2\cdot 6 + 21^2\cdot 8 + 27^2\cdot 7 + 33^2\cdot 17 +$$

$$+39^2\cdot 24 + 45^2\cdot 16 + 51^2\cdot 11 + 57^2\cdot 6) = 1459{,}08;$$

$$D_в = \overline{x^2}_в - \left(\overline{x}_в\right)^2 = 1459{,}08 - 36{,}12^2 = 154{,}4256;$$

$$\sigma_в = \sqrt{D_в} = \sqrt{154{,}4256} \approx 12{,}4268.$$

Імовірність того, що нормально розподілена випадкова величина потрапляє на відрізок $\lfloor x_{i-1}; x_i \rfloor$; обчислюється за формулою:

$$p_i = \Phi(z_i) - \Phi(z_{i-1}),\ \text{тут}\ z_{i-1} = \frac{x_{i-1}-\overline{x}_в}{\sigma_в}, z_i = \frac{x_i - \overline{x}_в}{\sigma_в}.$$

Обчислення теоретичних частот подано у таблиці:

x_{i-1}	x_i	n_i	z_{i-1}	z_i	$\Phi(z_{i-1})$	$\Phi(z_i)$	$n_i' = np_i$	$(n_i - np_i)^2$	$\frac{(n_i-np_i)^2}{np_i}$
$-\infty$	12,00	$\left.\begin{matrix}5\\6\end{matrix}\right\}=11$	$-\infty$	—1,94	—0,5000	—0,4738	$\left.\begin{matrix}2{,}62\\4{,}60\end{matrix}\right\}=7{,}22$	14,29	1,98
12,00	18,00		—1,94	—1,46	—0,4738	—0,4279			
18,00	24,00	8,00	—1,46	—0,98	—0,4279	—0,3365	9,14	1,30	0,14
24,00	30,00	7,00	—0,98	—0,49	—0,3365	—0,1879	14,85	61,66	4,15
30,00	36,00	17,00	—0,49	—0,01	—0,1879	—0,0040	18,39	1,94	0,11
36,00	42,00	24,00	—0,01	0,47	—0,0040	0,1808	18,48	30,46	1,65
42,00	48,00	16,00	0,47	0,96	0,1808	0,3315	15,06	0,87	0,06
48,00	54,00	11,00	0,96	1,44	0,3315	0,4251	9,36	2,69	0,29
54,00	$+\infty$	6,00	1,44	$+\infty$	0,4251	0,5000	7,49	2,23	0,30

Для перевірки обчислення теоретичних частот можна скористатися співвідношенням $\sum \text{np}_i = n$. У нашому випадку:

$$2{,}62 + 4{,}60 + 9{,}14 + 14{,}85 + 18{,}39 + 18{,}48 + 15{,}06 + 9{,}36 + 7{,}46 = 100.$$

Отже, теоретичні частоти обчислено правильно.

На перших двох інтервалах теоретичні частоти $n_i' = np_i < 5,$ тому об'єднаємо ці інтервали.

Обчислюємо спостережуване значення статистичного критерію:

$$\chi^2_{сп} = \sum_{i=1}^{8}\frac{(n_i - np_i)^2}{np_i} = 1{,}98 + 0{,}14 + 4{,}15 + 0{,}11 + 1{,}65 + 0{,}06 + 0{,}29 + 0{,}3 = 8{,}68.$$

Нормальний закон містить два параметри a та σ і ми об'єднали два перших інтервали, тому в нашому випадку число ступенів вільності буде дорівнювати 8 – 2 – 1 = 5.

За таблицею критичних точок розподілу χ^2 для рівня значущості $\alpha = 0{,}05$ і п'яти ступенів вільності знаходимо критичне значення статистичного критерію $\chi^2_{kp} = 11{,}1.$ Критична область правобічна, $\chi^2_{cn} < \chi^2_{kp}$, тому гіпотеза про нормальний закон розподілу приймається.

5.10.2. Перевірка гіпотези про розподіл генеральної сукупності за біноміальним законом

Для перевірки нульової гіпотези H_0 про біноміальний розподіл генеральної сукупності за рівня значущості α, виконуємо такі кроки:

1) висуваємо гіпотезу H_0 : емпіричний розподіл відповідає біноміальному; альтернативна гіпотеза H_1: емпіричний розподіл не відповідає біноміальному;

2) рівень значущості приймаємо α (якщо він не заданий за умовою);

3) застосовуємо критерій χ^2 Пірсона, для чого обчислимо $\chi^2_{\text{спост.}}$ за формулою:

$$\chi^2_{\text{спост.}} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'},$$

де $n_i' = np_i$, а p_i знаходимо за формулою Бернуллі $p_i = P(X = i) = C_n^i p^i q^{n-i}$;

4) знайдемо $\chi^2_{\text{крит}}$ за таблицею критичних точок розподілу $\chi^2(\alpha; v)$; де $v = s - r - 1$, а s — кількість груп вибірки, що залишилася після об'єднання малочисельних груп, $r = 1$ або $r = 0$ залежно від того, оцінювався параметр (p) за вибіркою, чи ні;

5) зіставляємо $\chi^2_{\text{кр}}$ і $\chi^2_{\text{спост.}}$. Якщо $\chi^2_{\text{кр}} > \chi^2_{\text{спост.}}$, то немає підстав, щоб відхилити нульову гіпотезу; якщо ж $\chi^2_{\text{кр}} \le \chi^2_{\text{спост.}}$, то нульову гіпотезу відхиляємо.

Зауваження. Якщо емпіричні частоти малочисельні (< 5), то їх і відповідні їм теоретичні частоти об'єднуємо з попередніми чи наступними значеннями. При цьому кількість груп вибірки зменшиться.

Приклад 5.33. У бібліотеці випадковим чином відібрано 300 комплектів по 6 книг. При цьому реєстрували кількість пошкоджених книг. У результаті одержали емпіричний розподіл, що його подано в таблиці:

x_i	0	1	2	3	4	5	6
n_i	85	93	65	29	16	9	3

Тут x_i — кількість пошкоджених книг в одному комплекті; n_i — частота, тобто кількість комплектів, які містять x_i пошкоджених книг.

Використовуючи критерій Пірсона, для рівня значущості 0,005 перевірити гіпотезу про те, що дискретна випадкова величина X — кількість пошкоджених книг, розподілена за біноміальним законом.

Р о з в ' я з а н н я.

1. Обчислимо частоту w та візьмемо її за оцінку ймовірності того, що випадково відібрана книга виявиться пошкодженою. За формулою Бернуллі $P_i = P(X = i) = C_n^i p^i q^{n-i}$ знайдемо ймовірності $P_i\left(i = \overline{0,6}\right)$ того, що подія, яку розглядаємо в задачі, з'явиться в $n = 6$ випробуваннях рівно i раз. Для цього всі дані занесемо до таблиці. Маємо:

$$p = w = \frac{\sum_{i=0}^{6} w_i \cdot i}{6} = \frac{\frac{1}{n}\sum_{i=0}^{6} n_i \cdot i}{6} = \frac{1}{300 \cdot 6} \cdot (0 \cdot 85 + 1 \cdot 93 + 2 \cdot 65 + 3 \cdot 29 + 4 \cdot 16 + 5 \cdot 9 + 6 \cdot 3) =$$

$$= \frac{1}{1800}(93 + 130 + 87 + 64 + 45 + 18) = \frac{437}{1800} \approx 0{,}24.$$

Оскільки $p \approx 0{,}24$, $q = 1 - p \approx 0{,}76$, тоді $p_0 = P_6(0) = C_6^0 p^0 \cdot q^6 = (0{,}76)^6 \approx 0{,}1927$;

$$p_1 = P_6(1) = C_6^1 p^1 q^5 = 6 \cdot 0{,}24 \cdot (0{,}76)^5 \approx 0{,}3651, \quad p_2 = C_6^2 p^2 q^4 = \frac{6 \cdot 5}{2} \cdot (0{,}24)^2 \cdot (0{,}76)^4 \approx 0{,}2882,$$

$$p_3 = \frac{6 \cdot 5 \cdot 4}{2 \cdot 3} \cdot (0{,}24)^3 \cdot (0{,}76)^3 \approx 0{,}1214, \qquad p_4 = \frac{6 \cdot 5}{1 \cdot 2} \cdot (0{,}24)^4 \cdot (0{,}76)^2 \approx 0{,}0287,$$

$$p_5 = 6 \cdot (0{,}24)^5 \cdot 0{,}76 \approx 0{,}0037; \; p_6 = (0{,}24)^6 \approx 0{,}0002.$$

2. Знайдемо $\chi^2_{\text{спост}} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'} = 128{,}717.$

Таблиця 5.4

i	n_i	$i \cdot n_i$	p_i	$n_i' = np_i = 300 \cdot p_i$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
0	85	0	0,1927	57,81	27,19	12,79
1	93	93	0,3651	109,53	273,24	2,495
2	65	130	0,2882	86,46	460,532	5,327
3	29	87	0,1214	36,42	55,056	1,512
4	16	64	0,0287	8,61	54,61	6,343
5 6	$\left.\begin{matrix}9\\3\end{matrix}\right\}12$	$\left.\begin{matrix}45\\18\end{matrix}\right\}63$	$\left.\begin{matrix}0{,}0037\\0{,}0002\end{matrix}\right\}0{,}0039$	$\left.\begin{matrix}1{,}11\\0{,}06\end{matrix}\right\}1{,}17$	117,29	100,25
$\sum$	300	200	1	300	—	128,717

3. Визначимо χ^2_{kp} за рівнем значущості $\alpha = 0{,}05$ і числом ступенів вільності $\nu = s - r - 1 = 6 - 1 - 1 = 4$ ($s = 6$, оскільки останні дві групи об'єднанні завдяки тому, що остання група малочисельна); тоді:

$$\chi^2_{\text{kp}} = \chi^2(0{,}05;4) = 9{,}5.$$

4. Оскільки $\chi^2_{kp} < \chi^2_{\text{спост}}$, то гіпотеза H_0 про відповідність емпіричного розподілу біноміальному відхиляється.

Розглянемо попередній приклад, змінивши числові дані задачі.

Приклад 5.34. У бібліотеці випадково відібрано 200 комплектів по 5 книг. Реєструвалася кількість пошкоджених книг (підкреслення, помарки тощо). У підсумку одержано варіаційний ряд:

Число пошкоджених книг в одній вибірці (x_i)	0	1	2	3	4	5
Частота n_i (кількість вибірок, що містять x_i пошкоджених книг)	72	77	34	14	2	1

Потрібно, використовуючи критерій Пірсона, для рівня значущості $\alpha = 0{,}05$, перевірити гіпотезу про те, що дискретна випадкова величина X (число пошкоджених книг) розподілена за біноміальним законом.

Р о з в ' я з а н н я.

1. Знайдемо частоту W і застосуємо її як оцінку ймовірності того, що навмання взята книга виявиться пошкодженою.

За формулою Бернуллі $p_i = C_n^i p^i q^{n-i}$ знайдемо ймовірність p_i $(i = 0,1,2,3,4,5)$ того, що подія, котра цікавить нас, з'явиться в $n = 5$ випробуваннях рівно i раз. Знайдемо

$$w = \frac{\sum_{i=0}^{s} w_i i}{5} = \frac{0 \cdot 72 + 1 \cdot 77 + 2 \cdot 34 + 3 \cdot 14 + 4 \cdot 2 + 5 \cdot 1}{200 \cdot 5} = \frac{200}{5 \cdot 200} = \frac{1}{5} = 0{,}2\,.$$

Таблиця 5.5

i	n_i	$i \cdot n_i$	p_i	$n_i' = np_i$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
0	72	0	0,32768	65,6	40,96	0,6244
1	77	77	0,4096	81,92	24,2064	0,2955
2	34	68	0,2048	40,96	48,4416	1,1827
3	14	42	0,0512	10,24	14,1376	1,3806
4 5	$\left.\begin{matrix}2\\1\end{matrix}\right\}3$	$\left.\begin{matrix}8\\5\end{matrix}\right\}13$	$\left.\begin{matrix}0{,}0064\\0{,}00032\end{matrix}\right\}0{,}00672$	$\left.\begin{matrix}1{,}28\\0{,}064\end{matrix}\right\}1{,}344$	$\left.\begin{matrix}0{,}5184\\0{,}8761\end{matrix}\right\}2{,}742$	2,04
Σ	200	200	1	200	—	5,5232

$$p_0 = C_5^0 p^0 q^5 = (0{,}8)^5 \approx 0{,}32768;\ p_1 = C_5^1 p^1 q^4 = 5 \cdot 0{,}2 \cdot (0{,}8)^4 = 0{,}4096;$$
$$p_2 = C_5^2 p^2 q^3 = 10 \cdot (0{,}2)^2 \cdot (0{,}8)^3 = 0{,}2048;\ p_3 = C_5^3 p^3 q^2 = 10 \cdot (0{,}2)^3 \cdot (0{,}8)^2 = 0{,}0512;$$
$$p_4 = C_5^4 p^4 q = 5 \cdot (0{,}2)^4 \cdot 0{,}8 = 0{,}0064;\ p_5 = C_5^5 p^5 q^0 = (0{,}2)^5 = 0{,}00032\,.$$

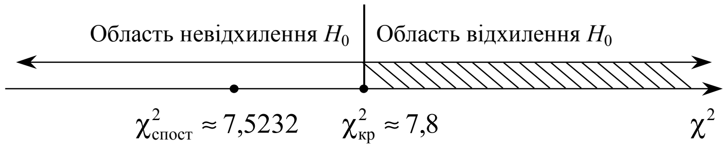
2. Для перевірки гіпотези висуваємо критерій:

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}\,,$$

де s — кількість груп вибірки, що залишилися після об'єднання.

3. Обчислимо $\chi^2_{\text{спост}}$. Під час складання розрахункової таблиці для порівняння емпіричних і теоретичних частот за допомогою критерію Пірсона ми об'єднали емпіричні частоти $(2+1)$ та відповідні імовірності $(0,0064 + 0,00032 \approx 0,00672)$. Після об'єднання кількість груп вибірки $s = 5$.

4. Знаходимо χ^2_{kp}. Один параметр (імовірність p) оцінювався за вибіркою, тобто $r = 1$, тому при визначенні числа ступенів вільності $\text{v} = s - r - 1 = 5 - 1 - 1 = 3$; $\chi^2_{kr}(0{,}05; 3) = 7{,}8$.



Оскільки $\chi^2_{\text{спост}} < \chi^2_{\text{крит}}$, то немає підстав відхилити нульову гіпотезу про біноміальний закон розподілу.

5.10.3. Перевірка гіпотези про розподіл випадкової величини за законом Пуассона

У цьому разі діємо аналогічно до попереднього випадку, тільки теоретичні частоти n_i' обчислюємо за формулою (5.50).

Приклад 5.35[1]. Проведено спостереження за кількістю викликів на телефонній станції. З цією метою протягом 100 випадково вибраних 5-секундних інтервалів часу реєструвалася кількість викликів. Одержали такий варіаційний ряд:

Кількість викликів x_i	0	1	2	3	4	5
Частоти n_i	8	28	31	18	9	6

Використовуючи критерій Пірсона, перевірити гіпотезу про те, що розподіл кількості викликів χ узгоджується із законом Пуассона. Рівень значущості прийняти $\alpha = 0{,}05$.

Розв'язання. $n = \sum n_i = 100$.

Крок 1. Гіпотеза H_0: випадкова величина X розподілена за законом Пуассона.

Крок 2. Обчислимо ймовірність m викликів протягом n випадково вибраних відрізків часу за формулою Пуассона:

$$P_n(m) = \begin{cases} \dfrac{\lambda^m}{m!} e^{-\lambda}, & \text{якщо } m = 0, 1, 2, \ldots, n, \\ 0, & \text{в усіх інших випадках.} \end{cases}$$

Крок 3. Знайдемо точкову оцінку параметра λ. Як відомо, точковою оцінкою для математичного сподівання генеральної сукупності є $\bar{x}_\text{в}$; обчислимо її:

$$\bar{x}_\text{в} = \frac{\sum_{i=1}^{6} x_i \cdot n_i}{n} = \frac{0 \cdot 8 + 1 \cdot 28 + 2 \cdot 31 + 3 \cdot 18 + 4 \cdot 9 + 5 \cdot 6}{100} = 2{,}1.$$

[1] *Л. И. Ниворожнина, З. А. Морозова* «Математическая статистика с элементами теории вероятностей в задачах с решениями». — М., 2005. — 608 с.

Тоді функція ймовірностей закону, який допускаємо у гіпотезі H_0, набуде вигляду:

$$P_n(m) = \frac{2{,}1^m \cdot e^{-2{,}1}}{m!}.$$

Крок 4. Застосовуємо критерій χ^2 Пірсона: $\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}$. Знайдемо $\chi^2_{\text{спост.}}$, для чого всі обчислення для зручності здійснимо в таблиці:

Кількість викликів x_i	n_i	$p_i = \frac{(2{,}1)^m \cdot e^{-2{,}1}}{m!}$	$np_i = n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
0	8	0,1225	12,25	18,0625	1,4745
1	28	0,2572	25,72	5,1984	0,2021
2	31	0,2700	27	16	0,5926
3	18	0,1890	18,9	0,81	0,0429
4	9	0,0992	9,92	0,8464	0,0853
5	6	0,0621	0,21	0,0441	0,0071
$\sum$	100	1	100	—	$\chi^2_{спот.} = 2{,}4045$

Крок 5. Із таблиці значень критичних точок розподілу χ^2 за даним рівнем значущості $\alpha = 0{,}05$ і числом ступенів вільності $v = s - r - 1 = 6 - 1 - 1 = 4$ знаходимо, що $\chi^2_{\text{кр}}(0{,}05; 4) = 9{,}5$.

Оскільки $\chi^2_{\text{спост.}} = 2{,}4045 < \chi^2_{\text{кр}} = 9{,}5$, тому немає підстав для відхилення нульової гіпотези про те, що кількість викликів на телефонній станції розподілена за законом Пуассона.

5.10.4. Перевірка гіпотези про розподіл випадкової величини за рівномірним законом

Для цього групуємо вибіркові дані, подаємо їх у вигляді інтервалів $[x_i, x_{i+1})$ та відповідних їм частот n_i, $i = \overline{1,k}$; $x_1 = a$, $x_{k+1} = b$. Після цього обчислимо ймовірності p_i — імовірності потрапляння випадкової величини X у частинні інтервали та визначимо теоретичні частоти n_i' за формулою:

$$n_i' = np_\text{i},$$

де $n = \sum_{i=1}^{k} n_i$ — обсяг вибірки, і знайдемо спостережуване значення критерію Пірсона $\chi^2_{\text{крит.}}$

Визначаємо $\chi^2_{\text{крит.}}$ за заданим рівнем значущості α та числом ступенів вільності $v = k - 1$ і приймаємо одне з таких рішень:

- якщо $\chi^2_{\text{спост.}} < \chi^2_{\text{кр}}$, то вважаємо, що немає достатніх підстав для відхилення гіпотези H_0 про рівномірний розподіл випадкової величини X на відрізку $[a; b]$;
- якщо $\chi^2_{\text{спост.}} \geq \chi^2_{\text{кр}}$, то гіпотезу H_0 відхиляємо.

Приклад 5.36. Автобус має п'ятихвилинний інтервал руху. Через кожну хвилину спостерігався такий розподіл кількості появ автобуса на зупинці:

Інтервал (у хв)	0—1	1—2	2—3	3—4	4—5
Кількість появ автобуса	34	36	40	39	41

Перевірити гіпотезу H_0 про рівномірний закон розподілу числа появ на зупинці автобуса.

Р о з в ' я з а н н я.

Крок 1. За заданим варіаційним рядом обчислюємо ймовірності p_i:

$$p_i = P(a_i < X < a_{i+1}) = \frac{x_{i+1} - x_i}{b - a}, \tag{5.52}$$

де $i = \overline{1{,}5}$; $x_1 = a = 0$, $x_5 = b = 5$.

Крок 2. Для перевірки гіпотези H_0 про те, що кількість появ автобуса на зупинці є рівномірно розподілена випадкова величина, застосовуємо критерій $\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}$.

Крок 3. Для зручності розрахунків зі знаходження $\chi^2_{\text{спост.}}$ складаємо таку таблицю:

Інтервали, хв	n_i	$p_i = \frac{x_{i+1} - x_i}{b-a} = \frac{x_{i+1} - x_i}{5}$	$n_i' = np_i$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
0—1	34	$\frac{1-0}{5} = 0{,}2$	38	16	0,4211
1—2	36	$\frac{2-1}{5} = 0{,}2$	38	4	0,1053
2—3	40	$\frac{3-2}{5} = 0{,}2$	38	4	0,1053
3—4	39	$\frac{4-3}{5} = 0{,}2$	38	1	0,0263
4—5	41	$\frac{5-4}{5} = 0{,}2$	38	9	0,2368
Σ	190	1	190	—	$\chi^2_{\text{спост.}} = 0{,}8948$

Крок 4. Визначимо $\chi^2_{\text{крит.}}$ за заданим рівнем значущості $\alpha = 0{,}005$ і числом ступенів вільності $v = k - 1 = 5 - 1 = 4$, тобто $\chi^2_{\text{крит.}} = 0{,}8948 < \chi^2_{\text{крит.}} = 9{,}5$, то нульова гіпотеза про рівномірний розподіл випадкової величини X на відрізку $[0; 5]$ приймається.

5.10.5. Перевірка гіпотези про розподіл неперервної випадкової величини за показниковим законом

Показниковий закон розподілу задається щільністю розподілу ймовірностей, тобто диференціальною функцією:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{для } x \geq 0, \\ 0 & \text{якщо } x < 0, \end{cases} \quad \text{де } \lambda > 0 \text{ — стала.}$$

Як і у попередніх п.п. 5.10.1—5.10.4, вихідні вибіркові дані групуємо, потім подаємо у вигляді послідовності частинних інтервалів і відповідних їм частот, а далі обчислюємо вибіркову середню $\overline{x_в}$. За оцінку параметра λ показникового розподілу беремо величину, обернену до $\overline{x_в}$: $\lambda = \frac{1}{\overline{x_в}}$.

Імовірності потрапляння випадкової величини X у частинні інтервали α_i, α_{i+1} знаходимо за формулою:

$$p_i = e^{\frac{-\alpha_{i+1}}{\bar{x}_в}} - e^{\frac{-\alpha_i}{\bar{x}_в}} = e^{-\lambda\alpha_{i+1}} - e^{-\lambda\alpha_i}. \tag{5.53}$$

Але якщо параметр λ заданий за умовою, то його оцінка замінюється точним значенням.

Критичне значення χ^2 визначаємо за заданим рівнем значущості та числом ступенів вільності $\text{v} = k - r - 1$, де r дорівнює 1 або 0 залежно від того, оцінювався чи ні за вибіркою параметр λ.

У результаті приймаємо одне з рішень: якщо $\chi^2_{\text{спост.}} < \chi^2_{\text{кр}}$, то вважаємо, що немає достатніх підстав відхилити нульову гіпотезу про розподіл випадкової величини X за показниковим законом; якщо ж $\chi^2_{\text{спост.}} < \chi^2_{\text{кр}}$, то нульову гіпотезу відхиляємо.

Приклад 5.37. Задано інтервальний ряд:

Інтервали	n_i
0—5	120
5—10	60
10—15	40
15—20	25
20—25	10
25—30	5
	$n=\sum_{i=1}^{5} n_i = 260$

Перевірити гіпотезу H_0 : випадкова величина χ розподілена за показниковим законом.

Р о з в ' я з а н н я. 1. Побудувавши гістограму частот, помітимо, що її вигляд нагадуватиме експонентну криву (рис. 5.36). Тому здійснимо «вирівнювання» статистичних даних за показниковим законом.

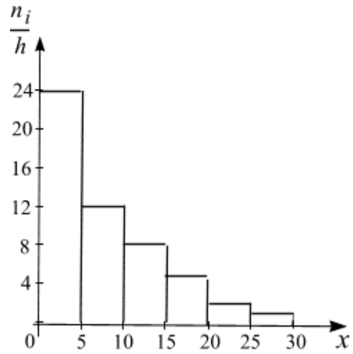


Рис. 5.36

Для знаходження точкової оцінки параметра $\lambda = \frac{1}{\overline{x}_{\text{в}}}$ обчислимо спочатку $\overline{x}_{\text{в}}$:

$$\overline{x}_{\text{в}} = \frac{\sum_{i=1}^{6} x_i n_i}{\sum_{i=1}^{6} n_i} = \frac{\sum_{i=1}^{6} x_i n_i}{260} =$$

$$= \frac{2{,}5\cdot 120 + 7{,}5\cdot 60 + 12{,}5\cdot 40 + 17{,}5\cdot 25 + 22{,}5\cdot 10 + 27{,}5\cdot 5}{260} = \frac{300+450+500+437{,}5+225+137{,}5}{260} \approx 7{,}8846 .$$

Отже, диференціальна функція запропонованого показникового закону розподілу має вигляд:

$$\begin{cases} f(x) = 0{,}1268\cdot e^{-0{,}1268x}, & x \ge 0; \\ 0, & x < 0. \end{cases}$$

2. Для перевірки відповідності емпіричних даних з передбачуваним показниковим законом розподілу застосуємо критерій згоди χ^2.

3. Обчислимо ймовірності потрапляння випадкової величини X у частинні інтервали (α_i, α_{i+1}) за формулою (5.53). Складемо допоміжну таблицю для знаходження $\chi^2_{\text{спост.}}$:

Таблиця 5.6

Інтервали	n_i	p_i	$n_i' = np_i$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
0—5	120	0,467	121,42	2,0164	0,0166
5—10	60	0,252	65,52	30,4704	0,4651
10—15	40	0,131	34,06	35,2836	1,0359
15—20	25	0,073	19,24	33,1776	1,7244
20—25	10	0,042	10,92	0,8464	0,0775
25—30	5	0,034	8,84	14,7456	1,668
	260	1	260	—	$\chi^2_{\text{спост.}} = 4{,}9876$

Тут $p_1 = e^0 - e^{-0,1268\cdot 5} = 1 - e^{-0,634} = 1 - 0,5326 = 0,4674$,

$p_2 = e^{-0,634} - e^{-0,1268\cdot 10} = 0,5326 - e^{-1,268} = 0,5326 - 0,2808 = 0,2518$,

$p_3 = e^{-0,1268\cdot 10} - e^{-0,1268\cdot 15} = e^{-1,268} - e^{-1,902} = 0,2808 - 0,1496 = 0,1312$,

$p_4 = e^{-1,902} - e^{-0,1268\cdot 20} = e^{-1,902} - e^{-2,536} \approx 0,074$,

$p_5 = e^{-2,536} - e^{-3,17} \approx 0,042$,

$p_6 = e^{-3,17} - e^{-3,804} \approx 0,034$.

Знайдемо з таблиці значень критичних точок χ^2 розподілу за рівнем значущості $\alpha = 0,05$ і числом ступенів вільності $v = k - r - 1 = 5 - 1 - 1 = 3$, критичне значення $\chi^2_{\text{крит.}}(0,05; 3) = 7,8$.

Оскільки $\chi^2_{\text{спост.}} = 4,9876 < 7,8$, то немає підстав для відхилення гіпотези про показниковий (експоненційний) закон розподілу.

Приклад 5.38. Встановити теоретичний закон розподілу ознаки X за заданим статистичним розподілом:

$[\alpha_i; \alpha_{i+1}]$	0—8	8—16	16—24	24—32	32—40
n_i	40	30	20	8	2

та за рівнем значущості $\alpha = 0,01$ перевірити правильність припущення.

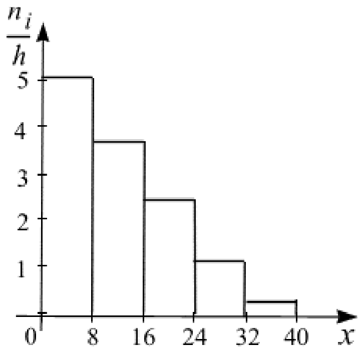


Рис. 5.37

Р о з в ' я з а н н я. Побудуємо гістограму частот (рис. 5.37). За формою одержаної гістограми припускаємо, що ознака X розподілена за показниковим законом. Для перевірки цього твердження застосовуємо критерій Пірсона. Теоретичні частоти n_i' обчислимо за формулою:

$$n_i' = n \cdot \left(e^{-\lambda\alpha_i} - e^{-\lambda\alpha_{i+1}}\right), \text{ де } \lambda = \frac{1}{\bar{x}_в}.$$

Перейдемо до дискретного розподілу та знайдемо $\bar{x}_в$:

x_i	4	12	20	28	36
n_i	40	30	20	8	2

$$\bar{x}_в = \frac{1}{n} \cdot \sum_{i=1}^{5} x_i \cdot n_i = \frac{1}{100}(4 \cdot 40 + 12 \cdot 30 + 20 \cdot 20 + 28 \cdot 8 + 36 \cdot 2) = 12,16;$$

тоді $\lambda = \dfrac{1}{12,16} \approx 0,0822$.

Обчислимо теоретичні частоти n_i', дані запишемо до табл. 5.7:

Таблиця 5.7

α_i	α_{i+1}	n_i	$e^{-\lambda\alpha_i} = e^{-0,082\alpha_i}$	$e^{-\lambda\alpha_{i+1}} = e^{-0,082\alpha_{i+1}}$	n'_i
0	8	40	1	0,5273	47,27
8	16	30	0,5273	0,2780	24,93
16	24	20	0,2780	0,1466	13,14
24	32	8	0,1466	0,083	6,36
32	40	2	0,083	0,06	2,3

Обчислимо спостережуване значення критерію (табл. 5.8):

Таблиця 5.8

	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
	40	42,27	—2,27	5,1529	0,1219
	30	24,93	5,07	25,7049	1,0311
	20	13,14	6,86	47,0596	3,5814
	8	6,36	1,64	2,6896	0,4229
	2	2,3	—0,3	0,09	0,0391
Σ	100	—	—	—	$\chi^2_{\text{спост}} = 5,1964$

Знайдемо критичну точку $\chi^2_{\text{кр}}$ для розподілу при $\alpha = 0,01$; $v = 5 - 1 - 1 = 3$; $\chi^2_{\text{кр}} = \chi^2(0,01;3) = 11,3$.

Оскільки $\chi^2_{\text{спост}} = 5,1964 < \chi^2_{\text{кр}} = 11,3$, то гіпотеза про показниковий закон розподілу ознаки X приймається.

5.10.6. Критерій серій

Критерій серій є непараметричною статистикою. Застосування цього критерію не вимагає обов'язкового знання закону розподілу, зокрема, що вибірка здійснена із нормально розподіленої сукупності. Критерій серій застосовується для розв'язання завдань, у яких потрібно встановити тенденції в розвитку досліджуваного процесу («тренду»). Надамо означення серій. Нехай є $n_1 + n_2$ елементів, серед яких n_1 елементів виду *«а»* та n_2 елементів виду *«в»*.

Серією називається частинна послідовність елементів одного виду, що входить в упорядковану послідовність елементів двох видів.

Кількість серій є дискретною величиною, якщо спосіб утворення послідовностей носить випадковий характер. Серією вважатимемо сукупність поряд розміщених чисел, які мають загальну властивість стосовно медіани: в одну групу поєднуються всі члени ряду, менші за медіану, в іншу — усі решта.

Порівнюючи кожне значення з вибірковою медіаною, можемо розбити всі значення на два типи (їх позначимо *«а»* та *«в»*), використовуючи те, що є значення більші або менші від медіани.

Перевіримо гіпотезу H_0: досліджувана вибірка є випадковою. Перевірку цієї гіпотези будемо здійснювати так. Спочатку спостережувані значення вибірки розташуємо в зростаючому порядку:

$$x_1 \le x_2 \le x_3 \le \ldots \le x_{n-1} \le x_n.$$

Потім визначимо медіану x_{Me} і кожному елементу $x_i (i = 1,2,\ldots,n)$ припишемо символ *«а»,* якщо $x_i < x_{\text{Me}}$, і символ *«в»*, якщо $x_i > x_{\text{Me}}$.

Одержимо послідовність із літер *«а»* та *«в»*. Позначимо через випадкову величину k — кількість серій отриманої послідовності. Доведено, що якщо нульова гіпотеза справедлива, то для досить великих n $(n > 10)$ розподіл кількості серій k є нормальним, $N(M(k);\sigma^2(k))$, де:

$$M(k) = \frac{2n_1n_2}{n_1 + n_2}; \quad \sigma(k) = \frac{2n_1n_2}{(n_1 + n_2)^{3/2}}.$$

Тоді нормована випадкова величина:

$$Z = \frac{k - \frac{2n_1n_2}{n_1 + n_2}}{\frac{2n_1n_2}{(n_1 + n_2)^{3/2}}} \to N(0;1^2). \tag{5.54}$$

Якщо ж із двох генеральних сукупностей з довільними функціями розподілу $F_1(x)$ і $F_2(x)$ витягнуті вибірки обсягами n_1 і n_2, то за допомогою критерію серій можна перевірити гіпотезу:

$$H_0 : F_1(x) = F_2(x), \tag{5.55}$$

тобто дві генеральні сукупності мають ту саму функцію розподілу. У цьому випадку результати двох вибірок поєднуються в одну неспадну послідовність, і спостереження з першої вибірки позначаються як *«а»*, а з другої — як *«в»*.

Отже, маємо таку схему застосування критерію серій:

- формулюється нуль—гіпотеза; а) H_0: вибірка випадкова;
  б) H_0: дві вибірки вибрані з однієї й тієї самої генеральної сукупності $\left(F_1(x) = F_2(x)\right)$.
- Альтернативна гіпотеза H_1, як правило, і у випадку а) і у випадку б) не формулюється при застосуванні критерію серій.
- Рівень значущості α для критерію, як правило, вибирають $\alpha = 0{,}05$, або $\alpha = 0{,}01$.
- Критерій (критеріальна статистика):

$$Z = \frac{k - \dfrac{2n_1n_2}{n_1n_2}}{\dfrac{2n_1n_2}{(n_1 + n_2)^{3/2}}},$$

де k — кількість серій.

- Критичні точки залежать від α. Оскільки альтернативні гіпотези не формулюються, то застосовується *z*-критерій з двобічною критичною областю. Маємо межі $\pm z_{\frac{\alpha}{2}}$, які відокремлюють критичні області від області прийняття гіпотези H_0. Якщо $\alpha = 0{,}05$, то критичні точки дорівнюють $\pm 1{,}96$; коли $\alpha = 0{,}01$, то — $\pm 2{,}575$. Для інших значень α критичні точки беруть із табл. 2 (додаток).

Гіпотеза H_0 приймається, якщо:

а) $|z_{\text{спост.}}| < z_{\alpha/2}$;

б) $|z_{\text{спост.}}| < z_{\alpha/2}$.

Приклад 5.39. Лікар рекомендував своїм пацієнтам, які мають зайву вагу, ліки «а» та «в». При цьому щоразу фіксував початкову вагу пацієнта в кг. У результаті він дістав таблицю:

а	66,5	83,0	67,8	75,6	81,6	98,0	57,6	100,7	59,7	73,3	100,3	92,1	$(n_1 = 12)$
в	81,4	73,1	71,0	70,1	66,3	59,4	73,8	72,2	73,5	102,1	71,8	—	$(n_2 = 11)$

(Очевидно, що якщо лікар пропонував пацієнтові одні з ліків «*а*» або «*в*» випадковим чином, то отримана послідовність із літер «*а*» та «*в*» буде випадковою).

Перевірити за допомогою критерію серій для рівнів значущості $\alpha = 0{,}05$ та $\alpha = 0{,}1$ гіпотезу про випадковість у призначенні ліків, тобто про вплив ліків «а» та «в» на зміну ваги пацієнтів.

Р о з в ' я з а н н я.

Крок 1. Сформулюємо гіпотезу H_0: досліджувана вибірка випадкова.

Крок 2. Утворимо спадну послідовність, у якій символом «*а*» будемо позначати значення з першої вибірки, символом «*в*» — значення із другої вибірки:

№ з/п	1	2	3	4	5	6	7	8	9	10	11
x_i	57,6	59,4	59,7	66,3	66,5	67,8	70,1	71,0	71,8	72,2	73,1
а;в	*а*	*в*	*а*	*в*	*а*	*а*	*в*	*в*	*в*	*в*	*в*
	1	2	3	4	5				6		

12	13	14	15	16	17	18	19	20	21	22	23
73,5	73,7	73,8	75,6	81,4	81,6	83,0	92,1	98,0	100,3	100,7	102,1
в	*а*	*в*	*а*	*в*	*а*	*а*	*а*	*а*	*а*	*а*	*в*
	7	8	9	10			11				12

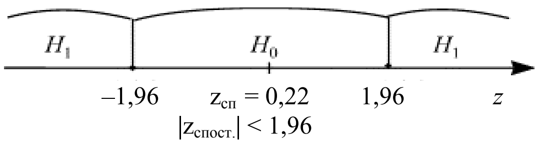
Загальна кількість серій $k = 12; n_1 = 12; n_2 = 11.$

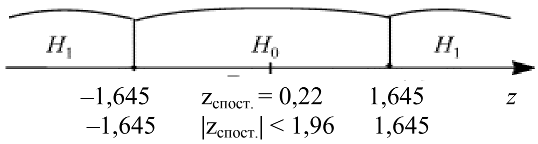
Крок 3. За формулою критерію (5.54) обчислимо спостережуване його значення:

$$Z_{\text{спост.}} = \frac{12 - 2 \cdot 12 \cdot 11/(12+11)}{2 \cdot 12 \cdot 11/(12+11)^{3/2}} \approx 0{,}22.$$

Крок 4. Для $\alpha = 0{,}05 \quad Z_{kp}(0{,}05) = 1{,}96.$ Для $\alpha = 0{,}1 \quad Z_{kp}(0{,}1) = 1{,}645.$

Оскільки в обох випадках $|Z_{\text{табл.}}| < Z_{kp}$, тому немає підстав відхиляти гіпотезу H_0 про випадковість вибірки, отже, вважаємо, що лікар призначав ліки *«а»* і *«в»* випадковим чином.





Немає підстав відхилити H_0 про випадковість вибірки.

5.10.7. Критерій знаків

Критерій знаків — один з найпростіших критеріїв непараметричної статистики, застосовується для **неперервних розподілів**. За допомогою цього критерію перевіряється нульова гіпотеза про те, що дві вибірки, вибрані з однієї й тієї самої генеральної сукупності, мають одну й ту саму функцію розподілу. При цьому використовуються відомості про характер розподілу (наприклад, не передбачається, що він нормальний).

За допомогою критерію знаків використовують знаки розбіжностей між двома числами, а не їхню кількісну міру.

Варто помітити, що якщо гіпотеза $H_0 : F(x) = F(y)$ справедлива, то немає підстав відхиляти гіпотезу $H_0 : \overline{X} = \overline{Y}$. Це свідчить про те, що критерій знаків можна застосовувати для перевірки гіпотези про рівність двох середніх генеральних сукупностей з невідомими законами розподілу. Нехай є дві вибірки $(x_1; x_2, ..., x_n)$ і $(y_1; y_2, ..., y_n)$ однакового обсягу *n*. Потрібно перевірити гіпотезу H_0: вибірки витягнуті з однієї й тієї самої сукупності, тобто $F(x) = F(y)$.

Передбачається, що випадкові величини *X* і *Y*, значення яких спостерігаються в *i*-тому випробуванні, незалежні одна від одної, а послідовні *n* спостережень незалежні між собою. Вибірки ранжирувані, тобто $x_1 < x_2 < ... < x_n$ і $y_1 < y_2 < ... < y_n$. Позначимо $x_i - y_i = r_i$. Чи можна вважати розбіжність між x_i і y_i значущою, істотною? Різниці $r_i = 0$ виключаємо.

Досліджуємо знаки різниць r_i і знайдемо кількість тих знаків, яких менше. Нехай їх виявилося r.

У випадку, коли нульова гіпотеза справедлива, різниці $x_i - y_i = r_i$ будуть симетрично розподілені відносно нуля, тобто знаки «+» та «–» рівноймовірні:

$$P("+") = P("-") = \frac{1}{2}, \text{ або } P(x - y > 0) = P(x - y < 0) = \frac{1}{2}.$$

Зрозуміло, що число r -дискретна випадкова величина, розподілена за біноміальним законом з параметрами *n* і $p = \frac{1}{2}$, тобто:

$$P_n(r) = C_n^r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{n-r} = C_n^r \left(\frac{1}{2}\right)^n = C_n^r 2^{-n}. \tag{5.56}$$

Критична область будується правобічна, якщо $H_1 : \overline{X} > \overline{Y}$; лівобічна, якщо $H_1 : \overline{X} < \overline{Y}$; двобічна, якщо $H_1 : \overline{X} \neq \overline{Y}$.

Позначимо через $r_\alpha -$ найменше значення *r*, для якого $P_n(r) \leq \alpha.$ Тоді гіпотеза H_0 відхиляється, коли $r_{\text{спост.}} \leq r_\alpha$, а для $r_{\text{спост.}} > r_\alpha$ немає підстав відхилити гіпотезу H_0.

Для перевірки гіпотез застосовується спеціальна таблиця критичних значень кількості знаків *r*, що відповідають заданому рівню α і обсягу вибірки *n* (додаток 12).

Величина r — розподілена за біноміальним законом. Виходить, що $M(r) = np = \frac{n}{2}$ і $D(r) = npq = \frac{n}{4}$. Оскільки для $n \to \infty$ біноміальний розподіл згідно з теоремою Муавра—Лапласа наближається до нормального розподілу, то $r \to N\left(\frac{n}{2}; \frac{n}{4}\right)$.

Для невеликих значень n і r імовірність (5.56) легко обчислити безпосередньо за формулою, але для більших $n\,(n > 30)$ і r можна використовувати нормальний розподіл.

Приклад 5.40.[1] У таблиці наведено результати обстеження 20 родин, що мають однаковий дохід. З'ясувалося, яку частину доходу кожна родина витрачає на транспорт і культурні потреби. Необхідно для рівня значущості 0,1 перевірити гіпотезу про те, що в генеральній сукупності середні частки витрат на транспорт і культурні потреби однакові. Завдання виконати за допомогою критерію знаків.

№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %	№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %
1	5,3	6,8	11	4,9	8,3
2	5,1	6,7	12	4,7	10,5
3	10,9	4,1	13	5,6	8,1
4	4,7	8,3	14	8,3	2,1
5	4,3	5,9	15	6,6	7,2
6	5,7	6,1	16	7,3	12,1
7	5,4	12,0	17	4,2	6,1
8	12,6	10,9	18	5,0	7,8
9	6,2	13,1	19	4,9	10,5
10	3,1	6,8	20	8,3	5,9

Р о з в ' я з а н н я.

Крок 1. Сформулюємо нульову гіпотезу $H_0 : F(x) = F(y)$, тобто частки витрат кожної родини на транспорт і культурні потреби однакові.

Позначимо знаком «+» більшу частку витрат на культурні потреби, а знаком «–» більшу частку витрат на транспорт, дістанемо таку послідовність знаків: + + – + + + + + – + + + + + + – + + + + + –. Кількість знаків «–» дорівнює 4, тобто $r = 4$.

Знайдемо з таблиці додаткf 12 критичних значень кількості знаків за заданим рівнем значущості $\alpha = 0{,}05$ та обсягу вибірки $n = 20$ критичне значення:

$$r_{\alpha;n} = r_{0,05;20} = 5.$$

Оскільки $r_{спост.} = 4 < 5$, то нульова гіпотеза відхиляється на користь альтернативної, тобто це означає, що частка витрат кожної родини на культурні потреби перевищує частку витрат на транспортні послуги.

5.10.8. Критерій згоди В. І. Романовського

Більш простим методом оцінювання близькості емпіричного розподілу до нормального є метод, який запропонував В. І. Романовський. Він використовував випадкову величину χ^2, але за його методом немає потреби знаходити значення $\chi^2_{крит.}$ із таблиць значень χ^2. За критерій (статистику) В. І. Романовський обрав величину:

$$Y_{Ром} = \frac{\chi^2 - k_{св}}{\sqrt{2k_{св}}}, \tag{5.57}$$

де $k_{св} = \nu$ — число ступенів вільності.

Якщо $|Y_{Ром}| \le 3$, то несуттєвою є розбіжність між емпіричним і теоретичним розподілом і емпіричний розподіл можна вважати приблизно нормальним; якщо ж $|Y_{Ром}| > 3$, то нульова гіпотеза про близькість емпіричного і теоретичного розподілів відхиляється.

[1] Приклад із посібника : В. *С. Мхотарян, Л. І. Трошин*. Математична статистика. — М. : МЄСИ. — С. 35—36.

Приклад 5.41. Маємо згруповані дані про денний виторг у магазині електротоварів (тис. грн).

Суми продажу	Кількість одиниць продажу
190—200	10
200—210	26
210—220	56
220—230	64
230—240	30
240—250	14

Перевірити нульову гіпотезу H_0 про те, що сума продажу (*X*) є випадковою величиною, яка розподілена за нормальним законом. Рівень значущості α прийняти за 0,5.

Р о з в ' я з а н н я. Для перевірки нульової гіпотези визначимо значення x_i^* — середини інтервалів та обчислимо точкові оцінки математичного сподівання *a* і середнього квадратичного відхилення σ^2 нормально розподіленої випадкової величини *X*, оскільки з умови слідує, що точкові оцінки гіпотетичного нормального закону невідомі. Отже, перевіряємо гіпотезу $H_0 : X \to N(a;\sigma^2)$. Для зручності розрахунки проведемо в таблиці

x_i^*	n_i	$x_i^* \cdot n_i$	$x_i^* - \bar{x}_в$	$(x_i^* - \bar{x}_в)^2 \cdot n_i$
195	10	1950	–26	6760
205	26	5330	–16	6656
215	56	12 040	–6	2016
225	64	14 400	16	1024
235	30	7050	14	5880
245	14	3430	24	8064
Σ	200	44 200	—	30 400

де $\bar{x}_B = \frac{1}{n}\sum_{i=1}^{6} x_i^* \cdot n = \frac{1}{200} \cdot 44\,200 < 221.$

Тоді $a \approx 221,\quad \sigma_в^2 = \frac{1}{200} \cdot 30\,400 = 152,$ а $\sigma_в = \sqrt{152} \approx 12{,}33.$

Зауважимо, що проведені розрахунки можна було б проводити з меншими числами, відзначаючи закономірності: усі x_i^* кратні 5, а частоти n_i — кратні 2. Можна було б розглянути $x_i' = \frac{x_i^*}{5}$, а $n_i' = \frac{n_i}{2}$. Тоді:

$$x_в' = \frac{1}{\sum_{i=1}^{6} n_i'} \cdot \sum_{i=1}^{6} x_i' \cdot n_i' = \frac{1}{100} \cdot 4420 = 44{,}2\,;$$

$$(\sigma_в')^2 = \frac{1}{\sum_{i=1}^{6} n_i'} \cdot \sum_{i=1}^{6} (x_i' - x_в')^2 \cdot n_i' = \frac{1}{100} \cdot 608 = 6{,}08\,,$$

тому $a \approx \bar{x}_в = 5 \cdot x_в' = 5 \cdot 44{,}2 = 221$, $\sigma_в^2 = 25 \cdot (\sigma')^2 = 25 \cdot 6{,}08 = 152$.

Тепер обчислимо теоретичні ймовірності p_i потрапляння випадкової величини $X \to N(221;152)$ у частинні інтервали ($x_i; x_{i+1}$) за формулою $p_i = \Phi(z_{i+1}) - \Phi(z_i)$, де $z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$.

Після цього обчислимо $\chi^2_{спост.} = \sum_{i=1}^{6} \frac{(n_i - n_i')^2}{n_i'}$. Розрахунки проводимо в таблиці:

Інтервали $(x_i; x_{i+1})$	Частоти n_i	Нормовані інтервали $(z_i; z_{i+1})$, де $z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$	$\Phi(z_i)$	$\Phi(z_{i+1})$	$p_i = \Phi(z_{i+1}) - \Phi(z_i)$
190—200	10	(– ∞; —1,7)	– 0,5	– 0,4554	0,0446
200—210	26	(–1,7; —0,89)	– 0,4554	– 0,3133	0,1421
210—220	56	(– 0,89; —0,08)	– 0,3133	– 0,0319	0,2814
220—230	64	(– 0,08; 0,73)	– 0,0319	0,2673	0,2992
230—240	30	(0,73; 1,54)	0,2673	0,4382	0,1709
240—250	14	(1,54; ∞)	0,4382	0,5	0,0618
Σ	$n = 200$	—	—	—	1

$n_i' = n \cdot p_i$	$(n_i - n_i')$	$\frac{(n_i - n_i')^2}{n_i'}$
8,92	1,08	0,1308
28,42	–2,42	0,206
56,28	– 0,28	0,0014
59,84	4,16	0,2892
34,18	– 4,18	0,5112
12,36	1,64	0,2176
200		$\chi^2_{спост.} = 1{,}3562$

Зауважимо, що найменше значення $\frac{x_i - \bar{x}_в}{\sigma_в} = \frac{190 - 221}{12{,}33} \approx -2{,}514$ замінено на «– ∞», а найбільше значення $\frac{x_i - \bar{x}_в}{\sigma_в} = \frac{250 - 221}{12{,}33} \approx -2{,}352$ замінено на «+ ∞».

Маємо, $\chi^2_{спост.} = 1{,}3562$, а число ступенів вільності $k_{св} = v = S - r - 1 = 6 - 2 - 1 = 3$.

За критерієм Романовського $|Y_{Ром}| = \left|\frac{1{,}3562 - 3}{\sqrt{6}}\right| \approx \left|\frac{-1{,}6438}{2{,}4495}\right| \approx 0{,}67 < 3$, тому немає підстав відхилити нульову гіпотезу. Отже, математичною моделлю заданого вибіркового розподілу можна вважати нормальний закон розподілу.

Зауваження. Відношення Романовського має підґрунтям те, що $M(\chi^2) = k_{св}$, а $D(\chi^2) = \sigma^2 = 2k_{св}$. Тому ймовірність відхилення χ^2 на $\sqrt{2k_{св}}$ близька до 1.

5.10.9 Критерій згоди Б. С. Ястремського

Як і критерій Романовського, критерій Ястремського:

$$Y_{ястр} = \frac{|C - k|}{\sqrt{2k + 4\Theta}} \tag{5.58}$$

застосовується без звернення до таблиць розподілу χ^2. У формулі (5.58) k — кількість груп, Θ — величина, яка залежить від k і C. Якщо $k < 20$ і $C = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i' \cdot (1 - p_i)}$, тоді $\Theta = 0{,}6$.

Якщо $|Y_{ястр.}| > 3$, то гіпотеза H_0 відхиляється; якщо ж $|Y_{ястр.}| \le 3$, то H_0 приймається.

Розв'яжемо приклад 5.41 за критерієм Ястремського. Для цього розрахунки виконуємо в наступній таблиці, де p_i — знайдені під час розв'язання прикладу 5.41.

p_i	$1-p_i$	n'_i	$\frac{(n_i-n'_i)^2}{n'_i}$	$\frac{(n_i-n'_i)^2}{n'_i\cdot(1-p_i)}$
0,0446 0,1421 0,2814 0,2992 0,1709 0,0618	0,9554 0,8579 0,7186 0,7008 0,8291 0,9382	8,92 28,42 56,28 59,84 34,18 12,36	0,1308 0,206 0,0014 0,2892 0,5112 0,2176	0,1369 0,2401 0,0019 0,4127 0,6166 0,2319
Σ				$C\approx 1{,}64$

тоді $\left|Y_{\text{ястр.}}\right| = \left|\frac{1{,}64-6}{\sqrt{12+4\cdot 0{,}6}}\right| \approx \left|\frac{-4{,}36}{3{,}7947}\right| \approx 1{,}143 < 3$.

Отже, немає підстав відхилити нульову гіпотезу про нормальний закон розподілу.

5.10.10. Критерій згоди А. Н. Колмогорова

Критерій Колмогорова застосовується для перевірки гіпотез про закони розподілу неперервних випадкових величин. На відміну від критерію Пірсона, де порівнюються емпіричні та теоретичні частоти, за критерієм Колмогорова порівнюються емпірична та теоретична функції розподілу, а також у критерії Колмогорова параметри функції розподілу $F_0(x)$ відомі, тимчасом як у критерії Пірсона теоретичні значення параметрів функції розподілу, що припускається, можна визначити за даними вибірки.

Нехай неперервна випадкова величина X має функцію розподілу $F(x)$, а з генеральної сукупності взято випадкову вибірку об'єму $n\ (n \ge 50)$. Потрібно перевірити нульову гіпотезу H_0 про рівність функції розподілу (емпіричної і теоретичної), $H_0: F(x) = F_0(x)$.

Схема перевірки H_0 за критерієм Колмогорова:

Крок 1. Результати спостережень розміщують у зростаючому (або неспадному) порядку: $x_1 \le x_2 \le x_3 \le ... \le x_n$, або подають у вигляді інтервального варіаційного ряду; потім знаходять емпіричну функцію розподілу $F_n^*(x) = \frac{n_x}{n}$ та значення припущеної функції розподілу $F(x)$.

Крок 2. За критерій обирається величина $\lambda = D\cdot\sqrt{n}$, де

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

λ називають критерієм згоди Колмогорова.

Крок 3. Обчислюємо $\lambda_{\text{спост.}}$

Крок 4. За заданим рівнем значущості α із таблиці критичних точок (квантилей) розподілу Колмогорова $K(\lambda)$ знаходимо $\lambda_{\text{крит}} = \lambda_\alpha$.

Крок 5. Робимо висновок :

- якщо $D\cdot\sqrt{n} \ge \lambda_\alpha$, то H_0 відхиляється;
- якщо $D\cdot\sqrt{n} < \lambda_\alpha$, то H_0 приймається.

Приклад 5.42. Є вибіркові дані про кількість угод, які підписані між брокерськими фірмами та конторами міста протягом місяця:

Кількість угод $(x_i - x_{i+1})$	0—10	10—20	20—30	30—40	40—50
Кількість брокерських фірм та контор (n_i)	22	25	12	8	3

Використовуючи критерії згоди Пірсона і Колмогорова, перевірити гіпотезу про нормальний закон розподілу для рівня значущості $\alpha = 0{,}05$.

Р о з в ' я з а н н я.
Розрахунки здійснимо в таблиці:

Інтервали	n_i	x_i^* — середини інтервалів	$x_i^* \cdot n_i$	$\left(x_i^*\right)^2 \cdot n_i$
0—10	22	5	110	550
10—20	25	15	375	5625
20—30	12	25	300	7500
30—40	8	35	280	9800
40—50	3	45	135	6075
Σ	70	—	1200	29 550

$\overline{x}_в = \frac{1}{n} \cdot \sum_{i=1}^{5} x_i^* \cdot n_i = \frac{1}{70} \cdot 120 \approx 17{,}1429$;

$\sigma_в^2 = \frac{1}{70} \cdot 29\,550 - (17{,}1429)^2 \approx 422{,}14285 - 293{,}87755 = 128{,}2653$,

тоді $\sigma_в = \sqrt{128{,}2653} \approx 11{,}3254$.

Крок 1. Гіпотеза H_0: $X \to N$ (17,1429; 128,2653).

Крок 2. Застосуємо критерій Пірсона $\chi^2 = \sum_{i=1}^{5} \frac{\left(n_i - n_i'\right)^2}{n_i'}$.

Крок 3. Складемо розрахункову таблицю для зручнішого обчислення $\chi^2_{спост.}$:

Інтервали	n_i	$z_i = \frac{x_i - \overline{x}_в}{\sigma_в}$	$z_{i+1} = \frac{x_{i+1} - \overline{x}_в}{\sigma_B}$	$\Phi(z_i)$	$\Phi(z_{i+1})$	$p_i = \Phi(z_{i+1}) - \Phi(z_i)$
0—10	22	$-\infty$	– 0,63	– 0,5	– 0,23565	0,26435
10—20	25	– 0,63	0,25	– 0,23565	0,09871	0,33436
20—30	12	0,25	1,14	0,09871	0,37286	0,27415
30—40	8	1,14	2,02	0,37286	0,47831	0,10545
40—50	3	2,02	∞	0,47831	0,5	0,02169
Σ	70	—	—	—	—	—

	$n_i' = np_i$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
	18,5045	12,21852	0,6603
	23,4052	2,54339	0,10867
	19,1905	51,70329	2,6942
	7,3815	0,38254	0,05182
	1,5183	2,19543	1,44598
$\sum$	70	—	$\chi^2_{спост} = 4{,}96097$

Крок 4. Знайдемо $\chi^2_{крит} = \chi^2(0{,}05; 5-2-1) = \chi^2(0{,}05; 2) = 6$.

Оскільки $\chi^2_{спост} = 4{,}96097 < \chi^2_{крит} = 6$, то підстав відхилити гіпотезу H_0 немає.

Крок 5. Застосуємо критерій Колмогорова. Обчислимо:

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

Для обчислення $\lambda_{спост}$ складемо таблицю:

Інтервали	n_i	Накопичені частоти	Нормовані інтервали $(z_i; z_{i+1})$	$F^*(x_{i+1})$	Теоретична функція розподілу $F(x) = 0{,}5 + \Phi(z_{i+1})$	$\left\|F^*(x) - F(x)\right\|$
0—10	22	22	$(-\infty; -0{,}63)$	0,31429	0,26435	0,04994
10—20	25	47	$(-0{,}63; 0{,}25)$	0,67143	0,59871	0,07272
20—30	12	59	$(0{,}25; 1{,}14)$	0,84286	0,87286	
30—40	8	67	$(1{,}14; 2{,}02)$	0,95714	0,97831	0,03
40—50	3	70	$(2{,}02; \infty)$	1	1	0,02117
						0
$\sum$	70	—	—			

$\max\left|F^*(x) - F(x)\right| = 0{,}07272$, отже, $D \approx 0{,}07$, тоді $\lambda_{\text{крит}(0,05)} = 1{,}358$; $\lambda_{\text{спост}} = 0{,}07 \cdot \sqrt{70} \approx 0{,}5857$.

Оскільки $\lambda_{\text{крит}} = 1{,}358 > \lambda_{\text{спост}} = 0{,}5857$, то немає підстав відхилити нульову гіпотезу H_0 про нормальний закон розподілу.

§ 6. ЕЛЕМЕНТИ КОРЕЛЯЦІЙНОГО ТА РЕГРЕСІЙНОГО АНАЛІЗУ

6.1. ПАРНА ЛІНІЙНА РЕГРЕСІЯ

Виходячи із сутності кореляційної залежності між двома змінними величинами [6], яка полягає в залежності умовного математичного сподівання однієї з них від значень, які приймає інша, цю залежність можна записати у вигляді:

$$M_x(Y) = \varphi(x), \text{ або } M_y(X) = \Psi(y). \qquad (6.1)$$

Перше з рівнянь — рівняння регресії *Y* на *X*, друге — *X* на *Y*.

Статистичні зв'язки між двома змінними базуються на основі регресійного і кореляційного аналізів. Головною метою першого з них є вивчення і встановлення форми залежності між змінними (залежності (6.1)), другого — виявлення зв'язку між випадковими величинами та оцінювання тісноти цього зв'язку.

У практиці статистики для побудови залежностей (6.1) маємо вибірки пар значень $(x_i; y_i)$, обмеженого обсягу n, але закони розподілу двовимірної випадкової величини $(X;Y)$ невідомі. Тому надалі будемо розглядати співвідношення (6.1) як оціночні, побудовані за вибірками $(x_i; y_i)$ рівняння регресії.

Як правило, для побудови вибіркових ліній регресії *Y* на *X* використовується метод найменших квадратів, для реалізації якого обирається співвідношення вигляду:

$$y_x = \varphi(x, b_0, b_1, ..., b_k), \qquad (6.2)$$

де y_x — умовна середня змінної *Y*, якщо змінна *X* набуває фіксованого значення $X = x$, а $b_0, b_1, ..., b_k$ — параметри, які визначають функцію φ.

Аналогічно можна розглядати вибіркове рівняння регресії *X* на *Y*:

$$x_y = \psi(y, c_o, c_1, c_2, ..., c_k). \qquad (6.3)$$

Сутність параметрів $c_0, c_1, ..., c_k$, як і для (6.2).

У подальшому, під час реалізації методу найменших квадратів (МНК) для побудови рівнянь регресії (ліній регресії) будуть використовуватися вибіркові аналоги початкових і центральних моментів випадкової величини.

Відповідні вибіркові числові характеристики для одновимірних випадкових величин (В.В.) розглянуто в § 2. Для розглядуваних тут двовимірних В.В. аналогічно до § 2 будуються вибіркові середні для X та Y:

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (6.4)$$

і виправлені дисперсії для В.В. X та Y:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2; \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2. \qquad (6.5)$$

Найбільш застосовувані в кореляційному аналізі числові характеристики зв'язку між двома В.В. — це вибірковий кореляційний момент (коваріація COV) ρ_{xy} і вибірковий коефіцієнт кореляції r_{xy}, які є мірою лінійного зв'язку між випадковими величинами *X* та *Y*. Ці величини обчислюються відповідно за формулами:

$$\rho_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad (6.6)$$

або $\rho_{xy} = \overline{xy} - \bar{x}\,\bar{y}$, де $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$;

$$r_{xy} = \frac{\frac{n-1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{6.7}$$

або $r_{xy} = \frac{\rho_{xy}}{s_x s_y}$.

Слід також відмітити наявність співвідношень, які випливають із (6.6), (6.7). Якщо $y = u + \mathrm{v}$, то:

$$\rho_{xy} = \rho_{xu} + \rho_{x\mathrm{v}}, \tag{6.8}$$

якщо $y = az, \quad a = \mathrm{const}$, то:

$$\rho_{xy} = a\rho_{xz} \tag{6.9}$$

при $y = \mathrm{const} \qquad \rho_{xy} = 0$.

Повернемось до розгляду рівняння регресії (6.2). У багатьох конкретних випадках залежності між випадковими величинами *(X ,Y)* для генеральної сукупності обирають $\varphi(x, b_0, b_1, ..., b_k)$ як лінійну функцію. Обмежимось на початку лінійною моделлю такої залежності:

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{6.10}$$

Параметри β_0, β_1 — невідомі, добавка ε — випадкова величина, розподілена нормально і має математичне сподівання $M(\varepsilon) = 0$ та дисперсію $D(\varepsilon) = \sigma^2$. Ця добавка ε ураховує те, що вибіркові значення $(x_i y_i)$ можуть не потрапляти на лінію регресії (6.10).

Співвідношення (6.10) називаються теоретичною лінійною регресійною моделлю; β_0, β_1 теоретичними параметрами моделі. Для їх знаходження треба знати всі значення В.В. (*X*, *Y*) генеральної сукупності, що практично неможливо. Тому, як правило, користуються обмеженими статистичними даними (x_i, y_i) з генеральної сукупності (*X*, *Y*) і будують емпіричне рівняння регресії:

$$y = b_0 + b_1 x, \tag{6.10a}$$

параметри якого b_0 та b_1 знаходять із системи рівнянь методу найменших квадратів:

$$\begin{cases} b_1 \sum_{i=1}^{n} x_i + n b_0 = \sum_{i=1}^{n} y_i, \\ b_1 \sum_{i=1}^{n} x_i^2 + b_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i. \end{cases} \tag{6.11}$$

Поділивши кожне з рівнянь системи (6.11) на *n*, з урахуванням можливих частот n_i для значень В.В. *X* та n_j для значень В.В. *Y* систему можна переписати у такому вигляді:

$$\begin{cases} b_1 \bar{x} + b_0 = \bar{y} \\ b_1 \overline{x^2} + b_0 \bar{x} = \overline{xy} \end{cases}, \tag{6.12}$$

де $\bar{x} = \frac{\sum_{i=1}^{l} x_i n_i}{n}$; $\quad \bar{y} = \frac{\sum_{j=1}^{m} y_j n_j}{n}$; $\quad \overline{xy} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{m} x_i y_j n_{ij}}{n}$; $\overline{x^2} = \frac{\sum_{i=1}^{l} x_i^2 n_i}{n}$; $\quad n = \sum_{i=1}^{l} n_i = \sum_{j=1}^{m} n_j$,

n_{ij} — відповідні частоти пар значень (x_i, y_j).

Розв'язуючи систему (6.11), знайдемо:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_x^2}, \tag{6.13}$$

де s_x^2 — вибіркова дисперсія випадкової величини *X*. З вигляду чисельника дробу (6.13) і (6.6) маємо вираз для знаходження вибіркового кореляційного моменту $\rho_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$; отже, $b_1 = \frac{\rho_{xy}}{s_x^2}$. Запишемо останній вираз у вигляді: $b_1 \cdot \frac{s_x}{s_y} = \frac{\rho_{xy}}{s_x s_y} = r_в$, де (згідно 6.7) $r_в$ — вибірковий коефіцієнт кореляції, серед властивостей якого $|r_в| \le 1.$ Тепер емпіричне рівняння регресії (6.10а) без урахування добавки ε, а також того, що $b_0 = \overline{y} - b_1\overline{x}$, можна записати у вигляді:

$$y - \overline{y} = r_в \frac{s_y}{s_x}(x - \overline{x}), \tag{6.14}$$

де $r_в \frac{s_y}{s_x}$ — вибірковий коефіцієнт регресії, або просто коефіцієнт регресії Y на X, який показує на скільки одиниць в середньому зміниться змінна Y, якщо змінна X збільшиться на одиницю.

Аналогічно можна побудувати лінійну регресійну модель X на Y у вигляді:

$$x - \overline{x} = r_в \frac{s_x}{s_y}(y - \overline{y}) \qquad .(6.15)$$

Слід зауважити, що лінії (6.14) і (6.15) перетинаються в точці $(\overline{x}; \overline{y})$, а вибірковий коефіцієнт кореляції $r_в$ (коефіцієнт кореляції) є показником тісноти зв'язку випадкових величин X і Y: чим ближче він до одиниці, за абсолютною величиною, тим ближче цей зв'язок до лінійного.

Приклад 6.1. У результаті деякого експерименту виявлені пари показників, які розташовані в табл. 6.1.

Таблиця 6.1

ДВАДЦЯТЬ П'ЯТЬ СПОСТЕРЕЖЕНЬ ЗМІННИХ *X* І *Y*

y	*x*	*y*	*x*
10,98	35,3	9,57	39,1
11,13	29,7	10,94	46,8
12,51	30,8	9,58	48,5
8,40	58,8	10,09	59,3
9,27	61,4	8,11	70,0
8,73	71,3	6,83	70,0
6,36	74,4	8,88	74,5
8,50	76,7	7,68	72,1
7,82	70,7	8,47	58,1
9,14	57,5	8,86	44,6
8,24	46,4	10,36	33,4
12,19	28,9	11,08	28,6
11,88	28,1		

Скласти рівняння лінії регресії.

Розв'язання. Кожній парі $(x_i; y_i)$ можна поставити у відповідність точку на площині *oxy*. Таким чином одержимо кореляційне поле залежності *Y* від *X*.

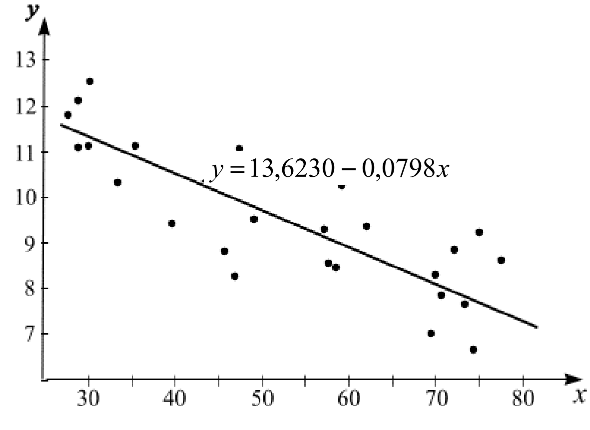


Рис 5.2

Обчислимо значення величин, що входять в емпіричне рівняння регресії (6.14), (6.15) для даних табл. 6.1.

$$\overline{x} = \frac{1}{n}\sum x_i = \frac{1}{25}(35,3 + 29,7 + ... + 28,6) = 52,6;$$

$$\overline{y} = \frac{1}{n}\sum y_i = \frac{1}{25}(10,98 + 11,13 + ... + 11,08) = 9,424;$$

$$\overline{xy} = \frac{1}{n}\sum x_i y_i = \frac{1}{25}(10,98 \cdot 35,3 + 11,13 \cdot 29,7 + ... + 11,08 \cdot 28,6) = 472,86;$$

$$\overline{x^2} = \frac{1}{n}\sum x_i^2 = \frac{1}{25}\left(35,3^2 + 29,7^2 + ... + 28,6^2\right) = 3052,94;$$

$$\overline{y^2} = \frac{1}{n}\sum y_i^2 = \frac{1}{25}\left(10,98^2 + 11,13^2 + ... + 11,08^2\right) = 91,38;$$

$$\sigma_x^2 = \overline{x^2} - \overline{x}^2 = 3052,94 - 52,6^2 = 286,18; \quad s_x \approx 16,9168;$$

$$\sigma_y^2 = \overline{y^2} - \overline{y}^2 = 91,38 - 9,424^2 = 2,5682; \quad s_y \approx 1,6026;$$

$$\rho_{xy} = \overline{xy} - \overline{x} \cdot \overline{y} = 472,86 - 52,6 \cdot 9,424 = -22,8424;$$

$$r_в = \frac{\rho_{xy}}{s_x s_y} = -0,8426.$$

Підставляючи дані в рівняння (6.14), дістанемо лінію регресії Y на X: $y = -0,07982x + 13,623$, яка зображена на рис 6.1; коефіцієнт кореляції $r_в = -0,8426$, що підкреслює достатньо тісний ($|r_в|$ — близький до 1) лінійний зв'язок між X і Y.

У табл. 6.2 показано різниці між істинними значеннями y_i і одержаних y з рівняння регресії.

Таблиця 6.2

РЕЗУЛЬТАТИ СПОСТЕРЕЖЕНЬ, РОЗРАХУНКОВІ ЗНАЧЕННЯ І ЗАЛИШКИ

y_i	$\tilde{y}_i$	$e_i = y_i - y$	y_i	$\tilde{y}_i$	$e_i = y_i - \tilde{y}_i$
10,98	10,81	0,17	11,88	11,38	0,50
11,13	11,25	–0,12	9,57	10,50	–0,93
12,51	11,17	1,34	10,94	9,89	1,05
8,40	8,93	–0,53	9,58	9,75	–0,17
9,27	8,72	0,55	10,09	8,89	1,20
8,73	7,93	0,80	8,11	8,03	0,08
6,36	7,68	–1,32	6,83	8,03	–1,20
8,50	7,50	1,00	8,88	7,68	1,20
7,82	7,98	–0,16	7,68	7,87	–0,19
9,14	9,03	0,11	8,47	8,98	–0,51
8,24	9,92	–1,68	8,86	10,06	–1,20
12,19	11,32	0,87	10,36	10,96	–0,60
			11,08	11,34	–0,26

Невеликі за модулем різниці $e_i = y_i - y$ вказують на досить точне, наближене рівняння регресії до результатів експерименту.

Виконання розрахунків для одержання емпіричного рівняння регресії у прикладі 6.1 потребує великої кількості обчислень, тому доцільно дані групувати і заносити їх у кореляційні таблиці.

Приклад 6.2. Розглянемо дані добового виготовлення продукції у (m) і величини основних виробничих фондів (ОВФ) x (млн грн) для сукупності з 50 однотипних підприємств [2]. У таблиці 6.3 наведено згруповані дані.

Таблиця 6.3

Величина ОВФ (X)	Середини інтервалів	Добові виробітки продукції, т (Y)					Усього n_i	Групова середня, т $\bar{y}_i$
	y_i	7—11	11—15	15—19	19—23	23—27		
	x_i	9	13	17	21	25		
20—25	22,5	2	1	—	—	—	3	10,3
25—30	27,5	3	6	4	—	—	13	13,3
30—35	32,5	—	3	11	7	—	21	17,8
35—40	37,5	—	1	2	6	2	11	20,3
40—45	42,5	—	—	—	1	1	2	23,0
Усього n_i		5	11	17	14	3	50	—
Групова середня x_j, млн грн		25,5	29,3	31,9	35,4	39,2	—	—

Скласти рівняння лінії регресії.

Р о з в ’ я з а н н я.

Обчислемо параметри рівняння регресії:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{l} x_i n_i = \frac{1}{50}(22,5\cdot 3 + 27,5\cdot 13 + 32,5\cdot 21 + 37,5\cdot 11 + 42,5\cdot 2) = 32,1;$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{l} x^2{}_i n_i = \frac{1}{50}\left(22,5^2\cdot 3 + 27,5^2\cdot 13 + 32,5^2\cdot 21 + 37,5^2\cdot 11 + 42,5^2\cdot 2\right) = 1052,25;$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} y_j n_j = \frac{1}{50}\left(9\cdot 5 + 13\cdot 11 + 17\cdot 17 + 21\cdot 14 + 25\cdot 3\right) = 16,92;$$

$$\overline{y^2} = \frac{1}{n}\sum_{j=1}^{l} y_j^2 n_j = \frac{1}{50}\left(9^2\cdot 5 + 13^2\cdot 11 + 17^2\cdot 17 + 21^2\cdot 14 + 25^2\cdot 3\right) = 304,52;$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{l} x_i y_j n_{ij} = \frac{1}{50}\left(22,5\cdot 9\cdot 2 + 22,5\cdot 1\cdot 13 + \ldots + 42,5\cdot 1\cdot 21 + 42,5\cdot 1\cdot 25\right) = 557,9;$$

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2 = 1052,25 - 1030,41 = 21,84,\ s_x \approx 4,6733;$$

$$\sigma_y^2 = \overline{y^2} - \bar{y}^2 = 304,52 - 286,2864 = 18,2336,\ s_y \approx 4,27;$$

$$\rho_{xy} = \overline{xy} - \bar{x}\cdot\bar{y} = 557,9 - 32,1\cdot 16,92 = 14,768;$$

$$r_в = \frac{\rho_{xy}}{s_x s_y} = \frac{14,7686}{4,6733\cdot 4,27} = 0,74.$$

Отже, емпіричне рівняння регресії Y на X: $y = 0,6761x - 4,79$, а X на Y: $x = 0,81y + 16,70$.

З першого рівняння випливає, що при збільшенні ОВФ на 1 млн грн добове виробництво продукції y збільшується в середньому на 0,6761 т, а з другого рівняння регресії (X на Y) випливає, що для збільшення добового виробництва продукції на 1 т треба в середньому збільшити ОВФ на 0,81 млн грн. Слід зазначити, що вільні члени в емпіричному рівнянні регресії в деяких випадках не мають реального змісту.

Варто також зазначити, що емпіричні коефіцієнти регресії b_0 і b_1 є лише оцінками теоретичних коефіцієнтів β_0, β_1 рівняння регресії (6.10) генеральної сукупності. Зрозуміло, що реальні значення змінних можуть відрізнятися від модельних, одержаних із (6.10а). Ці відхилення виражаються різницями між реальними значеннями y_i і одержаними з модельного рівняння $y(x_i)$; $e_i = y_i - y$. Але за певних умов рівняння регресії є досить якісним інструментом для аналізу і прогнозування економічних явищ.

Повернемось до кореляційного аналізу, основне завдання якого полягає у виявленні тісноти зв'язку між випадковими величинами генеральної сукупності, розподіленої, згідно з центральною граничною теоремою, за двовимірним нормальним законом. Цей зв'язок може бути виявлений шляхом різних оцінок вибіркових коефіцієнтів кореляції, b_0, b_1.

Під час розгляду двовимірних систем випадкових величин, розподілених за нормальним законом (§), розглядалася щільність нормального розподілу двох змінних x, y.

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-a_x}{\sigma_x}\right)^2 - 2\rho\frac{x-a_x}{\sigma_x}\cdot\frac{y-a_y}{\sigma_y} + \left(\frac{y-a_y}{\sigma_y}\right)^2\right)}, \tag{6.16}$$

де a_x, a_y — математичні сподівання змінних; σ_x, σ_y — дисперсії змінних в.в. X та Y і ρ — генеральний коефіцієнт кореляції, який визначається через кореляційний момент K_{xy} (коваріацію) за формулою:

$$\rho = \frac{K_{xy}}{\sigma_x\sigma_y} = \frac{M(XY) - a_x a_y}{\sigma_x\sigma_y}. \tag{6.17}$$

Величина ρ характеризує тісноту лінійного зв'язку між випадковими змінними X, Y у генеральній сукупності.

Рівняння (6.1) для спільного нормального розподілу випадкових величин X і Y є лінійні функції

$$M_x(y) = a_y + \rho\frac{\sigma_y}{\sigma_x}(x - a_x), \tag{6.18}$$

$$M_y(x) = a_x + \rho\frac{\sigma_x}{\sigma_y}(y - a_y). \tag{6.19}$$

1. $M(\varepsilon_k) = 0$;
2. $D(\varepsilon_k) = \sigma = const$;
3. $\sigma_{\varepsilon_i\varepsilon_j} = \operatorname{cov}(\varepsilon_i\varepsilon_j) = \begin{cases} 0\, i \neq j \\ \sigma\, i = j \end{cases}$ — відсутність залежності між відхиленнями;
4. $\sigma_{\varepsilon_i\varepsilon_j} = 0$ — відхилення незалежне від незалежної змінної. Як наголошувалося вище, випадкові відхилення ε_i за вибіркою не визначаються, тому під час аналізу надійності оцінок коефіцієнтів регресії вони замінюються на відхилення $e_i = y_i - b_0 - b_1 x$, а дисперсія $D(\varepsilon_i) = \sigma^2$ замінюється її незміщеною оцінкою $s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$.

Величину $s = \sqrt{s^2}$ називають стандартною похибкою (стандартною похибкою регресії). Можна показати, що $D(b_1) = s_{b_1}^2 = \frac{s^2}{s_x^2}$; $D(b_0) = s_{b_0}^2 = \bar{x}^2 s_{b_1}^2$, тоді відповідно $s_{b_0} = \sqrt{s_{b_0}^2}$; $s_{b_1} = \sqrt{s_{b_1}^2}$ — стандартні похибки коефіцієнтів регресії.

Для практики важливим є оцінювання коефіцієнта кореляції генеральної сукупності ρ, який, як правило, невідомий, за допомогою вибіркового коефіцієнта кореляції $r_в$. Розглянемо гіпотези $H_0 : \rho = 0$ і $H_1 : \rho \neq 0$. Для їх перевірки скористаємось випадковою величиною $t = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}}$ ($r_в = r_{xy}$ — випадкова величина), яка розподілена за законом Стьюдента з $v = n - 0$ 2 ступенями вільності. Гіпотеза H_0 відкидається, якщо $|t| < t(1-a, v)$, де $t(1-a, v)$ табличне значення критерію Стьюдента для a — рівня значущості і v — кількості ступенів вільності. Для прикладу 2 маємо $t = \frac{0{,}74\sqrt{50-2}}{\sqrt{1-0{,}74^2}} = 7{,}62$.

За таблицями для $\alpha = 0{,}05$ і $v = 48$ маємо $t_{кр}(0{,}95,\ 48) = 2{,}01$, оскільки $t > t_{кр}$, то гіпотезу H_0 відкидаємо; тобто, коефіцієнт кореляції між добовим виготовленням продукції Y і ОВФ X суттєво відрізняється від нуля.

Важливим також є одержання із заданою надійністю $\gamma = 1 - \alpha$ інтервалу, в якому знаходиться (інтервал покриває) невідомий коефіцієнт кореляції двовимірної генеральної сукупності ρ. Для побудови такого інтервалу, взагалі кажучи, треба знати розподіл випадкової величини $r_В$. Із [22] відомо, що цей розподіл повільно збігається до нормального, тому для невеликих об'сяг вибірки n будують спеціальні функції, які збігаються до відомих розподілів. У більшості випадків для цього обирають Z — перетворення Фішера:

$$Z = \frac{1}{2}\ln\ \frac{1+r_в}{1-r_в}. \tag{6.20}$$

Розподіл випадкової величини Z навіть за невеликих n можна вважати нормальним з математичним сподіванням:

$$\mathrm{M}(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \tag{6.21}$$

і дисперсією:

$$D(Z) = \frac{1}{n-3}.$$

Побудуємо довірчий інтервал для $M(Z)$ за формулою:

$$z - \frac{t_\gamma}{\sqrt{n-3}} \le M(Z) \le z + \frac{t_\gamma}{\sqrt{n-3}}, \tag{6.22}$$

де t_γ — розв'язок рівняння $2\Phi(t_\gamma) = \gamma$; $\Phi(t)$ — функція Лапласа.

Для одержання меж довірчого інтервалу для ρ знаходять z з рівняння:

$$\rho = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{6.23}$$

і далі із (6.22) одержують довірчий інтервал для коефіцієнта кореляції генеральної сукупності ρ.

Приклад 6.3. Для даних прикладу 2 знайти довірчі інтервали для ρ.

Розв'язання. З попереднього висновку про значимість коефіцієнта кореляції, є підстави будувати довірчий інтервал для коефіцієнта ρ генеральної сукупності (X, Y). Із (6.20) маємо $Z = \frac{1}{2}\ln\frac{1+0{,}74}{1-0{,}74} = 0{,}9505$, далі за таблицями знаходимо $t_\gamma = 1{,}96$ з $2\Phi(t_\gamma) = 0{,}95$; за виразом (6.22) маємо $0{,}6646 < M(Z) < 1{,}2364$ і далі розв'язуючи нерівності $0{,}6646 < \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) < 1{,}2364$, маємо довірчий інтервал $0{,}581 \le \rho \le 0{,}844$, який з надійністю $\gamma = 0{,}95$ накриває генеральний коефіцієнт кореляції ρ.

На початковому етапі оцінювання параметрів регресії можна безпосередньо одержати оцінки b_1 з розгляду гіпотез $H_0 : b_1 = 0$ $H_1 : b_1 \ne 0$. У разі відхилення гіпотези H_0 можна вважати, що дійсно існує лінійний зв'язок між В.В. X, Y (b_1 — статистично значущий). Для перевірки цих гіпотез будують випадкову величину $t = \frac{b_1}{s_{b_1}} = \frac{b_1}{\sqrt{s_{b_1}^2}}$, яка розподілена за законом Стьюдента з $n - 2$ ступенями вільності. Для рівня значущості α і $n - 2$ ступенів вільності знаходять критичне значення $t_{kp}\left(\frac{\alpha}{2}, n-2\right)$. Якщо $|t| > t_{kp}\left(\frac{\alpha}{2}, n-2\right)$, нульову гіпотезу відхиляють. Аналогічно для перевірки значущості b_0 будують $t = \frac{b_0}{s_{b_0}}$.

Так, для прикладу 1 для коефіцієнта b_1 маємо: $|t| = 7{,}52$, $t_{kp}(0{,}25;23) = 2{,}069$, тому слід прийняти H_1. Для аналогічних гіпотез для коефіцієнта $b_0 |t| = 2{,}238$; також слід прийняти H_1. У прикладі 2 для коефіцієнта b_1 маємо $|t| = 10{,}94$, $t_{kp}(0{,}25;23) = 2{,}015$. Коефіцієнт b_1 — значущий.

Повернемось до коефіцієнта b_1 у рівняннях регресії (6.10), (6.13), (6.14). b_1 — вибірковий коефіцієнт регресії. Якщо коефіцієнт кореляції значимий, то з огляду на (6.14) коефіцієнт регресії b_1 також значимий. Нехай β_1 — коефіцієнт регресії генеральної сукупності, тоді для нього можна побудувати довірчий інтервал [1].

$$b_1 - t(\gamma, n-2)\frac{s_y\sqrt{1-r_{\text{в}}^2}}{s_x\sqrt{n-2}} < \beta_1 < b_1 + t(\gamma, n-2)\frac{s_y\sqrt{1-r_{\text{в}}^2}}{s_x\sqrt{n-2}}, \tag{6.24}$$

де $t(\gamma, n-2)$ — табличне значення розподілу Стьюдента (γ — надійність, $n-2$ — ступені вільності).

За даними прикладу 1, маємо:

$$-0{,}07982 - 2{,}069 \cdot \frac{1{,}6026 \cdot \sqrt{1-(0{,}8426)^2}}{16{,}9168\sqrt{25-2}} < \beta_1 < -0{,}07982 + 2{,}069\frac{1{,}6026\sqrt{1-(0{,}8426)^2}}{16{,}9168\sqrt{25-2}},$$

або остаточно маємо $-0{,}1015 < \beta_1 < -0{,}0581$ тобто істинне значення β_1 лежить в інтервалі (–0,1015; –0,0581) і це встановлено з надійністю 95 %.

Важливою характеристикою кореляційного аналізу вважається також кореляційне відношення, яке обчислюють за формулою:

$$\eta_{yx} = \sqrt{\frac{D_m}{s_y^2}}, \tag{6.25}$$

де $D_m = \frac{\sum_{j=1}^{l}\left(\overline{y_j} - \overline{y}\right)^2 n_j}{n}$ — міжгрупова дисперсія, $\overline{y_j}$ — середнє j-ої групи, l — кількість груп, і коефіцієнт детермінації:

$$R = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{m}\left(\overline{y_i} - \overline{y}\right)^2}{s_y^2}} = |r|, \tag{6.26}$$

де m — кількість значень y, r — коефіцієнт кореляції, $\overline{y_i}$ — розрахункове значення $y(x_i)$ з рівняння регресії. Емпіричне кореляційне відношення y на x (η_{yx}) характеризує ступінь впливу змінності x на зміну y по відношенню до неврахованих факторів. Чим ближче значення η_{yx} до 1 тим більший вплив змінності x на змінність y. При $\eta_{yx} = 1$, зв'язок між x і y детермінований (функціональний), для $\eta_{yx} = 0$ — кореляційний зв'язок відсутній.

Коефіцієнт детермінації R^2 (для парної лінійної регресії він дорівнює r^2) показує ступінь змінності y зумовлену змінністю x. Чим ближче R^2 до одиниці, тим тісніше спостережувані значення (x_i, y_i) розташовані до лінії регресії, і, відповідно, рівняння регресії краще описує лінійну залежність змінних.

Так, наприклад, якщо рівняння регресії описує зв'язок тижневої кількості виробів з відповідними затратами на їх виробництво і при цьому $R^2 = 0{,}755$, то це свідчить про те, що 75,5 % варіації тижневих затрат пояснюється кількістю виробів, виготовлених за тиждень. Решта 24,5 % варіації залишається не поясненою. Вони, можливо, пояснюються іншими, не врахованими в регресійній моделі факторами, або не достатньою адекватністю моделі.

Для перевірки значущості кореляційного відношення генеральної сукупності використовують F — розподіл Фішера-Снедекера. Для цього обчислюють

$$F = \frac{\eta_{yx}^2(n-m)}{\left(1-\eta_{yx}^2\right)(m-1)}, \tag{6.27}$$

де m — кількість інтервалів (сформованих за груповими ознаками). Кореляційне відношення η_{yx} суттєво відрізняється від 0, якщо $F > F_{\text{кр}}$, де $F = F(\alpha, \text{v}_1, \text{v}_2)$ — табличне значення критерію для рівня значущості α; $\text{v}_1 = m-1)$, $\text{v}_2 = n-m)$ — ступені вільності.

Для визначення значущості коефіцієнта детермінації будуємо випадкову величину:

$$F = \frac{R^2(n-2)}{1-R^2}, \tag{6.28}$$

якщо $F > F_{кр}$, то R — статистично значущий. Тут $F_{кр} = F(\alpha, v_1, v_2)$, ступені вільності $v_1 = 1$, $v_2 = n - 2$.

Приклад 6.4. В таблиці наведено вихідні дані для обчислення η_{yx} і R^2. Перевірити їх значимість.

x_i	n_j	$\overline{y}_j$	$\left(\overline{y}_j - \overline{y}\right)^2 n_j$	$\overline{y}_j$	$\left(y_i - \overline{y}\right)^2 n_i$
22,5	3	10,3	131,5	10,4	127,5
27,5	13	13,3	170,4	13,8	126,5
32,5	21	17,8	16,3	17,2	1,6
37,5	11	20,3	125,7	20,6	149,0
42,5	2	23,0	73,9	23,9	97,4
	$\sum$		517,8	—	502,0

Загальна середня $\overline{y} = 16{,}92$; $s_y^2 = 18{,}23$, (див. пр. 1) групові середні $\overline{y_j}$ і y_i — очікувані розрахункові значення $y_i = y(x_i)$.

За формулою (6.25) маємо:

$$\eta_{yx} = \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{5}\left(y_j - \overline{y}\right)^2 n_j}{s_y^2}} = \sqrt{\frac{10{,}36}{18{,}23}} = 0{,}754.$$

Підставимо одержані з таблиці дані у формулу (6.18), тоді знайдемо $R = \sqrt{\frac{10{,}64}{18{,}23}} \approx 0{,}742 \approx r$; $R^2 = 0{,}551$ — коефіцієнт детермінації. Він показує, що 55,1 % варіації y обумовлено варіацією змінної x, а 44,9 % — іншими факторами.

Нехай $H_0 : \eta_{yx} = 0;\ H_1 : \eta_{yx} \neq 0$. Обчислимо F за (6.27):

$$F = \frac{0{,}754(50-5)}{(1-0{,}754)(5-1)} = 14{,}82.\ F_{кр} = F(0{,}05;4;95) = 2{,}57.\ F > F_{кр},$$

тому гіпотезу H_0 відхиляємо, тобто кореляційне відношення η_{yx} значуще.

За аналогічних гіпотез для коефіцієнта детермінації знаходимо F за формулою (6.28): $F = \frac{0{,}557 \cdot (50-2)}{1-0{,}551} = 5{,}89$; $F_{кр}(0{,}5;1;48) = 4{,}04$, $F > F_{кр}$. Коефіцієнт детермінації також значущий.

6.2. МНОЖИННА ЛІНІЙНА РЕГРЕСІЯ

Зазвичай на будь-який економічний показник впливають не один, а кілька факторів. Тому модель парної лінійної регресії є дещо ідеалізованою і не враховує багато інших факторів впливу. Тому далі розглянемо модель множинної регресії, в якій взаємозв'язок випадкових величин Y і $X_1, X_2, ..., X_m$ генеральної сукупності подається у вигляді теоретичного лінійного рівняння

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon, \tag{6.29}$$

де $\vec{\beta}$ — вектор невідомих параметрів $\vec{\beta} = \left(\beta_0, \beta_1, ..., \beta_m\right)$, а ε — випадкове відхилення, Як і для парної лінійної регресії (6.10), для оцінки вектора $\vec{\beta}$ маємо вибірку з генеральної сукупності у вигляді:

$$\vec{X} = (X_1, X_2, ... X_m) \text{ і } Y = (y_1, y_2, ..., y_n), \tag{6.30}$$

де $X_i = \left(x_{i1}, x_{i2}, ... x_{im}\right);\quad i = 1, 2, ... n.$

Відповідне (6.21) емпіричне рівняння регресії подамо у вигляді:

$$Y = b_0 + b_1 X_1 + b_1 X_2 + ... + b_m X_m + e, \tag{6.31}$$

де вектор $\vec{b} = \left(b_0, b_1, b_2, ..., b_m\right)$ — оцінка теоретичних значень вектора $\overline{\beta}$, e — оцінка відхилень ε.

Значення вектора $\vec{b}$ одержимо за допомогою методу найменших квадратів (МНК), вважаючи, що для компонент вектора ε виконуються умови: $M(\varepsilon_i)=0$; $D(\varepsilon_i)=\sigma^2=const$; $\rho_{\varepsilon_i\varepsilon_j}=\begin{cases}0 & i\neq j\\ \sigma^2 & i=j\end{cases}$; $i=1\div n$, $j=1\div n$; відсутній сильний лінійний зв'язок (мультиколінеарність) між компонентами вектора $\vec{X}$ а компоненти вектора ε розподілені нормально. За виконання цих умов оцінки $b_0, b_1, b_{2,} ..., b_m$ параметрів $\beta_0, \beta_1, ..., \beta_m$ множинної лінійної регресії (6.29) за МНК будуть незміщеними, ефективними.

Подамо вибіркові дані та відповідні коефіцієнти в матричній формі:

$$Y=\begin{pmatrix}y_1\\ y_2\\ .\\ .\\ .\\ y_n\end{pmatrix},\quad X=\begin{pmatrix}1 & x_{11} & x_{12} & ... & x_{1m}\\ 1 & x_{21} & x_{22} & ... & x_{2m}\\ . & . & . & ... & .\\ 1 & x_{n1} & x_{n2} & ... & x_{nm}\end{pmatrix};\quad B=\begin{pmatrix}b_0\\ b_1\\ .\\ .\\ .\\ b_m\end{pmatrix};\quad e=\begin{pmatrix}e_1\\ e_2\\ .\\ .\\ .\\ e_w\end{pmatrix}$$

Y —матриця розмірності $n\times 1$ вибірки значень випадкової величини *Y*;

X — матриця розмірності $n\times(m+1)$, в *і*-ому рядку якої спостережуваені значення вектора X_i; *В* — матриця розмірності $m\times 1$ параметрів регресії (6.31); *е* — матриця розмірності $n\times 1$ відхилень вибіркових значень y_i, від значень $\tilde{y}_i$, одержаних з рівняння регресії:

$$\tilde{y}_i=b_0+b_1x_{i1}+b_2x_{i2}+...+b_mx_{im}\,. \tag{6.32}$$

Тоді система рівнянь для знаходження матриці *В* має матричний вигляді:

$$\left(X^T\cdot X\right)B=X^TY\,, \tag{6.33}$$

де X^T — матриця транспонована по відношенню до матриці X, а її розв'язок:

$$B=\left(X^T\cdot X\right)^{-1}X^TY\,, \tag{6.34}$$

де $\left(X^T\cdot X\right)^{-1}$ — обернена матриця до матриці $\left(X^T\cdot X\right)$.

Так, для $m=2$, рівняння регресії $y=b_0+b_1x_1+b_2x_2$ маємо систему:

$$\begin{cases} n\;\; b_0+b_1\sum\limits_{i=1}^{n}x_{i1}+b_2\sum\limits_{i=1}^{n}x_{i2}=\sum\limits_{i=1}^{n}y_i\\ b_0\;\;\sum\limits_{i=1}^{n}x_{i1}+b_1\sum\limits_{i=1}^{n}x_{i1}^2+b_2\sum\limits_{i=1}^{n}x_{i1}x_{i2}=\sum\limits_{i=1}^{n}y_ix_{i1}\\ b_0\;\;\sum\limits_{i=1}^{n}x_{i2}+b_1\sum\limits_{i=1}^{n}x_{i1}x_{i2}+b_2\sum\limits_{i=1}^{n}x_{i2}^2=\sum\limits_{i=1}^{n}y_ix_{i2}\,.\end{cases} \tag{6.35}$$

Позначимо Z_{jj} j-й діагональний елемент матриці $Z=\left(X^TX\right)^{-1}$. Оскільки значення дисперсії $D(\bar{\varepsilon})=\sigma^2$ за вибіркою з генеральної сукупності визначити неможливо, то замінимо її відповідною незміщеною оцінкою:

$$s^2=\frac{\sum\limits_{i=1}^{n}\left(y_i-b_0-b_1x_{i1}-...-b_mx_{im}\right)^2}{n-m-1}=\frac{\sum\limits_{i=1}^{n}e_i^2}{n-m-1}\,. \tag{6.36}$$

Тоді вибіркові дисперсії емпіричних коефіцієнтів регресії можна визначити за формулами:

$$s_{bj}^2=s^2Z_{jj}=\frac{Z_{jj}\sum e_i^2}{n-m-1},\quad j=1\div m\,, \tag{6.37}$$

а величина $s_{bj}=\sqrt{s_{bj}^2}$ — стандартна похибка відповідного коефіцієнта регресії. Для рівняння регресії $Y=b_0+b_1X_1+b_2X_2$ дисперсії і стандартні похибки обчислюють за формулами:

$$s_{b_0}^2=\left(\frac{1}{n}+\frac{\overline{x}_1^2 s_{x_2}^2+\overline{x}_2^2 s_{x_1}^2-2\overline{x}_1\overline{x}_2\rho_{x_1x_2}}{s_{x_1}^2\cdot s_{x_2}^2-\rho_{x_1x_2}^2}\right)s^2; \qquad (6.38)$$

$$s_{b_1}^2=\frac{s_{x_2}^2\cdot s^2}{s_{x_1}^2 s_{x_2}^2-\rho_{x_1x_2}^2};\ s_{b_2}^2=\frac{s_{x_1}^2 s^2}{s_{x_1}^2 s_{x_2}^2-\rho_{x_1x_2}^2},\text{ де } s_{x_1}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i_1}-\overline{x}_1\right)^2,$$

$$s_{x_2}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i_2}-\overline{x}_2\right)^2,\ \rho_{x_1x_2}=\frac{1}{n}\sum_{i=1}^{n}\left(x_{i_1}-\overline{x}_1\right)\left(x_{i_2}-\overline{x}_2\right),$$

відповідно вибіркові дисперсії і кореляційний момент (коваріація).

Кореляційний момент між b_1 і b_2 можна обчислити за формулою:

$$s_{b_1b_2}=\frac{-r_{x_1x_2}\cdot s^2}{\left(1-r_{x_1x_2}^2\right)s_{x_1}s_{x_2}}, \qquad (6.39)$$

де $r_{x_1x_2}=\dfrac{\rho_{x_1x_2}}{s_{x_1}s_{x_2}}$ — вибірковий коефіцієнт кореляції між змінними X_1 і X_2.

Для одержання інтервальних оцінок коефіцієнтів β_j теоретичного рівняння регресії (6.29) побудуємо випадкову величину:

$$t=\frac{b_j-\beta_j}{s_{b_j}}, \qquad (6.40)$$

яка має розподіл Стьюдента з $n-m-1$ ступенями вільності. За рівняя значущості α і за $n-m-1$ ступенями вільності знайдемо критичну точку розподілу Стьюдента $t_{kp}\left(\dfrac{\alpha}{2};n-m-1\right)$ і з умови $|t|<t_{kp}\left(\dfrac{\alpha}{2};n-m-1\right)$ дістанемо довірчий інтервал для β_j:

$$b_j-s_{b_j}t_{kp}\left(\frac{\alpha}{2};n-m-1\right)<\beta_j<b_j+s_{b_j}t_{kp}\left(\frac{\alpha}{2};n-m-1\right). \qquad (6.41)$$

Перевірити гіпотези $H_0:b_j=0$; $H_1:b_j>0$ можна побудувавши випадкову величину $t=\dfrac{b_j}{s_{b_j}}$, яка розподілена за Стьюдентом з $n-m-1$ ступенями вільності. Якщо $|t|>t_{kp}\left(\dfrac{\alpha}{2};n-m-1\right)$, то H_0 відхиляється і коефіцієнт b_j вважається статистично значущим. За $|t|\le t_{kp}\left(\dfrac{\alpha}{2};n-m-1\right)$ коефіцієнт b_j статистично не значущий, а наявність відповідної змінної X_j у рівнянні регресії з погляду статистики не виправдане. Його можна позбутися в рівнянні регресії. Для уточнення вигляду рівняння множинної регресії використовують коефіцієнт детермінації:

$$\overline{R}^2=1-\frac{\frac{1}{n-m-1}\sum_{i=1}^{n}e_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}. \qquad (6.42)$$

Чим ближчий цей коефіцієнт до одиниці, тим краще рівняння регресії пояснює поведінку Y, у загальному випадку $0\le\overline{R}^2\le 1$. Доведено, що додавати в модель множинної регресії додаткові змінні доцільно лише в тому випадку, поки зростає $\overline{R}^2$.

Для перевірки значущості коефіцієнта детермінації побудуємо випадкову величину:

$$F = \frac{\frac{n-m-1}{m}\overline{R}^2 + 1}{1-\overline{R}^2}, \qquad (6.43)$$

яка має розподіл Фішера з m і $n-m-1$ ступенями вільності. Для рівня значущості α і ступенів вільності m і $n-m-1$ за таблицями розподілу Фішера знаходимо $F_{kp}(\alpha; m; n-m-1)$, якщо $F > F_{kp}(\alpha; m; n-m-1)$, то $\overline{R}^2 > 0$, тобто статистично значущий.

За наявності у вибірках трендів (особливо це можливо в аналізі числових рядів), або автокореляції, порушення умов незалежності між собою відхилень e_i, навіть за гарних показників значущості b_j і $\overline{R}^2$, рівняння регресії може не відображати реального взаємозв'язку між змінними.

На практиці для аналізу наявності або відсутності автокореляції використовують статистику Дарбіна—Уотсона DW, яка розраховується за формулою:

$$DW = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \qquad (6.44)$$

За відсутності автокореляції поведінка відхилень e_i випадкова, тому можна вважати, що в половині випадків $e_i \approx e_{i-1}$, а в другій половині $e_i \approx -e_{i-1}$. Тому з урахуванням (6.44) маємо:

$$DW = \frac{\sum_{i=1}^{n}\frac{1}{2}(2e_i)^2}{\sum_{i=1}^{n} e_i^2} = 2\frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} e_i^2} = 2.$$

Отже, необхідною умовою незалежності випадкових відхилень є близькість до двійки статистики DW. У таблицях критичних точок статистики DW для рівня значущості α обсяг вибірки n і кількості змінних моделі регресії m указується два числа d_1 і d_2 — відповідно верхня і нижня межі.

Якщо : $DW < d_1$ — є додатня автокореляція залишків;

$DW > 4 - d_1$ — є від'ємна автокореляція залишків;

$d_2 < DW < 4 - d_2$ — автокореляції немає;

$d_1 < DW < d_2$, або $4 - d_2 < DW < 4 - d_1$, то автокореляція може бути, а може й не бути.

При наявності автокореляції залишків запропонована модель регресії вважається незадовільною.

Приклад 6.5. У [22] аналізується обсяги Y заощаджень домогосподарством за 10 років. Уважається його розмір y_i в певному році залежить від величини x_{i-1} доходу X_1 у попередньому році та відсоткової справки X_2 в поточному році. Дані наведено в таблиці.

Рік	80	81	82	83	84	85	86	87	88	89	90
X_1, тис.	100	110	140	150	160	160	180	200	230	250	260
X_2, %	2	2	3	2	3	4	4	3	4	5	5
Y, тис.	20	25	30	30	35	38	40	38	44	50	55

1) За даними МНК оцінити коефіцієнти лінійної регресії $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.
2) Оцінити статистичну значущість коефіцієнтів регресії b_0, b_1, b_2.
3) Для $\alpha = 0{,}05$ побудувати довірчі інтервали для знайдених коефіцієнтів.
4) Обчислити коефіцієнт детермінації $\overline{R}^2$ і оцінити його статистичну значущість $\alpha = 0{,}05$.
5) Обчислити статистику Дарбіна—Уотсона DW і оцінити наявність автокореляції.
6) Зробити висновки про якість моделі.
7) Визначити, як зміниться обсяг збережень зі ростанням відсоткової ставки, чи буде відповідь статистично обґрунтована;
8) Зробити прогноз на наступний рік обсягу збережень, якщо дохід становить 270 тис. грн, а відсоткова ставка буде 5,5 %.

Проміжні результати запишемо в таблицю:

Рік	x_1	x_2	y	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
80	100	2	20	10 000	4	200	2000	40
81	110	2	25	12 100	4	220	2750	50
82	140	3	30	19 600	9	420	4200	90
83	150	2	30	22 500	4	300	4500	60
84	160	3	35	25 600	9	480	5600	105
85	160	4	38	25 600	16	640	6080	152
86	180	4	40	32 400	16	720	7200	160
87	200	3	38	40 000	9	600	7600	114
88	230	4	44	52 900	16	920	10 120	176
89	250	5	50	62 500	25	1250	12 500	250
90	260	5	55	67 600	25	1300	14 300	275
Сума	1940	37	405	370 800	137	7050	76 850	1472
Середнє	176,3636	3,3636	36,8182	33 709,0909	12,4546	640,9091	6986,3636	133,8182
$\sum\left(x_{i1}-x\right)^2$ 28 654,55	$\sum\left(x_{i2}-\overline{x}_2\right)^2$ $\sum\left(y_i-\overline{y}\right)^2$ $\sum\left(x_{i1}-\overline{x}_1\right)\left(x_{i2}-\overline{x}_2\right)^2$ $\sum\left(x_{i1}-\overline{x}_1\right)\left(y_i-\overline{y}\right)$ $\sum\left(x_{i2}-\overline{x}_2\right)\left(y_i-\overline{y}\right)$ 12, 5455 1087,636 524,5451 5422,727 109,7272							

Коефіцієнт знайдемо з розв'язку системи рівнянь (6.35):

$$b_0 = 2{,}962\ ;\ b_1 = 0{,}1242\,;\ b_2 = 3{,}554\,.$$

Отже, емпіричне рівняння регресії має вигляд

$$y = 2{,}962 + 0{,}1242x_1 + 3{,}554x_2\,.$$

З рівняння знайдемо модельне значення $\widetilde{y}_i\left(x_{i1}, x_{i2}\right)$ залежної величини Y і обчислимо відхилення e_i реальних значень від модельних. Дані наведено в таблиці:

Рік	Y	$\widetilde{y}$	e_i	e_i^2	$e_i - e_{i-1}$	$\left(e_i - e_{i-1}\right)^2$
80	20	22,48852	–2,48852	6,19273	—	—
81	25	23,73041	1,269594	1,61187	3,75811	14,12339
82	30	31,00991	–1,00991	1,01992	–2,27950	5,19612
83	30	28,69796	1,30204	1,69523	2,31194	5,34507
84	35	33,49369	1,50631	2,26896	0,20427	0,04173
85	38	37,04753	0,95247	0,90719	–0,55384	0,30674
86	40	39,53131	0,46869	0,21967	–0,48378	0,23404
87	38	38,46125	–0,46125	0,21275	–0,92994	0,86479
88	44	45,74076	–1,74076	3,03024	–1,27951	1,63714
89	50	51,77838	–1,77838	3,16263	–0,03762	0,00141
90	55	53,02027	1,97973	3,91933	3,75811	14,12332
Сума	405	405	≈ 0	24,24058	—	41,87375
Середнє	36,81818	36,81818	—	—	—	—

Для аналізу значущості коефіцієнтів регресії знайдемо дисперсію регресії за формулою (6.28):

$$s^2 = \frac{24{,}2406}{8} \approx 3{,}03$$

і стандартні відхилення коефіцієнтів за формулами (6.30):

$$s_{b_0}^2=\left(\frac{1}{11}+\frac{31104{,}12\cdot 12{,}55+11{,}31\cdot 28654{,}5-2\cdot 176{,}36\cdot 3{,}36\cdot 524{,}54}{28654{,}5\cdot 12{,}55-275147{,}56}\right)3{,}03=3{,}583;$$

$$s_{b_1}^2=\frac{12{,}55\cdot 3{,}03}{84466{,}415}=0{,}00054;\ s_{b_2}^2=\frac{28654{,}5\cdot 303}{84466{,}415}=1{,}03;$$

$$s_{b_0}=1{,}893;\ S_{b_1}=0{,}0212;\ s_{b_2}=1{,}0146.$$

Для перевірки гіпотез $\mathrm{H}_0: b_j=0$; $\mathrm{H}_1: b_j>0$ обчислимо статистики $t=\dfrac{b_j}{S_{b_j}}$, маємо $t_{b_0}=1{,}565$; $t_{b_1}=5{,}858$; $t_{b_2}=3{,}503$.

Для рівня значущості $\alpha=0{,}05$ маємо $t_{\text{кр}}(0{,}025\ ,8\ )=2{,}306$, отже, $|t_{b_1}|>t_{\text{кр}}$; $|t_{b_2}|>t_{\text{кр}}$, а $|t_{b_0}|<t_{\text{кр}}$ і коефіцієнти b_1,b_2 — значущі, а b_0 — ні. З економічного погляду b_0 відображає екзогенне середовище, тому його краще використати в моделі, а не відкинути.

За формулами (6.33) знаходимо довірчі інтервали для коефіцієнтів:

$$-1{,}403<\beta_0<7{,}327;\ 0{,}0753<\beta_1<0{,}1731;\ 1{,}2142<\beta_2<5{,}8935.$$

Коефіцієнт детермінації (скорегований) з (6.34) $\overline{R}^2=0{,}9721$. Для F статистики, формула (6.35), маємо $F=175{,}21$. Критичне значення $F_{\text{кр}}(0{,}05;2;8)=4{,}46$ $F>F_{\text{кр}}$, тому $\overline{R}^2$ — значущий.

Статистика DW за формулою (6.36) дорівнює $DW=\dfrac{41{,}874}{24{,}2406}=1{,}727$.

За таблицями для $\alpha=0{,}05$, $n=11$ знаходимо $d_1=0{,}658$; $d_2=1{,}604$. Отже, маємо $1{,}604<DW<4-1{,}504=2{,}396$, тобто маємо підстави вважати, що автокореляція відсутня.

За всіма статистичними показниками можна зробити висновок, що запропонована модель доволі адекватна статистичним даним і може бути використана для аналізу та прогнозу. У зв'язку з цим можна передбачити, що зі зростанням відсоткової ставки буде збільшуватися обсяг заощаджень (коефіцієнт b_2 додатній і значущий, а через обсяг заощаджень у наступному році за відсоткової ставки 5,5 % і доходу 270 тис. буде становити близько 56,04 тис.

Приклади для самостійного розв'зування

1. За даними таблиці побудувати три регресійні моделі

Y	X_1	X_2
1	0	2
3	1	1
5	3	0
11	4	—2

1) $Y=\alpha_0+\alpha_1X_1+\varepsilon$,

2) $Y=\gamma_0+\gamma_2X_2+\varepsilon$,

3) $Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$.

1) Чи будуть справедливими гіпотези $H_0:\alpha_1=\beta_1$; $H_0:\gamma_2=\beta_2$?

2) Зробіть висновки щодо побудованих моделей.

2. Обсяг Q пропозиції деякого товару на конкретному ринку залежить від ціни P товару і заробітної плати співробітників W. За даними таблиці спостережень за 16 місяців

Q	20	35	30	45	60	69	75	90	105	110	120	130	130	130	135	140
P	10	15	20	25	40	37	43	35	38	55	50	35	40	55	45	65
W	12	10	9	9	8	8	6	4	4	5	3	1	2	3	1	2

1) Оцінити за МНК коефіцієнти рівняння регресії $Q=\beta_0+\beta_1P+\beta_2W+\varepsilon$.
2) Визначити інтервальні оцінки коефіцієнтів для рівня значущості $\alpha=0{,}1$.
3) Перевірити гіпотези про статистичну значущість коефіцієнтів рівняння регресії.
4) Перевірити статистичну значущість коефіцієнта детермінації $\overline{R}^2$.

5) Перевірити гіпотезу про відсутність автокореляції залишків.

6) Зробіть висновки з приводу побудованої моделі.

3. Для пояснення змін ВВП за 10 років (Y) будується регресійна модель з змінними X_1 — споживанням і X_2 — інвестиціями за статистичними даними:

X_1 млрд. $	8	9,5	11	12	13	14	15	16,5	17	18
X_2 млрд. $	1,65	1,8	2,0	2,1	2,2	2,4	2,65	2,85	3,2	3,55
ВВП, млрд. $	14	16	18	20	23	23,5	25	26,5	28,5	30,5

1) Оцінити коефіцієнти лінійної регресії $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

2) Оцініть стандартну помилку регресії і стандартні помилки коефіцієнтів.

3) Перевірте гіпотези значущості коефіцієнтів для $\alpha = 0{,}05$.

4) Перевірте статистичну значущість коефіцієнта детермінації $\overline{R}^2$; $\alpha = 0{,}05$.

5) Знайдіть значення статистики DW. Чи має місце автокореляція?

6) Зробіть висновки щодо якості моделі.

7) Через три роки очікуються рівні споживання $X_1 = 22$, інвестицій $X_2 = 3{,}8$. Яким буде очікуваний рівень ВВП (Y)?

4. За 20 спостережень одержано такі результати:

$\sum x_{i1} = 4{,}88; \sum x_{i1}^2 = 2{,}518; \sum x_{i2} = 26{,}7;$

$\sum x_{i2}^2 = 75{,}15; \sum y_i = 44{,}7; \sum x_{i1} x_{i2} = 13{,}75;$

$\sum x_{i1} y_i = 22{,}1; \sum x_{i2} y_i = 125{,}75; \sum y_i^2 = 210{,}4; \sum e_i^2 = 0{,}015.$

1) Оцінити коефіцієнт лінійної регресії $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

2) Визначити стандартні похибки коефіцієнтів.

3) Для $\alpha = 0{,}05$ знайдіть довірчі інтервали для β_1, β_2.

4) Оцінити статистичну значущість для коефіцієнтів регресії і детермінації для $\alpha = 0{,}05$.

6.3. НЕЛІНІЙНА РЕГРЕСІЯ

Розглянуті в попередніх параграфах моделі парної і множинної лінійних регресій у багатьох практичних випадках дають задовільний результат для встановлення залежностей між економічними показниками і можуть бути використані для аналізу та прогнозування. Однак існує багато випадків залежностей між економічними параметрами, в яких використання лінійних моделей неможливе. Тому зупинимось тут на розгляді деяких нелінійних моделей, зокрема, логарифмічних, поліноміальних, експоненціальних, обернених, які використовуються в багатьох галузях економіки для більш реального відображення залежностей.

У більшості випадків розглядувані нижче нелінійні моделі шляхом відповідного перетворення змінних зводяться до лінійних стосовно до інших проміжних змінних. Розглянемо деякі випадки таких перетворень. Нехай деяка економічна залежність моделюється співвідношенням:

$$Y = AX^{\beta}, \tag{6.45}$$

де A, β — параметри. Залежно від параметрів ця модель відображає залежність попиту від ціни $(\beta < 0)$, або від доходу $(\beta > 0)$. Загалом (6.45) називають функцією Енгеля. Якщо прологарифмувати вираз (6.45), дістанемо:

$$\ln Y = \ln A + \beta \ln X, \tag{6.46}$$

в якому, зробивши заміни $\ln Y = Y^*; \ln A = \beta_0$ і $\ln X = X^*$, перейдемо для нових змінних $(X^*; Y^*)$ до звичайної лінійної моделі:

$$Y^* = \beta_0 + \beta X^* + \varepsilon. \tag{6.47}$$

Якщо виконуються умови застосування МНК, з неї можна одержати незміщені оцінки для β_0 і β. Слід також зазначити, що в моделі (6.45) параметр β відіграє роль еластичності змінної Y по змінній X. Із (6.46) маємо:

$$\frac{y'}{y} = \frac{\beta}{x} \Rightarrow \beta = y' \frac{x}{y} = E_x(y).$$

Виконавши логарифмування, наприклад для виробничої функції Коба-Дугласа $Y = AK^{\alpha}L^{\beta}$, дістанемо:

$$\ln Y = \ln A + \alpha \ln K + \beta \ln L \tag{6.48}$$

після замін $\ln Y = Z; \ln A = \beta_0; \ln K = X_1; \ln L = X_2$ одержимо лінійну регресії модель:

$$Z = \beta_0 + \alpha X_2 + \beta X_2 .$$

У банківському і фінансовому аналізі часто використовується модель вигляду:

$$Y_t = Y_0 (1 + r)^t , \tag{6.49}$$

де Y_0 — початкове значення змінної (наприклад початковий внесок у банк); Y_t — значення змінної в момент часу t, r — складний темп приросту Y (відсоткова ставка). Після логарифмування маємо:

$$\ln Y_t = \ln Y_0 + t \ln(1 + r). \tag{6.50}$$

Позначивши $\ln Y_t = Y ; \ln Y_0 = \beta_0 ; \ln(1 + r) = \beta$, будемо мати лінійну модель:

$$Y = \beta_0 + \beta t + \varepsilon_t . \tag{6.51}$$

У моделях, які вивчають залежності між обсягами випуску продукції та середніми витратами виробництва, між доходами і попитом, рівнем безробіття і процентними змінами заробітної плати, часто застосовується обернена модель:

$$Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon , \tag{6.52}$$

яка заміною $\frac{1}{X} = X_1$ зводиться до лінійної.

Аналогічно можна звести до лінійної поліноміальну модель:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + ... + \beta_m X^m + \varepsilon , \tag{6.53}$$

показникову модель $Y = b_0 e^{BX}$, або більш складну модель Коба-Дугласа з урахуванням технічного прогресу:

$$Y = AK^{\alpha}L^{\beta}e^{\gamma t} , \tag{6.54}$$

Приклади для самостійного розв'язування:

1. Аналізується прибуток Y (млн \$), залежно від витрат на рекламу X (млн \$) за 9 років одержано дані:

Y	5	7	13	15	20	25	22	20	17
X	0,8	1,0	1,8	2,5	4,0	5,7	7,5	8,3	8,8

1) Побудувати кореляційне поле і надайте пропозиції щодо формули залежності між показниками.
2) Оцінити за МНК коефіцієнт лінійної регресії $Y = \beta_0 + \beta_1 X + \varepsilon$.
3) Оцінити якість моделі.
4) Оцінити за МНК коефіцієнти квадратичної регресії $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.
5) Оцінити якість побудованої регресії. Яка з моделей має переваги, в чому?

2. За статистичними даними:

t	1	4	16
Y	0,85	0,45	0,4

оцінити параметри β, γ моделі $Y = \beta e^{\gamma t} + \varepsilon$.

3. Аналізується індекс споживчих цін Y за обсягом грошової маси X (млрд \$) на основі даних з 1981 по 1997 роки.

Рік	81	82	83	84	85	86	87	88	89	90
Y	65	68	72,5	77,5	82	85,5	88,5	91	95	100
X	110	125	132	137	160	177	192	215	235	240

Рік	91	92	93	94	95	96	97
Y	106,5	112	115,5	118,5	120	120,5	121
X	245	250	275	285	295	320	344

1) Побудувати кореляційне поле.
2) Побудувати регресії Y на X; $\ln Y$ на $\ln X, \ln Y$ на $\ln X$.
3) Інтерпретувати коефіцієнти регресії для кожної з моделей.
4) Для кожної з моделей знайти еластичність Y по X.
5) Яка з моделей найкраща?
6) За статистичними даними досліджується залежність між темпом приросту заробітної плати $\dot{w}_t = \frac{w_t - w_{t-1}}{w_t} / 100\,\%$ і рівнем безробіття u_t

Рік	70	71	72	73	74	75	76	77	78	79
w_t	1,61	1,66	1,80	1,95	2,05	2,12	2,25	2,45	2,55	2,67
u_t	1,0	1,38	1,15	1,50	1,55	1,20	1,1	1,0	1,35	1,8

Рік	80	81	82	83	84	85	86	87	88
w_t	2,73	2,80	2,93	3,02	3,15	3,27	3,45	3,60	3,80
u_t	1,9	1,45	1,85	1,2	1,5	1,25	1,4	1,3	1,6

Для одержання модельної залежності використовується крива Філіппа:

$$\dot{w}_t = \beta_0 + \beta_1 \frac{1}{u_t} + \varepsilon_t .$$

1) Зведіть модель до лінійної і оцініть за МНК β_0, β_1.
2) Побудуйте кореляційне поле.
3) Визначте довірчі інтервали для коефіцієнтів β_0, β_1 і $\alpha = 0{,}05$.
4) За розташуванням точок на кореляційному полі спробуйте підібрати іншу модель залежності $\dot{w}_t$ від u_t. Знайдіть оцінки параметрів запропонованих моделей.

ДОДАТКИ

Таблиця 1

ТАБЛИЦЯ ЗНАЧЕНЬ ФУНКЦІЇ ГАУСА—ЛАПЛАСА $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3918
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3825
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3725	0,3712	0,3697
0,4	0,3683	0,3668	0,3653	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3538
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3352
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3144
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2920
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2685
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1965
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1736
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1518
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1315
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0957
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0804
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0669
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0396	0,0387	0,0379	0,0371	0,0363
2,2	0,0355	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0290
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0229
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0180
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0139
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0107
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0081
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0061
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0025
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0018
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0013
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006
3,6	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0004
3,7	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
3,8	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002
3,9	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001

Таблиця 2

ТАБЛИЦЯ ЗНАЧЕНЬ ФУНКЦІЇ ЛАПЛАСА $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Таблиця 3

ІМОВІРНОСТІ $P_a(X=m)=\frac{a^m}{m!}e^{-a}$, $m=0,1,2,\ldots$ **РОЗПОДІЛУ ПУАССОНА**

m \ a	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494
4		0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111
5				0,0001	0,0002	0,0004	0,0007	0,0012	0,0020
6							0,0001	0,0002	0,0003

m \ a	1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
0	0,3679	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496
1	0,3679	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842
2	0,1839	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700
3	0,0613	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710
4	0,0153	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812
5	0,0031	0,0045	0,0062	0,0084	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309
6	0,0005	0,0008	0,0012	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098
7	0,0001	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027
8				0,0001	0,0001	0,0001	0,0002	0,0003	0,0005	0,0006
9								0,0001	0,0001	0,0001

m \ a	2	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9
0	0,1353	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550
1	0,2707	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596
2	0,2707	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314
3	0,1804	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237
4	0,0902	0,0992	0,1082	0,1169	0,1254	0,1336	0,1414	0,1488	0,1557	0,1622
5	0,0361	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940
6	0,0120	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455
7	0,0034	0,0044	0,0055	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188
8	0,0009	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068
9	0,0002	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022
10		0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006
11							0,0001	0,0001	0,0001	0,0002

Продовження табл. 3

a / m	3	3,5	4	4,5	5	6	7	8	9
0	0,0498	0,0302	0,0183	0,0111	0,0067	0,0025	0,0009	0,0003	0,0003
1	0,1494	0,1057	0,0733	0,0500	0,0337	0,0149	0,0064	0,0027	0,0027
2	0,2240	0,1850	0,1465	0,1125	0,0842	0,0446	0,0223	0,0107	0,0107
3	0,2240	0,2158	0,1954	0,1687	0,1404	0,0892	0,0521	0,0286	0,0286
4	0,1680	0,1888	0,1954	0,1898	0,1755	0,1339	0,0912	0,0573	0,0573
5	0,1008	0,1322	0,1563	0,1708	0,1755	0,1606	0,1277	0,0916	0,0916
6	0,0504	0,0771	0,1042	0,1281	0,1462	0,1606	0,1490	0,1221	0,1221
7	0,0216	0,0385	0,0595	0,0824	0,1044	0,1377	0,1490	0,1396	0,1396
8	0,0081	0,0169	0,0298	0,0463	0,0653	0,1033	0,1304	0,1396	0,1396
9	0,0027	0,0066	0,0132	0,0232	0,0363	0,0688	0,1014	0,1241	0,1241
10	0,0008	0,0023	0,0053	0,0104	0,0181	0,0413	0,0710	0,0993	0,0993
11	0,0002	0,0007	0,0019	0,0043	0,0082	0,0225	0,0452	0,0722	0,0722
12	0,0001	0,0002	0,0006	0,0016	0,0034	0,0113	0,0263	0,0481	0,0481
13		0,0001	0,0002	0,0006	0,0013	0,0052	0,0142	0,0296	0,0296
14			0,0001	0,0002	0,0005	0,0022	0,0071	0,0169	0,0169
15				0,0001	0,0002	0,0009	0,0033	0,0090	0,0090
16						0,0003	0,0014	0,0045	0,0045
17						0,0001	0,0006	0,0021	0,0021
18							0,0002	0,0009	0,0009
19							0,0001	0,0004	0,0004
20								0,0002	0,0002
21								0,0001	0,0001

Закінчення табл. 3

m \ a	10	11	12	13	14	15	16	17	18	19	20
1	0,0005	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0023	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0076	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0189	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0000	0,0000
5	0,0378	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0631	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0901	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0019	0,0010	0,0005
8	0,1126	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1251	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0083	0,0050	0,0029
10	0,1251	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1137	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,0948	0,1094	0,1144	0,1099	0,0984	0,0829	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0729	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0521	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0387
15	0,0347	0,0534	0,0724	0,0885	0,0989	0,1024	0,0992	0,0906	0,0786	0,0650	0,0516
16	0,0217	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0128	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0936	0,0863	0,0760
18	0,0071	0,0145	0,0255	0,0397	0,0554	0,0706	0,0830	0,0909	0,0936	0,0911	0,0844
19	0,0037	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0019	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0866	0,0888
21	0,0009	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0004	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0002	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669
24	0,0001	0,0003	0,0008	0,0020	0,0043	0,0083	0,0144	0,0226	0,0328	0,0442	0,0557
25		0,0001	0,0004	0,0010	0,0024	0,0050	0,0092	0,0154	0,0237	0,0336	0,0446
26			0,0002	0,0005	0,0013	0,0029	0,0057	0,0101	0,0164	0,0246	0,0343
27			0,0001	0,0002	0,0007	0,0016	0,0034	0,0063	0,0109	0,0173	0,0254
28				0,0001	0,0003	0,0009	0,0019	0,0038	0,0070	0,0117	0,0181
29				0,0001	0,0002	0,0004	0,0011	0,0023	0,0044	0,0077	0,0125
30					0,0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31						0,0001	0,0003	0,0007	0,0015	0,0030	0,0054
32						0,0001	0,0001	0,0004	0,0009	0,0018	0,0034
33							0,0001	0,0002	0,0005	0,0010	0,0020
34								0,0001	0,0002	0,0006	0,0012
35									0,0001	0,0003	0,0007
36									0,0001	0,0002	0,0004
37										0,0001	0,0002
38											0,0001
39											0,0001

Таблиця 4

ТАБЛИЦЯ ЗНАЧЕНЬ ФУНКЦІЇ e^{-x}

x	0	1	2	3	4	5	6	7	8	9
0,0	1,0000	0,9900	0,9802	0,9704	0,9608	0,9512	0,9418	0,9324	0,9231	0,9139
0,1	0,9048	0,8958	0,8869	0,8781	0,8694	0,8607	0,8521	0,8437	0,8353	0,8270
0,2	0,8187	0,8106	0,8025	0,7945	0,7866	0,7788	0,7711	0,7634	0,7558	0,7483
0,3	0,7408	0,7334	0,7261	0,7189	0,7118	0,7047	0,6977	0,6907	0,6839	0,6771
0,4	0,6703	0,6637	0,6570	0,6505	0,6440	0,6376	0,6313	0,6250	0,6188	0,6126
0,5	0,6065	0,6005	0,5945	0,5886	0,5827	0,5769	0,5712	0,5655	0,5599	0,5543
0,6	0,5488	0,5434	0,5379	0,5326	0,5273	0,5220	0,5169	0,5117	0,5066	0,5016
0,7	0,4966	0,4916	0,4868	0,4819	0,4771	0,4724	0,4677	0,4630	0,4584	0,4538
0,8	0,4493	0,4449	0,4404	0,4360	0,4317	0,4274	0,4232	0,4190	0,4148	0,4107
0,9	0,4066	0,4025	0,3985	0,3946	0,3906	0,3867	0,3829	0,3791	0,3753	0,3716
1,0	0,3679	0,3642	0,3606	0,3570	0,3535	0,3499	0,3465	0,3430	0,3396	0,3362
1,1	0,3329	0,3296	0,3263	0,3230	0,3198	0,3166	0,3135	0,3104	0,3073	0,3042
1,2	0,3012	0,2982	0,2952	0,2923	0,2894	0,2865	0,2837	0,2808	0,2780	0,2753
1,3	0,2725	0,2698	0,2671	0,2645	0,2618	0,2592	0,2567	0,2541	0,2516	0,2491
1,4	0,2466	0,2441	0,2417	0,2393	0,2369	0,2346	0,2322	0,2299	0,2276	0,2254
1,5	0,2231	0,2209	0,2187	0,2165	0,2144	0,2122	0,2101	0,2080	0,2060	0,2039
1,6	0,2019	0,1999	0,1979	0,1959	0,1940	0,1920	0,1901	0,1882	0,1864	0,1845
1,7	0,1827	0,1809	0,1791	0,1773	0,1755	0,1738	0,1720	0,1703	0,1686	0,1670
1,8	0,1653	0,1637	0,1620	0,1604	0,1588	0,1572	0,1557	0,1541	0,1526	0,1511
1,9	0,1496	0,1481	0,1466	0,1451	0,1437	0,1423	0,1409	0,1395	0,1381	0,1367
2,0	0,1353	0,1340	0,1327	0,1313	0,1300	0,1287	0,1275	0,1262	0,1249	0,1237
2,1	0,1225	0,1212	0,1200	0,1188	0,1177	0,1165	0,1153	0,1142	0,1130	0,1119
2,2	0,1108	0,1097	0,1086	0,1075	0,1065	0,1054	0,1044	0,1033	0,1023	0,1013
2,3	0,1003	0,0993	0,0983	0,0973	0,0963	0,0954	0,0944	0,0935	0,0926	0,0916
2,4	0,0907	0,0898	0,0889	0,0880	0,0872	0,0863	0,0854	0,0846	0,0837	0,0829
2,5	0,0821	0,0813	0,0805	0,0797	0,0789	0,0781	0,0773	0,0765	0,0758	0,0750
2,6	0,0743	0,0735	0,0728	0,0721	0,0714	0,0707	0,0699	0,0693	0,0686	0,0679
2,7	0,0672	0,0665	0,0659	0,0652	0,0646	0,0639	0,0633	0,0627	0,0620	0,0614
2,8	0,0608	0,0602	0,0596	0,0590	0,0584	0,0578	0,0573	0,0567	0,0561	0,0556
2,9	0,0550	0,0545	0,0539	0,0534	0,0529	0,0523	0,0518	0,0513	0,0508	0,0503
3,0	0,0498	0,0493	0,0488	0,0483	0,0478	0,0474	0,0469	0,0464	0,0460	0,0455
3,1	0,0450	0,0446	0,0442	0,0437	0,0433	0,0429	0,0424	0,0420	0,0416	0,0412
3,2	0,0408	0,0404	0,0400	0,0396	0,0392	0,0388	0,0384	0,0380	0,0376	0,0373
3,3	0,0369	0,0365	0,0362	0,0358	0,0354	0,0351	0,0347	0,0344	0,0340	0,0337
3,4	0,0334	0,0330	0,0327	0,0324	0,0321	0,0317	0,0314	0,0311	0,0308	0,0305
3,5	0,0302	0,0299	0,0296	0,0293	0,0290	0,0287	0,0284	0,0282	0,0279	0,0276
3,6	0,0273	0,0271	0,0268	0,0265	0,0263	0,0260	0,0257	0,0255	0,0252	0,0250
3,7	0,0247	0,0245	0,0242	0,0240	0,0238	0,0235	0,0233	0,0231	0,0228	0,0226
3,8	0,0224	0,0221	0,0219	0,0217	0,0215	0,0213	0,0211	0,0209	0,0207	0,0204
3,9	0,0202	0,0200	0,0198	0,0196	0,0194	0,0193	0,0191	0,0189	0,0187	0,0185
4,0	0,0183	0,0181	0,0180	0,0178	0,0176	0,0174	0,0172	0,0171	0,0169	0,0167
4,1	0,0166	0,0164	0,0162	0,0161	0,0159	0,0158	0,0156	0,0155	0,0153	0,0151
4,2	0,0150	0,0148	0,0147	0,0146	0,0144	0,0143	0,0141	0,0140	0,0138	0,0137
4,3	0,0136	0,0134	0,0133	0,0132	0,0130	0,0129	0,0128	0,0127	0,0125	0,0124
4,4	0,0123	0,0122	0,0120	0,0119	0,0118	0,0117	0,0116	0,0114	0,0113	0,0112
4,5	0,0111	0,0110	0,0109	0,0108	0,0107	0,0106	0,0105	0,0104	0,0103	0,0102
4,6	0,0101	0,0100	0,0099	0,0098	0,0097	0,0096	0,0095	0,0094	0,0093	0,0092
4,7	0,0091	0,0090	0,0089	0,0088	0,0087	0,0087	0,0086	0,0085	0,0084	0,0083
4,8	0,0082	0,0081	0,0081	0,0080	0,0079	0,0078	0,0078	0,0077	0,0076	0,0075
4,9	0,0074	0,0074	0,0073	0,0072	0,0072	0,0071	0,0070	0,0069	0,0069	0,0068

Таблиця 5

ІМОВІРНОСТІ БІНОМІАЛЬНОГО РОЗПОДІЛУ $P(x=m)=C_n^m p^m q^{n-m}$

n	*m* \ *p*	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95
2	**0**	0,903	0,810	0,640	0,490	0,360	0,250	0,160	0,090	0,040	0,010	0,003
	1	0,095	0,180	0,320	0,420	0,480	0,500	0,480	0,420	0,320	0,180	0,095
	2	0,003	0,010	0,040	0,090	0,160	0,250	0,360	0,490	0,640	0,810	0,903
3	**0**	0,857	0,729	0,512	0,343	0,216	0,125	0,064	0,027	0,008	0,001	0+
	1	0,135	0,243	0,384	0,441	0,432	0,375	0,288	0,189	0,096	0,027	0,007
	2	0,007	0,027	0,096	0,189	0,288	0,375	0,432	0,441	0,384	0,243	0,135
	3	0+	0,001	0,008	0,027	0,064	0,125	0,216	0,343	0,512	0,729	0,857
4	**0**	0,815	0,656	0,410	0,240	0,130	0,063	0,026	0,008	0,002	0+	0+
	1	0,171	0,292	0,410	0,412	0,346	0,250	0,154	0,076	0,026	0,004	0+
	2	0,014	0,049	0,154	0,265	0,346	0,375	0,346	0,265	0,154	0,049	0,014
	3	0+	0,004	0,026	0,076	0,154	0,250	0,346	0,412	0,410	0,292	0,171
	4	0+	0+	0,002	0,008	0,026	0,063	0,130	0,240	0,410	0,656	0,815
5	**0**	0,774	0,590	0,328	0,168	0,078	0,031	0,010	0,002	0+	0+	0+
	1	0,204	0,328	0,410	0,360	0,259	0,156	0,077	0,028	0,006	0+	0+
	2	0,021	0,073	0,205	0,309	0,346	0,313	0,230	0,132	0,051	0,008	0,001
	3	0,001	0,008	0,051	0,132	0,230	0,313	0,346	0,309	0,205	0,073	0,021
	4	0+	0+	0,006	0,028	0,077	0,156	0,259	0,360	0,410	0,328	0,204
	5	0+	0+	0+	0,002	0,010	0,031	0,078	0,168	0,328	0,590	0,774
6	**0**	0,735	0,531	0,262	0,118	0,047	0,016	0,004	0,001	0+	0+	0+
	1	0,232	0,354	0,393	0,303	0,187	0,094	0,037	0,010	0,002	0+	0+
	2	0,031	0,098	0,246	0,324	0,311	0,234	0,138	0,060	0,015	0,001	0+
	3	0,002	0,015	0,082	0,185	0,276	0,313	0,276	0,185	0,082	0,015	0,002
	4	0+	0,001	0,015	0,060	0,138	0,234	0,311	0,324	0,246	0,098	0,031
	5	0+	0+	0,002	0,010	0,037	0,094	0,187	0,303	0,393	0,354	0,232
	6	0+	0+	0+	0,001	0,004	0,016	0,047	0,118	0,262	0,531	0,735
7	**0**	0,698	0,478	0,210	0,082	0,028	0,008	0,002	0+	0+	0+	0+
	1	0,257	0,372	0,367	0,247	0,131	0,055	0,017	0,004	0+	0+	0+
	2	0,041	0,124	0,275	0,318	0,261	0,164	0,077	0,025	0,004	0+	0+
	3	0,004	0,023	0,115	0,227	0,290	0,273	0,194	0,097	0,029	0,003	0+
	4	0+	0,003	0,029	0,097	0,194	0,273	0,290	0,227	0,115	0,023	0,004
	5	0+	0+	0,004	0,025	0,077	0,164	0,261	0,318	0,275	0,124	0,041
	6	0+	0+	0+	0,004	0,017	0,055	0,131	0,247	0,367	0,372	0,257
	7	0+	0+	0+	0+	0,002	0,008	0,028	0,082	0,210	0,478	0,698
8	**0**	0,663	0,430	0,168	0,058	0,017	0,004	0,001	0+	0+	0+	0+
	1	0,279	0,383	0,336	0,198	0,090	0,031	0,008	0,001	0+	0+	0+
	2	0,051	0,149	0,294	0,296	0,209	0,109	0,041	0,010	0,001	0+	0+
	3	0,005	0,033	0,147	0,254	0,279	0,219	0,124	0,047	0,009	0+	0+
	4	0+	0,005	0,046	0,136	0,232	0,273	0,232	0,136	0,046	0,005	0+
	5	0+	0+	0,009	0,047	0,124	0,219	0,279	0,254	0,147	0,033	0,005
	6	0+	0+	0,001	0,010	0,041	0,109	0,209	0,296	0,294	0,149	0,051
	7	0+	0+	0+	0,001	0,008	0,031	0,090	0,198	0,336	0,383	0,279
	8	0+	0+	0+	0+	0,001	0,004	0,017	0,058	0,168	0,430	0,663

Закінчення табл. 5

n	p / m	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95
9	**0**	0,630	0,387	0,134	0,040	0,010	0,002	0+	0+	0+	0+	0+
	1	0,299	0,387	0,302	0,156	0,060	0,018	0,004	0+	0+	0+	0+
	2	0,063	0,172	0,302	0,267	0,161	0,070	0,021	0,004	0+	0+	0+
	3	0,008	0,045	0,176	0,267	0,251	0,164	0,074	0,021	0,003	0+	0+
	4	0,001	0,007	0,066	0,172	0,251	0,246	0,167	0,074	0,017	0,001	0+
	5	0+	0,001	0,017	0,074	0,167	0,246	0,251	0,172	0,066	0,007	0,001
	6	0+	0+	0,003	0,021	0,074	0,164	0,251	0,267	0,176	0,045	0,008
	7	0+	0+	0+	0,004	0,021	0,070	0,161	0,267	0,302	0,172	0,063
	8	0+	0+	0+	0+	0,004	0,018	0,060	0,156	0,302	0,387	0,299
	9	0+	0+	0+	0+	0+	0,002	0,010	0,040	0,134	0,387	0,630
10	**0**	0,599	0,349	0,107	0,028	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,315	0,387	0,268	0,121	0,040	0,010	0,002	0,000	0,000	0,000	0,000
	2	0,075	0,194	0,302	0,233	0,121	0,044	0,011	0,001	0,000	0,000	0,000
	3	0,010	0,057	0,201	0,267	0,215	0,117	0,042	0,009	0,001	0,000	0,000
	4	0,001	0,011	0,088	0,200	0,251	0,205	0,111	0,037	0,006	0,000	0,000
	5	0+	0,001	0,026	0,103	0,201	0,246	0,201	0,103	0,026	0,001	0,000
	6	0+	0+	0,006	0,037	0,111	0,205	0,251	0,200	0,088	0,011	0,001
	7	0+	0+	0,001	0,009	0,042	0,117	0,215	0,267	0,201	0,057	0,010
	8	0+	0+	0+	0,001	0,011	0,044	0,121	0,233	0,302	0,194	0,075
	9	0+	0+	0+	0+	0,002	0,010	0,040	0,121	0,268	0,387	0,315
	10	0+	0+	0+	0+	0+	0,001	0,006	0,028	0,107	0,349	0,599
11	**0**	0,569	0,314	0,086	0,020	0,004	0+	0+	0+	0+	0+	0+
	1	0,329	0,384	0,236	0,093	0,027	0,005	0,001	0+	0+	0+	0+
	2	0,087	0,213	0,295	0,200	0,089	0,027	0,005	0,001	0+	0+	0+
	3	0,014	0,071	0,221	0,257	0,177	0,081	0,023	0,004	0+	0+	0+
	4	0,001	0,016	0,111	0,220	0,236	0,161	0,070	0,017	0,002	0+	0+
	5	0+	0,002	0,039	0,132	0,221	0,226	0,147	0,057	0,010	0+	0+
	6	0+	0+	0,010	0,057	0,147	0,226	0,221	0,132	0,039	0,002	0+
	7	0+	0+	0,002	0,017	0,070	0,161	0,236	0,220	0,111	0,016	0,001
	8	0+	0+	0+	0,004	0,023	0,081	0,177	0,257	0,221	0,071	0,014
	9	0+	0+	0+	0,001	0,005	0,027	0,089	0,200	0,295	0,213	0,087
	10	0+	0+	0+	0+	0,001	0,005	0,027	0,093	0,236	0,384	0,329
	11	0+	0+	0+	0+	0+	0+	0,004	0,020	0,086	0,314	0,569
12	**0**	0,540	0,282	0,069	0,014	0,002	0+	0+	0+	0+	0+	0+
	1	0,341	0,377	0,206	0,071	0,017	0,003	0+	0+	0+	0+	0+
	2	0,099	0,230	0,283	0,168	0,064	0,016	0,002	0+	0+	0+	0+
	3	0,017	0,085	0,236	0,240	0,142	0,054	0,012	0,001	0+	0+	0+
	4	0,002	0,021	0,133	0,231	0,213	0,121	0,042	0,008	0,001	0+	0+
	5	0+	0,004	0,053	0,158	0,227	0,193	0,101	0,029	0,003	0+	0+
	6	0+	0+	0,016	0,079	0,177	0,226	0,177	0,079	0,016	0+	0+
	7	0+	0+	0,003	0,029	0,101	0,193	0,227	0,158	0,053	0,004	0+
	8	0+	0+	0,001	0,008	0,042	0,121	0,213	0,231	0,133	0,021	0,002
	9	0+	0+	0+	0,001	0,012	0,054	0,142	0,240	0,236	0,085	0,017
	10	0+	0+	0+	0+	0,002	0,016	0,064	0,168	0,283	0,230	0,099
	11	0+	0+	0+	0+	0+	0,003	0,017	0,071	0,206	0,377	0,341
	12	0+	0+	0+	0+	0+	0+	0,002	0,014	0,069	0,282	0,540

Таблиця 6

ЗНАЧЕННЯ $t_{\gamma,\nu}$, ВИЗНАЧАЮТЬСЯ РІВНІСТЮ: $2\int_0^t f(x)dx = \gamma$,

ДЕ $f(x)$ — ЩІЛЬНІСТЬ РОЗПОДІЛУ СТЬЮДЕНТА З ν СТУПЕНЯМИ ВІЛЬНОСТІ

ν \ γ	0,75	0,8	0,85	0,9	0,95	0,99	0,999
1	2,414	3,078	4,165	6,314	12,706	63,657	636,619
2	1,604	1,886	2,282	2,920	4,303	9,925	31,599
3	1,423	1,638	1,924	2,353	3,182	5,841	12,924
4	1,344	1,533	1,778	2,132	2,776	4,604	8,610
5	1,301	1,476	1,699	2,015	2,571	4,032	6,869
6	1,273	1,440	1,650	1,943	2,447	3,707	5,959
7	1,254	1,415	1,617	1,895	2,365	3,499	5,408
8	1,240	1,397	1,592	1,860	2,306	3,355	5,041
9	1,230	1,383	1,574	1,833	2,262	3,250	4,781
10	1,221	1,372	1,559	1,812	2,228	3,169	4,587
11	1,214	1,363	1, 548	1,796	2,201	3,106	4,437
12	1,209	1,356	1,538	1,782	2,179	3,055	4,318
13	1,204	1,350	1,530	1,771	2,160	3,012	4,221
14	1,200	1,345	1,523	1,761	2,145	2,977	4,140
15	1,197	1,341	1,517	1,753	2,131	2,947	4,073
16	1,194	1,337	1,512	1,746	2,120	2,921	4,015
17	1,191	1,333	1,508	1,740	2,110	2,898	3,965
18	1,189	1,330	1,504	1,734	2,101	2,878	3,922
19	1,187	1,328	1,500	1,729	2,093	2,861	3,883
20	1,185	1,325	1,497	1,725	2,086	2,845	3,850
21	1,183	1,323	1,494	1,721	2,080	2,831	3,819
22	1,182	1,321	1,492	1,717	2,074	2,819	3,792
23	1,180	1,319	1,489	1,714	2,069	2,807	3,768
24	1,179	1,318	1,487	1,711	2,064	2,797	3,745
25	1,178	1,316	1,485	1,708	2,060	2,787	3,725
26	1,177	1,315	1,483	1,706	2,056	2,779	3,707
27	1,176	1,314	1,482	1,703	2,052	2,771	3,690
28	1,175	1,313	1,480	1,701	2,048	2,763	3,674
29	1,174	1,311	1,479	1,699	2,045	2,756	3,659
30	1,173	1,310	1,477	1,697	2,042	2,750	3,646

Таблиця 7

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ χ^2

ν \ α	**0,001**	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**	**0,999**
1	10,828	7,879	6,635	5,024	3,841	2,706	0,016	0,004	0,001	0,000	0,000	0,000
2	13,816	10,597	9,210	7,378	5,991	4,605	0,211	0,103	0,051	0,020	0,010	0,002
3	16,266	12,838	11,345	9,348	7,815	6,251	0,584	0,352	0,216	0,115	0,072	0,024
4	18,467	14,860	13,277	11,143	9,488	7,779	1,064	0,711	0,484	0,297	0,207	0,091
5	20,515	16,750	15,086	12,833	11,070	9,236	1,610	1,145	0,831	0,554	0,412	0,210
6	22,458	18,548	16,812	14,449	12,592	10,645	2,204	1,635	1,237	0,872	0,676	0,381
7	24,322	20,278	18,475	16,013	14,067	12,017	2,833	2,167	1,690	1,239	0,989	0,598
8	26,124	21,955	20,090	17,535	15,507	13,362	3,490	2,733	2,180	1,646	1,344	0,857
9	27,877	23,589	21,666	19,023	16,919	14,684	4,168	3,325	2,700	2,088	1,735	1,152
10	29,588	25,188	23,209	20,483	18,307	15,987	4,865	3,940	3,247	2,558	2,156	1,479
11	31,264	26,757	24,725	21,920	19,675	17,275	5,578	4,575	3,816	3,053	2,603	1,834
12	32,909	28,300	26,217	23,337	21,026	18,549	6,304	5,226	4,404	3,571	3,074	2,214
13	34,528	29,819	27,688	24,736	22,362	19,812	7,042	5,892	5,009	4,107	3,565	2,617
14	36,123	31,319	29,141	26,119	23,685	21,064	7,790	6,571	5,629	4,660	4,075	3,041
15	37,697	32,801	30,578	27,488	24,996	22,307	8,547	7,261	6,262	5,229	4,601	3,483
16	39,252	34,267	32,000	28,845	26,296	23,542	9,312	7,962	6,908	5,812	5,142	3,942
17	40,790	35,718	33,409	30,191	27,587	24,769	10,085	8,672	7,564	6,408	5,697	4,416
18	42,312	37,156	34,805	31,526	28,869	25,989	10,865	9,390	8,231	7,015	6,265	4,905
19	43,820	38,582	36,191	32,852	30,144	27,204	11,651	10,117	8,907	7,633	6,844	5,407
20	45,315	39,997	37,566	34,170	31,410	28,412	12,443	10,851	9,591	8,260	7,434	5,921
21	46,797	41,401	38,932	35,479	32,671	29,615	13,240	11,591	10,283	8,897	8,034	6,447
22	48,268	42,796	40,289	36,781	33,924	30,813	14,041	12,338	10,982	9,542	8,643	6,983
23	49,728	44,181	41,638	38,076	35,172	32,007	14,848	13,091	11,689	10,196	9,260	7,529
24	51,179	45,559	42,980	39,364	36,415	33,196	15,659	13,848	12,401	10,856	9,886	8,085
25	52,620	46,928	44,314	40,646	37,652	34,382	16,473	14,611	13,120	11,524	10,520	8,649
26	54,052	48,290	45,642	41,923	38,885	35,563	17,292	15,379	13,844	12,198	11,160	9,222
27	55,476	49,645	46,963	43,195	40,113	36,741	18,114	16,151	14,573	12,879	11,808	9,803
28	56,892	50,993	48,278	44,461	41,337	37,916	18,939	16,928	15,308	13,565	12,461	10,391
29	58,301	52,336	49,588	45,722	42,557	39,087	19,768	17,708	16,047	14,256	13,121	10,986
30	59,703	53,672	50,892	46,979	43,773	40,256	20,599	18,493	16,791	14,953	13,787	11,588
31	61,098	55,003	52,191	48,232	44,985	41,422	21,434	19,281	17,539	15,655	14,458	12,196
32	62,487	56,328	53,486	49,480	46,194	42,585	22,271	20,072	18,291	16,362	15,134	12,811
33	63,870	57,648	54,776	50,725	47,400	43,745	23,110	20,867	19,047	17,074	15,815	13,431
34	65,247	58,964	56,061	51,966	48,602	44,903	23,952	21,664	19,806	17,789	16,501	14,057
35	66,619	60,275	57,342	53,203	49,802	46,059	24,797	22,465	20,569	18,509	17,192	14,688
36	67,985	61,581	58,619	54,437	50,998	47,212	25,643	23,269	21,336	19,233	17,887	15,324
37	69,346	62,883	59,893	55,668	52,192	48,363	26,492	24,075	22,106	19,960	18,586	15,965
38	70,703	64,181	61,162	56,896	53,384	49,513	27,343	24,884	22,878	20,691	19,289	16,611
39	72,055	65,476	62,428	58,120	54,572	50,660	28,196	25,695	23,654	21,426	19,996	17,262

Закінчення табл. 7

α / ν	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
40	73,402	66,766	63,691	59,342	55,758	51,805	29,051	26,509	24,433	22,164	20,707	17,916
41	74,745	68,053	64,950	60,561	56,942	52,949	29,907	27,326	25,215	22,906	21,421	18,575
42	76,084	69,336	66,206	61,777	58,124	54,090	30,765	28,144	25,999	23,650	22,138	19,239
43	77,419	70,616	67,459	62,990	59,304	55,230	31,625	28,965	26,785	24,398	22,859	19,906
44	78,750	71,893	68,710	64,201	60,481	56,369	32,487	29,787	27,575	25,148	23,584	20,576
45	80,077	73,166	69,957	65,410	61,656	57,505	33,350	30,612	28,366	25,901	24,311	21,251
46	81,400	74,437	71,201	66,617	62,830	58,641	34,215	31,439	29,160	26,657	25,041	21,929
47	82,720	75,704	72,443	67,821	64,001	59,774	35,081	32,268	29,956	27,416	25,775	22,610
48	84,037	76,969	73,683	69,023	65,171	60,907	35,949	33,098	30,755	28,177	26,511	23,295
49	85,351	78,231	74,919	70,222	66,339	62,038	36,818	33,930	31,555	28,941	27,249	23,983
50	86,661	79,490	76,154	71,420	67,505	63,167	37,689	34,764	32,357	29,707	27,991	24,674
51	87,968	80,747	77,386	72,616	68,669	64,295	38,560	35,600	33,162	30,475	28,735	25,368
52	89,272	82,001	78,616	73,810	69,832	65,422	39,433	36,437	33,968	31,246	29,481	26,065
53	90,573	83,253	79,843	75,002	70,993	66,548	40,308	37,276	34,776	32,018	30,230	26,765
54	91,872	84,502	81,069	76,192	72,153	67,673	41,183	38,116	35,586	32,793	30,981	27,468
55	93,168	85,749	82,292	77,380	73,311	68,796	42,060	38,958	36,398	33,570	31,735	28,173
56	94,461	86,994	83,513	78,567	74,468	69,919	42,937	39,801	37,212	34,350	32,490	28,881
57	95,751	88,236	84,733	79,752	75,624	71,040	43,816	40,646	38,027	35,131	33,248	29,592
58	97,039	89,477	85,950	80,936	76,778	72,160	44,696	41,492	38,844	35,913	34,008	30,305
59	98,324	90,715	87,166	82,117	77,931	73,279	45,577	42,339	39,662	36,698	34,770	31,020
60	99,607	91,952	88,379	83,298	79,082	74,397	46,459	43,188	40,482	37,485	35,534	31,738
65	105,99	98,11	94,42	89,18	84,82	79,97	50,88	47,45	44,60	41,44	39,38	35,36
70	112,32	104,21	100,43	95,02	90,53	85,53	55,33	51,74	48,76	45,44	43,28	39,04
75	118,60	110,29	106,39	100,84	96,22	91,06	59,79	56,05	52,94	49,48	47,21	42,76
80	124,84	116,32	112,33	106,63	101,88	96,58	64,28	60,39	57,15	53,54	51,17	46,52
85	131,04	122,32	118,24	112,39	107,52	102,08	68,78	64,75	61,39	57,63	55,17	50,32
90	137,21	128,30	124,12	118,14	113,15	107,57	73,29	69,13	65,65	61,75	59,20	54,16
95	143,34	134,25	129,97	123,86	118,75	113,04	77,82	73,52	69,92	65,90	63,25	58,02
100	149,45	140,17	135,81	129,56	124,34	118,50	82,36	77,93	74,22	70,06	67,33	61,92

Таблиця 8

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ СТЬЮДЕНТА

ν \ α	**0,001**	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**
1	318,309	63,657	31,821	12,706	6,314	3,078
2	22,327	9,925	6,965	4,303	2,920	1,886
3	10,215	5,841	4,541	3,182	2,353	1,638
4	7,173	4,604	3,747	2,776	2,132	1,533
5	5,893	4,032	3,365	2,571	2,015	1,476
6	5,208	3,707	3,143	2,447	1,943	1,440
7	4,785	3,499	2,998	2,365	1,895	1,415
8	4,501	3,355	2,896	2,306	1,860	1,397
9	4,297	3,250	2,821	2,262	1,833	1,383
10	4,144	3,169	2,764	2,228	1,812	1,372
11	4,025	3,106	2,718	2,201	1,796	1,363
12	3,930	3,055	2,681	2,179	1,782	1,356
13	3,852	3,012	2,650	2,160	1,771	1,350
14	3,787	2,977	2,624	2,145	1,761	1,345
15	3,733	2,947	2,602	2,131	1,753	1,341
16	3,686	2,921	2,583	2,120	1,746	1,337
17	3,646	2,898	2,567	2,110	1,740	1,333
18	3,610	2,878	2,552	2,101	1,734	1,330
19	3,579	2,861	2,539	2,093	1,729	1,328
20	3,552	2,845	2,528	2,086	1,725	1,325
21	3,527	2,831	2,518	2,080	1,721	1,323
22	3,505	2,819	2,508	2,074	1,717	1,321
23	3,485	2,807	2,500	2,069	1,714	1,319
24	3,467	2,797	2,492	2,064	1,711	1,318
25	3,450	2,787	2,485	2,060	1,708	1,316
26	3,435	2,779	2,479	2,056	1,706	1,315
27	3,421	2,771	2,473	2,052	1,703	1,314
28	3,408	2,763	2,467	2,048	1,701	1,313
29	3,396	2,756	2,462	2,045	1,699	1,311
30	3,385	2,750	2,457	2,042	1,697	1,310
40	3,307	2,704	2,423	2,021	1,684	1,303
50	3,261	2,678	2,403	2,009	1,676	1,299
60	3,232	2,660	2,390	2,000	1,671	1,296
120	3,160	2,617	2,358	1,980	1,658	1,289
∞	3,090	2,576	2,326	1,960	1,645	1,282

Таблиця 9

КРИТИЧНІ ТОЧКИ РОЗПОДІЛУ ФІШЕРА-СНЕДЕКОРА $F(\nu_1, \nu_2)$

(ν_1 — КІЛЬКІСТЬ СТУПЕНІВ ВІЛЬНОСТІ ЧИСЕЛЬНИКА, ν_2 — КІЛЬКІСТЬ СТУПЕНІВ ВІЛЬНОСТІ ЗНАМЕННИКА)

Рівень значущості 0,01																			
ν_2	ν_1																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Рівень значущості 0,025																			
ν_2	ν_1																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28	968,63	976,71	984,87	993,10	997,25	1001	1006	1010	1014	1018
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

Рівень значущості 0,05																			
ν_2	ν_1																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Рівень значущості 0,1																			
v2	***v1***																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61

Закінчення таблиці 9

Рівень значущості 0,1																			
v2	***v1***																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	∞
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

Таблиця 10

ЗНАЧЕННЯ $q = q(\gamma, n)$

n	γ			**n**	γ		
	0,95	**0,99**	**0,999**		**0,95**	**0,99**	**0,999**
5	1,37	2,67	5,64	**20**	0,37	0,58	0,88
6	1,09	2,01	3,88	**25**	0,32	0,49	0,73
7	0,92	1,62	2,98	**30**	0,28	0,43	0,63
8	0,80	1,38	2,42	**35**	0,26	0,38	0,56
9	0,71	1,20	2,06	**40**	0,24	0,35	0,50
10	0,65	1,08	1,80	**45**	0,22	0,32	0,46
11	0,59	0,98	1,60	**50**	0,21	0,30	0,43
12	0,55	0,90	1,45	**60**	0,188	0,269	0,38
13	0,52	0,83	1,33	**70**	0,174	0,245	0,34
14	0,48	0,78	1,23	**80**	0,161	0,226	0,31
15	0,46	0,73	1,15	**90**	0,151	0,211	0,29
16	0,44	0,70	1,07	**100**	0,143	0,198	0,27
17	0,42	0,66	1,01	**150**	0,115	0,160	0,211
18	0,40	0,63	0,96	**200**	0,099	0,136	0,185
19	0,39	0,60	0,92	**250**	0,089	0,120	0,162

Таблиця 11

КРИТИЧНІ ТОЧКИ λ_α **РОЗПОДІЛУ КОЛМОГОРОВА:** $P(\lambda > \lambda_\alpha) = \alpha$

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

Таблиця 12

КРИТИЧНІ ТОЧКИ R_α **КРИТЕРІЮ ЗНАКІВ:** $P(r \le R_\alpha) = \alpha$

n	α			**n**	α			**n**	α		
	0,01	**0,025**	**0,05**		**0,01**	**0,025**	**0,05**		**0,01**	**0,025**	**0,05**
5			0	**34**	9	10	11	**63**	21	23	24
6		0	0	**35**	10	11	12	**64**	22	23	24
7	0	0	0	**36**	10	11	12	**65**	22	24	25
8	0	0	1	**37**	10	12	13	**66**	23	24	25
9	0	1	1	**38**	11	12	13	**67**	23	25	26
10	0	1	1	**39**	11	12	13	**68**	23	25	26
11	1	1	2	**40**	12	13	14	**69**	24	25	27
12	1	2	2	**41**	12	13	14	**70**	24	26	27
13	1	2	3	**42**	13	14	15	**71**	25	26	28
14	2	2	3	**43**	13	14	15	**72**	25	27	28
15	2	3	3	**44**	13	15	16	**73**	26	27	28
16	2	3	4	**45**	14	15	16	**74**	26	28	29
17	3	4	4	**46**	14	15	16	**75**	26	28	29
18	3	4	5	**47**	15	16	17	**76**	27	28	30
19	4	4	5	**48**	15	16	17	**77**	27	29	30
20	4	5	5	**49**	15	17	18	**78**	28	29	31
21	4	5	6	**50**	16	17	18	**79**	28	30	31
22	5	5	6	**51**	16	18	19	**80**	29	30	32
23	5	6	7	**52**	17	18	19	**81**	29	31	32
24	5	6	7	**53**	17	18	20	**82**	30	31	33
25	6	7	7	**54**	18	19	20	**83**	30	32	33
26	6	7	8	**55**	18	19	20	**84**	30	32	33
27	7	7	8	**56**	18	20	21	**85**	31	32	34
28	7	8	9	**57**	19	20	21	**86**	31	33	34
29	7	8	9	**58**	19	21	22	**87**	32	33	35
30	8	9	10	**59**	20	21	22	**88**	32	34	35
31	8	9	10	**60**	20	21	23	**89**	33	34	36
32	8	9	10	**61**	20	22	23	**90**	33	35	36
33	9	10	11	**62**	21	22	24	**91**	33	35	37

ПРЕДМЕТНИЙ ПОКАЖЧИК

ЛІТЕРАТУРА

1. *Айвазян С.А.* Прикладная статистика: Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М. : Финансы и статистика, 1983. — 472 с.
2. *Айвазян С. А.* Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. — М. : Изд. об. «ЮНИТИ», 1998. — 1080 с.
3. *Барковський В. В.* Теорія ймовірностей та математична статистика / В. В. Барковський, Н. В. Барковська, О. К. Лопатін. — К. : ЦНЛ, 2002. — 448 с.
4. *Боровков А. А.* Математическая статистика : — учебник / А. А. Боровков. — М. : Наука. Главная редакция физ.-матю лит.-ры, 1984. — 472 с.
5. *Ван-дер-Варден Б. Л.* Математическая статистика / Б. Л. Ван-дер-Варден; пер. с нем. — М. : ИЛ, 1960. — 434 с.
6. *Григулич С. М., Лісовська В. П., Макаренко О. І., Пахомов І. І., Стасюк В. Д., Черніс І. М.* Теорія ймовірності для економістів. — К. : КНЕУ, 2012. — 307 с.
7. *Гмурман В. Е.* Руководство к решению задач по теории вероятностей и математической статистике : учеб. пособие для студ. вузов / В. Е. Гмурман. — М. : Высш. шк., 2005. — 404 с.
8. *Гмурман В. Е.* Теория вероятностей и математическая статистика : учеб. пособие / В. Е. Гмурман. — 12-е изд. — М. : Высшее образование, 2007 — 479 с.
9. *Дунин-Барковский И. В., Смирнов Н. В.* Теория вероятностей и математическая статистика в технике / И. В. Дунин-Барковский, Н. В. Смирнов. — М. : Гос. изд-во технико-теоретич. лит-ры, 1955. — 555 с.
10. *Елисеева И. И.* Общая теория статистики : учебник / И. И. Елисеева, М. М. Юзбашев; под ред. И. И. Елисееваой. — 5-е изд., перераб. и доп. — М. : Финансы и статистика, 2004. — 656 с.
11. *Ивченко Г. И.* Математическая статистика : учебн. пособие для вузов / Г. И. Ивченко, Ю. И. Медведев. — М. : Высш. шк., 1984. — 248 с.
12. *Кармелюк Г. І.* Теорія ймовірностей та математична статистика. Посібник з розв'язування задач : навч. посіб. / Г. І. Кармелюк. — К. : Центр учбової літератури, 2007 — 576 с.
13. *Карташов М. В.* Імовірність, процеси, статистика : посібник / М. В. Карташов. — К. : Видавполіграфіч. центр «Київський університет», 2008. — 494 с.
14. *Кендалл М.* Статистические выводы и связи / М. Кендалл, А. Стьюарт. — Главная ред. физ.-математич. лит-ры изд-ва «Наука», 1973. — 899 с.
15. *Кибзун А. И.* Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами : учебн. Пособие / А. И. Кибзун, Е. Р. Горяинова, А. В. Наумов, А. Н. Сиротин. — М. : ФИЗМАТЛИТ, 2002. — 224 с.
16. *Ковальчук Т. М.* — Теорія ймовірностей і математична статистика для менеджерів / [Т. М. Ковальчук, В. А.Ковальчук, Є. К. Бабець та ін.] — К. : Вид. дім «Професіонал», 2006. — 256 с.
17. *Козлов М. В.* Введение в математическую статистику / М. В. Козлов. — М. : Изд-во МГУ, 1987. — 264 с.
18. *Колемаев В. А.* Теория вероятностей и математическая статистика / В. А. Колемаев, В. Н. Калинина : учебник / под ред. В. А.Колемаева. — М. : ИНФРА-М, 1997. — 302 с.
19. *Крамер Г.* Математические методы статистики. — 2-е изд., стер. / *Г. Крамер* — М. : Из-во Мир, 1975. — 648 с.
20. *Кремер Н. Ш.* Теория вероятностей и математическая статистика : учебник для вузов / Н. Ш. Кремер. — М. : ЮНИТИ — ДАНА, 2003. — 543 с.
21. *Мхитарян В. С.* Математическая статистика / В. С. Мхитарян, Л. И. Трошин. — М. : МЭСИ, 1975.
22. *Ниворожкина Л. П.* Основы статистики с элементами теории вероятностей для економистов: Руководство для решения задач / [Л. П. Ниворожкина, М. З. А.орозова, И. А. Герасимова, И. В. Житников] — Ростов н/Д. : Феникс, 1999. — 320 с.
23. *Норманн Дрейер.* Прикладной регрессионный анализ / Н. Дрейер. Г. Смит. — Диалектика, 2007. — 912 с.
24. *Осипчук М. М.* Теорія ймовірностей, випадкові процеси та математична статистика 4.2. Математична статистика : Конспект лекцій / М. М. Осипчук. — Івано-Франківськ : Факел, 2003. — 85 с.
25. *Панков А. Р.* Практикум по математической статистике : учеб. пособие / А. Р. Панков, Е. Н. Платонов. — М. : Изд-во МАИ, 2006. — 86 с.
26. *Письменный Д. Т.* Конспект лекций по теории вероятностей, математической статистике и случайным процессам / Д. Т. Письменный. — М. : Айрис-пресс, 2006. — 288 с.
27. *Пугачев В. С.* Теория вероятностей и математическая статистика : учеб. пособие. — 2-е изд. исправл. и дополн / В. С. Пугачев. — М.: ФИЗМАТЛИТ, 2002. — 496 с.
28. *Романовский В. И.* Элементарный курс математической статистики / В. И. Романовский. — М.: ГОСПЛАиздат, 1939. — 360 с.
29. *Чернова Н. И.* Математическая статистика : учебн. Пособие / Н. И. Чернова. — Новосибирск : Изд-во Новосиб. гос. ун-та, —2007. — 148 с.

ЗМІСТ

Навчальне видання

ГРИГУЛИЧ Світлана Миколаївна
ЛІСОВСЬКА Валентина Петрівна
МАКАРЕНКО Олександр Іванович та ін.

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

Редактор *Л. Гордієнко*
Коректор *Л. Гримальська*
Верстка *І. Грибанової*

Підп. до друку 00.00.15. Формат 60×84/8
Друк. арк. 14,96. Зам. № 12-4448

Державний вищий навчальний заклад
«Київський національний економічний університет імені Вадима Гетьмана»
03680, м. Київ, проспект Перемоги, 54/1

Свідоцтво про внесення до Державного реєстру
суб'єктів видавничої справи (серія ДК, № 235 від 07.11.2000)

Тел./факс (044) 537-61-41; тел. (044) 537-61-44
E-mail: publish@kneu.kiev.ua