МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«ЗАПОРІЗЬКА ПОЛІТЕХНІКА»

І. М. Килимник

ПРАКТИКУМ З ЕЛЕМЕНТІВ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Навчальний посібник

Запоріжжя • НУ «Запорізька політехніка» • 2026

УДК 517.91(075.8)
К 39

Рекомендовано до друку Вченою радою
Національного університету «Запорізька політехніка»,
(Протокол № 12 від 26.05.2026 р.)

Рецензенти:

Туровцев Г. В., доктор фізико-математичних наук, професор, ректор Запорізького інституту економіки та інформаційних технологій;

Гребенюк С. М., доктор технічних наук, професор, завідувач кафедри фундаментальної та прикладної математики Запорізького національного університету.

К39 **Килимник І. М.**

Практикум з елементів математичної статистики : навч. посіб. Запоріжжя : НУ «Запорізька політехніка», 2026. 302 с.

ISBN 978-617-529-554-0

У посібнику системно викладено курс математичної статистики, що відповідає типовій програмі вищої математики для технічних спеціальностей. Видання має автономний характер: теоретичний матеріал представлений у повному обсязі, що дозволяє опанувати дисципліну без опрацювання додаткової літератури.

Кожна тема супроводжується розгорнутим розбором типових задач, а для закріплення навичок запропоновано комплекс вправ для самостійного виконання з відповідями.

Навчальний посібник «Практикум з елементів математичної статистики» розрахований на студентів денної та заочної (дистанційної) форм навчання інженерних факультетів. Посібник також може бути корисним для здобувачів інших спеціальностей, навчальний план яких передбачає вивчення методів математичної статистики.

УДК 517.91(075.8)

ISBN 978-617-529-554-0

ЗМІСТ

ВСТУП

Математична статистика – це прикладна, практична наука, що вивчає великі сукупності однотипних об'єктів. У 19-му столітті вона виділилася з теорії ймовірностей і з тих пір вважається самостійною наукою.

Математична статистика користується методами різних розділів математики. Але насамперед вона користується методами теорії ймовірностей, яка є для неї основною теоретичною базою. Теорія ймовірностей надає математичну основу аналізу випадкових подій.

Математична статистика займається збором, аналізом, інтерпретацією, поданням та організацією даних. Вона дозволяє робити висновки та приймати рішення на основі аналізу великих обсягів інформації.

Деякі ключові аспекти математичної статистики:

- *Збір даних*: Математична статистика розробляє методи збору даних, щоб забезпечити їх репрезентативність та надійність. Це може включати проведення опитувань, експериментів або спостережень.

- *Аналіз даних*: Статистичні методи використовуються для аналізу даних та виявлення закономірностей, зв'язків та тенденцій. Це може включати розрахунок середніх значень, дисперсій, кореляцій та інших статистичних показників.

- *Інтерпретація результатів*: Математична статистика допомагає інтерпретувати результати аналізу даних та робити висновки про генеральну сукупність на основі вибірки.

- *Прийняття рішень*: Статистичні методи використовуються для прийняття обґрунтованих рішень у різних галузях, таких як наука, медицина, економіка та бізнес.

Основне завдання математичної статистики полягає у розробці методів аналізу статистичних даних відповідно до мети дослідження, а саме:

1) методологія збору та систематизації: розробка способів накопичення та групування статистичної інформації (зокрема для великих масивів даних);

2) ідентифікація розподілу: визначення закону розподілу випадкової величини або системи величин на основі отриманих даних;

3) оцінювання параметрів: знаходження точкових та інтервальних оцінок невідомих параметрів розподілу;

4) перевірка статистичних гіпотез: підтвердження або спростування припущень щодо виду закону розподілу, характеру зв'язку між величинами чи значень їхніх параметрів.

Основні завдання математичної статистики:

- Статистичне оцінювання параметрів законів розподілу.
- Статистична перевірка гіпотез.

Застосування математичної статистики. Математична статистика знаходить застосування у багатьох областях, включаючи:

-*Наукові дослідження*: для аналізу результатів експериментів та спостережень.

-*Медицина*: для оцінки ефективності лікування та діагностики захворювань.

-*Економіка*: аналіз економічних показників, прогнозування ринкових тенденцій.

-*Бізнес*: для аналізу ринку, оцінки ризиків та прийняття управлінських рішень.

-*Соціологія*: аналіз результатів опитувань, вивчення соціальних тенденцій.

-*Психологія*: обробка даних психологічних тестів, вивчення впливу різних чинників поведінки.

-*Біологія*: опрацювання результатів біологічних експериментів, аналіз генетичних даних.

-*Інформатика*: аналіз даних, машинне навчання, штучний інтелект.

Студенти використовують математичну статистику у різних аспектах своєї навчальної та дослідницької діяльності. Наприклад:

Написання курсових та дипломних(кваліфікаційних) робіт:

Студенти, які займаються дослідженнями у різних галузях (соціологія, психологія, економіка, біологія та ін.), застосовують методи математичної статистики для аналізу зібраних даних, перевірки гіпотез та отримання достовірних результатів.

Вони використовують статистичні методи для обробки результатів опитувань, експериментів та спостережень.

Аналіз даних у навчальних проєктах:

У багатьох навчальних курсах студенти виконують проєкти, які потребують аналізу даних. Математична статистика допомагає їм виявляти закономірності, будувати моделі та робити висновки на основі даних.

Наукові дослідження:

Під час проведення наукових досліджень студенти повинні вміти правильно зібрати, обробити та проаналізувати дані. Математична статистика надає необхідний інструментарій.

Загалом математична статистика є важливим інструментом для студентів, що дозволяє їм проводити якісні дослідження, аналізувати дані та робити обґрунтовані висновки.

Автор висловлює подяку рецензентам доктору фізико-математичних наук, професору Туровцеву Г. В., доктору технічних наук, професору Гребенюку С. М. за корисні зауваження та поради, що сприяли покращенню цього видання.

ЕЛЕМЕНТИ МАТЕМАТИЧНОЇ СТАТИСТИКИ

1 Поняття генеральної сукупності та вибірки. Вибірковий метод

Поняття генеральної сукупності (ГС) та вибірки є фундаментальними у статистиці та дослідженнях. Вони використовуються для проведення аналізу та формування висновків про великі групи об'єктів чи явищ, коли вивчення кожного елемента неможливе чи недоцільне.

Означення 1.1. *Генеральна сукупність* (англ. population) (або *статистична сукупність*) – це безліч усіх об'єктів, одиниць чи спостережень, щодо яких передбачається робити висновки у межах конкретного дослідження.

Об'єкти, що входять у генеральну сукупність, називаються її *елементами*, загальна кількість яких число N – її *обсягом*. Поняття генеральної сукупності у певному сенсі аналогічне поняттю випадкової величини (закону розподілу ймовірностей, імовірнісному простору), оскільки цілком зумовлено певним комплексом умов.

Ключові характеристики генеральної сукупності:

• *Всеосяжність*: Вона включає абсолютно всі елементи, що володіють заданими характеристиками, які є предметом вивчення.

• *Гіпотетичність*: Часто генеральна сукупність є гіпотетичною чи надто великою, щоб її можна було повністю дослідити. Наприклад, "всі жителі Землі" або "всі можливі наслідки кидка монети".

• *Визначається метою дослідження*: Те, що буде генеральною сукупністю, безпосередньо залежить від поставленого завдання. Наприклад, якщо ми вивчаємо думку студентів НУ "ЗП", то генеральною сукупністю будуть усі студенти НУ "ЗП". Якщо ж ми вивчаємо думку студентів по всій Україні, то генеральною сукупністю будуть усі студенти України.

Приклади генеральної сукупності:

• Усі дорослі жителі міста Запоріжжя.

• Усі товари певного типу, вироблені на конкретному заводі протягом року.

• Усі медичні карти пацієнтів із певним захворюванням у лікарні.

• Усі можливі результати експерименту з підкидання грального кубика.

Означення 1.2. *Вибірка* (англ. *sample*) (або *вибіркова сукупність*) – це частина генеральної сукупності, яка охоплюється експериментом, спостереженням, опитуванням чи іншим методом дослідження.

Це підмножина об'єктів (елементів), що фактично вивчається, щоб зробити висновки про всю генеральну сукупність. Кількість об'єктів (елементів) n у вибірці називається її *обсягом* (або *об'ємом*).

Вибірки поділяють на повторні і безповторні.

Означення 1.3. *Повторна вибірка* - це вибірка, коли відібраний випадковим чином об'єкт обов'язково повертається у генеральну сукупність перед відбором наступного об'єкта.

Означення 1.4. *Безповторна вибірка* - це вибірка, коли відібраний випадковим чином об'єкт більше у генеральну сукупність не повертається.

Ключові характеристики вибірки:

• *Частина ГС*: Вибір завжди є підмножиною генеральної сукупності.

• *Мета – репрезентативність*: Головна мета формування вибірки – щоб вона була репрезентативною. Це означає, що вибірка повинна відображати основні характеристики генеральної сукупності у тих самих пропорціях. Наприклад, якщо в генеральній сукупності 55% жінок і 45% чоловіків, то й у репрезентативній вибірці має бути таке саме співвідношення.

• *Використовується для висновків*: Аналізуючи дані, отримані з вибірки, дослідники прагнуть зробити узагальнення та висновки, які будуть застосовні до всієї генеральної сукупності.

Надалі будемо розглядати безповторні і репрезентативні вибірки.

Щоб вибірка була репрезентативною, використовують різні методи формування вибірки. Основні принципи:

• Випадковість: вибір елементів у вибірку має бути випадковим, щоб кожен елемент генеральної сукупності мав рівні шанси бути включеним. Це допомагає уникнути упередженості.

• Розмір вибірки: Достатній розмір вибірки є важливим для отримання статистично значущих результатів. Занадто маленька вибірка може не відображати різноманітність генеральної сукупності.

• Стратифікація (за потреби): Якщо генеральна сукупність складається з різних підгруп (страт), які мають важливе значення для дослідження (наприклад, за віком, статтю, доходом), то вибірка може бути стратифікованою. Це означає, що з кожної страти відбирається певна кількість елементів пропорційна їх частці в генеральній сукупності.

Методи формування вибірки: Існують різні методи формування вибірки, які поділяються на *ймовірнісні* (випадкові) та *неймовірнісні* (невипадкові).

Імовірнісні методи (забезпечують репрезентативність):

• Проста випадкова вибірка: Кожен елемент генеральної сукупності має рівні шанси потрапити у вибірку. Це може бути реалізовано за допомогою генератора випадкових чисел або лотереї.

• Систематична вибірка: Вибирається кожен k-й елемент із впорядкованого списку генеральної сукупності, починаючи з випадково вибраної точки.

• Стратифікована (типологічна) вибірка: Генеральна сукупність поділяється на однорідні підгрупи (страти), а потім із кожної страти випадково відбираються елементи. Це гарантує представництво всіх важливих груп.

• Кластерна (гніздова) вибірка: Генеральна сукупність ділиться на кластери (групи), а потім випадково вибираються кілька кластерів, і всі елементи всередині цих вибраних кластерів включаються у вибірку. Ефективно для великих географічно розподілених сукупностей.

Неймовірнісні методи (не завжди гарантують репрезентативність, але можуть бути корисні у пошуковому дослідженні):

• Зручна (випадкова) вибірка: Вибираються ті елементи, які є найбільш доступними та зручними для дослідника.

• Цільова (передбачувана) вибірка: Дослідник вибирає елементи, які, на його думку, найкраще відповідають цілям дослідження.

• Квотна вибірка: Дослідник встановлює квоти для різних груп населення та набирає необхідну кількість респондентів для кожної квоти.

• Вибір "снігової грудки": Перші респонденти вибираються випадково, а потім вони рекомендують інших респондентів з тієї ж групи. Використовується для недоступних груп.

Приклади вибірки:

• 500 випадково обраних дорослих мешканців Запоріжжя, опитаних про їхнє ставлення до місцевих новин.

• 100 випадкових зразків товарів із виробничої лінії, перевірені на дефекти.

• Медичні дані 30 пацієнтів, які брали участь у клінічному випробуванні нових ліків.

Взаємозв'язок генеральної сукупності та вибірки

Взаємозв'язок між генеральною сукупністю та вибіркою є основним у статистичному аналізі:

1) *Вибірка витягується з генеральної сукупності*: Для проведення дослідження, коли немає можливості чи необхідності вивчати всю генеральну сукупність, з неї формується вибірка.

2) *Висновки про ГС робляться з урахуванням вибірки*: Мета вивчення вибірки – отримати інформацію, яку можна екстраполювати (перенести) на всю генеральну сукупність.

3) *Репрезентативність – ключ до достовірності*: Чим репрезентативною є вибірка, тим точнішими та надійнішими будуть висновки, зроблені про генеральну сукупність. Якщо вибірка нерепрезентативна (зміщена), висновки можуть бути помилковими.

Вибірковий метод

Найчастіше, при вивченні якоїсь характеристики великої групи об'єктів (людей, товарів, подій), званої генеральної сукупністю, неможливо чи недоцільно досліджувати кожен об'єкт. Наприклад, неможливо протестувати на міцність усі вироблені лампочки (це їх зруйнує). У таких ситуаціях на допомогу приходить вибірковий метод.

Означення 1.5. *Вибірковий метод* – це статистичний метод дослідження загальних властивостей генеральної сукупності з урахуванням вивчення властивостей лише частини цих об'єктів, званої вибіркою.

Вибірковий метод потрібен для:

• *Економія ресурсів*: Дослідити всю сукупність часто дуже дорого та трудомістко.

• *Економія часу*: Отримання даних з усієї сукупності може тривати надто багато часу.

• *Практична неможливість*: У деяких випадках генеральна сукупність може бути нескінченною або доступ до неї обмежений (наприклад, тестування ліків на всіх людях у світі).

• *Руйнівний контроль*: Іноді дослідження об'єкта призводить до його псування або знищення (наприклад, перевірка якості лампочок на термін служби).

• *Репрезентативність вибірки:* Дуже важливо, щоб вибірка була репрезентативною, тобто адекватно відображала властивості генеральної сукупності. Якщо вибірка не є репрезентативною, висновки, зроблені на її основі, можуть бути помилковими.

Основні поняття вибіркового методу:

• Генеральна сукупність (ГС): Вся сукупність об'єктів, явищ або подій, про які ми хочемо зробити висновки. Її обсяг N зазвичай дуже великий чи нескінченний.

• Вибірка (вибіркова сукупність): частина об'єктів, відібраних із генеральної сукупності для безпосереднього вивчення. Її обсяг n значно менший за N.

• Репрезентативність вибірки: Це властивість вибірки, що означає, що вона адекватно відбиває структуру та властивості генеральної сукупності. Якщо вибірка репрезентативна, то висновки, зроблені на її основі, можна з достатньою мірою впевненості поширити на всю генеральну сукупність. Для забезпечення репрезентативності використовують різні методи формування вибірки (наприклад, випадковий відбір, стратифікований відбір і т.д.).

Цілі вибіркового методу:

• Отримання достовірної інформації про генеральну сукупність за мінімальних витрат часу, ресурсів та зусиль.

• Можливість вивчення руйнівних об'єктів.

Приклад використання вибіркового методу: великий виробник чіпсів хоче оцінити середню масу чіпсів у пачці, що випускається на конвеєрі. Виробляються мільйони пачок на день.

Тоді

• генеральна сукупність: усі пачки чіпсів, вироблені протягом дня (чи певний період);

• проблема: неможливо зважити кожну пачку;

• вибірковий метод: виробник вибирає (наприклад, випадково, кожну 100 пачку) 1000 пачок чіпсів із загального потоку.

• вибірка: ці 1000 пачок.

• вивчення вибірки: зважується кожна з 1000 пачок, розраховується середня маса, розкид значень.

На основі даних по цих 1000 пачках робляться висновки про середню масу та мінливість маси для всіх вироблених пачок, а також про те, чи відповідають вони встановленим стандартам. Якщо вибірка була репрезентативною, ці висновки будуть досить точними.

2 Способи збирання інформації

Способи збирання інформації – це основа будь-якої статистичної роботи, адже даних немає і статистики. Вибір методу збору даних критично важливий, оскільки він впливає на якість, достовірність та застосовність отриманих результатів.

У математичній статистиці (і не тільки) використовуються різні методи збирання інформації, які можна розділити на кілька основних категорій:

1) Спостереження

Означення 2.1. *Спостереження* – це цілеспрямоване, планомірне та систематичне сприйняття явищ, процесів та фактів для отримання первинних даних.

Воно може бути:

• Безпосереднє спостереження. Дослідник особисто присутній і фіксує те, що відбувається. Наприклад, спостереження за поведінкою покупців у магазині; підрахунок кількості автомобілів, які проїжджають через певне перехрестя; етнографічні дослідження, коли вчений живе серед групи, що вивчається.

• Опосередковане спостереження. Збір даних здійснюється не безпосередньо, а через прилади, записи чи інші джерела. Наприклад, збирання даних про температуру повітря з метеостанцій; використання відеокамер для аналізу трафіку; аналіз даних із державних реєстрів чи баз даних.

Переваги: Дозволяє отримати дані про реальну поведінку без впливу на об'єкт, що спостерігається, часто надає об'єктивну інформацію.

Недоліки: Пасивний метод, що не дозволяє виявити причинно-наслідкові зв'язки, може бути трудомістким і вимагати багато часу.

2) Експеримент

Означення 2.2. *Експеримент* – це активний метод збирання інформації, у якому дослідник цілеспрямовано створює чи змінює умови вивчення впливу цих змін на об'єкт чи явище. Мета експерименту – встановити причинно-наслідкові зв'язки.

Наприклад, клінічні випробування нових ліків (контрольна група отримує плацебо, експериментальна – препарат); зміна однієї змінної у виробничому процесі для визначення її впливу на якість продукції; маркетинговий експеримент, коли у різних регіонах тестується різна реклама продукту.

Переваги: Дозволяє виявити причинно-наслідкові зв'язки, високу контрольованість умов.

Недоліки: можуть бути етичні обмеження, штучні умови можуть спотворювати результати, складність у відтворенні реальних умов.

3) Опитування

Означення 2.3. *Опитування* – це метод збору первинної інформації шляхом з'ясування думок, знань, уявлень та установок людей з питань, що цікавлять.

Опитування можуть бути:

• Усні (інтерв'ю). Наприклад, особисте інтерв'ю (розмова з респондентом віч-на-віч; телефонне інтерв'ю (спілкування по телефону).

Переваги: Гнучкість, можливість уточнення питань, висока глибина інформації.
Недоліки: Дорожнеча, трудомісткість, можливий вплив інтерв'юера.

• Письмові (анкетування). Наприклад, паперові анкети (лунають або розсилаються поштою); онлайн-опитування (через веб-форми, електронну пошту, соціальні мережі).

Переваги: Охоплення великої аудиторії, анонімність (може підвищити чесність відповідей), економічність.
Недоліки: низький відсоток повернення, неможливість уточнення, ризик неправильного розуміння питань.

• Типи питань у опитуваннях. Наприклад, закриті (з запропонованими варіантами відповідей (так/ні, вибір зі списку); відкриті (вимагають розгорнутої відповіді).

4) Аналіз документів (контент-аналіз)

Означення 2.4. *Аналіз документів* – це метод збирання даних шляхом вивчення існуючих письмових, друкованих чи електронних джерел інформації.

Наприклад, аналіз фінансових звітів компаній (вивчення протоколів засідань, стенограм, законодавчих актів); контент-аналіз новинних статей, постів у соціальних мережах для виявлення певних тенденцій чи тональностей; дослідження медичних карток пацієнтів.

Переваги: Доступність даних, часто дешевша за інші методи, можливість вивчення історичних даних.
Недоліки: Документи можуть бути неповними, застарілими, упередженими або не відповідати цілям дослідження.

5) Панельні дослідження

Означення 2.5. *Панельні дослідження* – це багаторазовий збір інформації в однієї й тієї групи об'єктів (панелі) через певні проміжки часу. Дозволяють відстежувати зміни та динаміку. Наприклад, опитування однієї і тієї ж групи домогосподарств про їх споживчі звички протягом року; моніторинг стану здоров'я тих самих пацієнтів протягом кількох років.

Переваги: Дозволяє аналізувати динаміку, виявляти тенденції та причинно-наслідкові зв'язки у часі.

Недоліки: "Зменшення" панелі (втрата учасників), ефект "старіння" панелі, висока вартість.

6) Реєстрація даних (автоматизований збір)

З розвитком технологій дедалі актуальнішим стає автоматизований збір даних. Наприклад, датчики у виробничому устаткуванні, що збирають дані про роботу машин; веб-аналітика (Google Analytics), що збирає дані щодо поведінки користувачів на сайті; дані з фітнес-трекерів, розумного годинника; транзакційні дані з банківських систем або систем продажів.

Переваги: Висока точність, великий обсяг даних, безперервність збирання, мінімальне людське втручання.

Недоліки: Необхідність у спеціальному устаткуванні / ПЗ, питання конфіденційності даних, ризик технічних збоїв.

Вибір методу збору інформації залежить від:

• Цілей дослідження: Що саме ви хочете дізнатися?

• Характер об'єкта, що вивчається: Які дані доступні?

• Наявні ресурси: Час, бюджет, персонал.

• Необхідна точність і глибина даних.

Часто в одному дослідженні використовуються комбінації кількох методів для отримання повнішої і достовірнішої картини.

3 Статистичне оброблення результатів спостережень

Нехай для об'єктів генеральної сукупності визначено певну ознаку чи числову характеристику, яку можна заміряти. Ця характеристика –

випадкова величина X , що приймає на кожному об'єкті певне числове значення. З вибірки обсягу n отримуємо значення цієї випадкової величини у вигляді ряду з n чисел:

$$x_1, x_2, \ldots, x_n. \tag{3.1}$$

Значення xi вибірки називаються ***варіантами***.

Означення 3.1. *Статистичний ряд розподілу* (або *варіаційний ряд*) – це спосіб представлення даних, який показує, як одиниці сукупності, що вивчається, розподіляються за значеннями певної ознаки.

Він складається з двох основних елементів:

1. Варіанти (значення ознаки): це окремі значення ознаки, що варіюється (змінюється). Наприклад, якщо ми вивчаємо зріст студентів, варіантами будуть конкретні значення зросту (170 см, 175 см, 180 см тощо).

2. Частоти (чи чисельності): це кількість разів, скільки зустрічається кожне конкретне значення ознаки у цій сукупності. Якщо 5 студентів мають зріст 170 см, то частота для варіанта "170 см" дорівнюватиме 5.

Також можуть використовуватися відносні частоти (або, зокрема, окремі випадки або елементи в рамках загальної статистичної сукупності, які розглядаються для більш детального аналізу, а також "частини" можуть вказувати на конкретні аспекти або ознаки, що характеризують ці елементи) - це частка кожної частоти від загального обсягу сукупності, виражена в частках одиниці.

Статистичні ряди розподілу можуть бути:

• Дискретними: коли ознака набуває окремих ізольованих значень (наприклад, кількість дітей у сім'ї).

• Інтервальні: коли ознака є неперервною і її значення групуються в інтервали (наприклад, вік людей, згрупований за інтервалами: 18-25 років, 26-35 років тощо). У цьому випадку вказуються інтервали та відповідні частоти.

Означення 3.2. *Статистичний розподіл вибірки* – це, по суті, те саме, як і статистичний ряд розподілу, але стосовно вибірки (частини генеральної сукупності), а не до всієї генеральної сукупності.

Коли ми проводимо дослідження, часто не можемо вивчити всіх представників генеральної сукупності (наприклад, усіх жителів країни). Натомість ми беремо вибірку – підмножина генеральної сукупності.

Статистичний розподіл вибірки дає нам початкове уявлення про закономірності, властивості генеральної сукупності, виходячи з даних, отриманих із цієї вибірки.

Він також являє собою перелік варіантів (значень ознаки, що спостерігаються) і відповідних їм частот або відносних частот, зазвичай у вигляді таблиці. Мета його побудови – узагальнити та систематизувати дані, отримані в результаті спостереження, щоб виявити основні тенденції та особливості розподілу ознаки у досліджуваній вибірці.

Узагальнення:

• Статистичний ряд розподілу – це загальний термін для впорядкованого представлення даних про розподіл ознаки, чи то в генеральній сукупності, чи у вибірці.

• Статистичне розподілення вибірки – це конкретний вид статистичного ряду розподілу, побудований на основі даних, зібраних з вибірки з генеральної сукупності.

Обидва поняття є фундаментальними для подальшого статистичного аналізу, оскільки дозволяють візуалізувати та аналізувати структуру даних.

Коли ми отримуємо дані вибірки (наприклад, результати опитування, вимірювання зросту людей, оцінки студентів), вони часто є несгрупованим набором чисел. Серед чисел ряду (3.1) можуть бути однакові числа. Щоб ці дані стали зрозумілими, наочними та придатними для подальшого аналізу, їх необхідно згрупувати.

Нехай випадкова величина X набуває значень $x_1, x_2, \ldots, x_k$, причому значення x_1 спостерігалося n_1 раз, значення x_2 – n_2 раз, … значення x_k – n_k раз ($n_1+ n_2+\ldots+n_k=n$ – обсяг вибірки).

Угруповання вибіркових даних – це процес упорядкування та систематизації вихідних даних з метою виявлення закономірностей, структури та розподілу ознаки, що вивчається.

Воно дозволяє:

• Зменшити обсяг даних для зручності сприйняття.

• Наочно уявити розподіл ознаки.
• Виявити основні тенденції та особливості у даних.
• Розрахувати різні статистичні показники (середні значення, дисперсію тощо).

Варіаційні ряди дають можливість визначити характер розподілу варіант сукупності за тією чи іншою кількісною ознакою. Варіаційні ряди розподілу залежно від групувальної ознаки поділяють на *ранжовані*, *дискретні* та *інтервальні*.

Означення 3.3. *Ранжований варіаційний ряд* – це ряд впорядкованих варіант сукупності в порядку зростання чи спадання ознаки, що досліджується.

3.1 Дискретний варіаційний ряд

Дискретний варіаційний ряд використовується, коли ознака, що вивчається, приймає кінцеву або лічильну кількість значень, які зазвичай є цілими числами або чітко фіксованими категоріями (наприклад, кількість дітей в сім'ї, бали на іспиті, число дефектів в партії товару).

Означення 3.4. *Дискретним варіаційним рядом* (*варіацією*) називають упорядковану за зростанням значень x_i послідовність варіант із вказівкою їх відповідних частот n_i або відносних частот w_i і подану у вигляді таблиці.

Означення 3.5. *Частотою* (Frequency) або *абсолютною частотою* n_i називають число, що показує скільки разів певне значення чи категорія зустрічається у наборі даних.

Іншими словами, це просто підрахунок того, скільки разів сталася та чи інша подія. Наприклад. Якщо при опитуванні 20 осіб про їхній улюблений колір 5 з них назвали жовтий, то частота для жовтого кольору дорівнює 5.

Означення 3.6. *Відносна частота* (Relative Frequency) або *частость* w_i – це частка чи відсоток, який певне значення чи категорія займає від загальної кількості спостережень.

Відносна частота (частость) показує, як часто подія відбувається щодо всього набору даних. Вона розраховується як відношення абсолютної частоти n_i до загального числа n спостережень: $w_i = \frac{n_i}{n}$.

Наприклад. У попередньому прикладі, якщо 5 із 20 осіб назвали жовтий, то відносна частота жовтого кольору буде 5/20=0,25 або 25%.

При вивченні варіаційних рядів поряд із поняттям частоти використовується поняття накопиченої частоти (позначаємо n_i').

Означення 3.7. *Накопичена частота* (Cumulative Frequency) або *кумулятивна частота* n_i' – це сума частот для даного значення та всіх попередніх значень (або категорій) у впорядкованому наборі даних, тобто для $X \le x_i$.

Вона показує загальну кількість спостережень, які менші або дорівнюють певному значенню. *Відмінність від частоти*: частота показує скільки разів зустрічається конкретне значення, в накопичена частота показує скільки разів трапляються всі значення до певного включно.

Наприклад. Маємо дані на 12 людей віком 10, 20, 30 років та їх частоти: 3, 5, 4 відповідно. Тоді матимемо. Вік 10: частота 3, накопичена частота 3. Вік 20: частота 5, накопичена частота 3+5=8 (це означає, що 8 осіб мають вік 20 років або менше). Вік 30 частота 4, накопичена частота 8+4=12.

Означення 3.8. *Накопичена відносна частота* (Cumulative Relative Frequency) або *накопичена* частость w_i' – це сума відносних частот для даного значення та всіх попередніх значень (або категорій) у впорядкованому наборі даних, тобто для $X \le x_i$.

Вона розраховується як $w_i' = \frac{n_i'}{n}$ або в процентах $w_i' = \frac{n_i'}{n} \cdot 100\%$ і показує частку чи відсоток спостережень, які менші чи рівні певному значенню. Останнє значення в стовпці накопиченої відносної частоти завжди дорівнюватиме 1 (або 100%).

Наприклад (продовжуючи попередній). Вік 10: відносна частота 3/12=0,25, накопичена відносна частота 0,25. Вік 20: відносна частота 5/12≈0,4167, накопичена відносна частота 0,25+0,4167=0,6667 (це

означає, що 66,67% осіб мають вік 20 років або менше). Вік 30 відносна частота 4/12≈0,3333, накопичена відносна частота 0,6667+0,3333=1,0.

Накопичена частота та накопичена частость в описовій статистиці, дозволяють зрозуміти кумулятивний розподіл даних. Вони показують, скільки чи яка частка спостережень перебуває нижче чи дорівнює певному значенню.

Де застосовуються ці поняття:

1) Накопичена частота та накопичена частость особливо корисні для: а) побудови емпіричної функції розподілу (дозволяють зрозуміти, як розподілені дані і яка ймовірність того, що випадкова величина набуде значення, що не перевищує певний поріг); б) визначення медіани, квартилів та інших процентилів (за їх допомогою легко знайти значення, нижче яких знаходиться певний відсоток даних, наприклад, 50% накопиченої частоти вказують на медіану); в) порівняння розподілів (дозволяють порівнювати, яка частка даних знаходиться нижче за певний рівень у різних наборах даних); г) контроль якості та статистичного управління процесами (використовуються для побудови контрольних "значень" та оцінки стабільності процесів).

2) Частота та відносна частота використовуються для: а) опису даних (дають базове уявлення про те, які значення зустрічаються найчастіше); б) побудови гістограм та стовпчастих діаграм (візуалізація розподілу даних); в) розрахунку основних статистичних показників (середнє значення, дисперсії тощо); г) оцінки ймовірності (відносна частота є емпіричною оцінкою ймовірності події: чим більше спостережень, тим ближча відносна частота до істинної ймовірності).

Таким чином, якщо частота та відносна частота дають "миттєвий знімок" розподілу, то накопичені показники дають "кумулятивний" погляд, дозволяючи зрозуміти, скільки даних лежить до певного "значення".

Структура дискретного варіаційного ряду:

Таблиця, якою визначається дискретний варіаційний ряд, складається з двох або трьох рядків:

1) Варіанта (x_i): Значення ознаки, розташовані в порядку зростання.

2) Частота (n_i): Сума всіх частот повинна дорівнювати обсягу вибірки: $\sum_{i=1}^{k} x_i = n$ (табл. 3.1).

3) Відносна частота (w_i) (опціонально, тобто не є обов'язковим, а надається на вибір або за бажанням): Сума відносних частот повинна дорівнювати 1: $\sum_{i=1}^{k} w_i = 1$ (табл. 3.2).

Таблиця 3.1

Варіанта x_i	x_1	x_2	…	x_k	
Частота варіанти n_i	n_1	n_2	…	n_k	$\sum_{i=1}^{k} x_i = n$

Таблиця 3.2

Варіанта x_i	x_1	x_2	…	x_k	
Відносна частота w_i	w_1	w_2	…	w_k	$\sum_{i=1}^{k} w_i = 1$

Приклад 1. Опитали 20 студентів про кількість п'ятірок, отриманих за семестр, та отримали такі дані: 3, 2, 4, 3, 5, 2, 3, 4, 3, 3, 5, 2, 4, 3, 3, 2, 4, 5, 3, 4. Побудувати дискретний варіаційний ряд.

Розв'язання. Впорядкуємо дані (необов'язково, але корисно): 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5. Підрахуємо частоту кожної варіанти: «дві п'ятірки»: 4 студенти, «три п'ятірки»: 8 студентів, «чотири п'ятірки»: 5 студентів, «п'ять п'ятірок»: 3 студенти.

Оскільки дискретний варіаційний ряд визначається у вигляді таблиці, то складемо її (табл. 3.3):

Таблиця 3.3

Кількість п'ятірок x_i	2	3	4	5	Разом
Частота n_i	4	8	5	3	20
Відносна частота $w_i = \frac{n_i}{n}$	0,2	0,4	0.25	0,15	1

3.2 Інтервальний варіаційний ряд

Інтервальний варіаційний ряд використовується, коли ознака, що вивчається, є неперервною (наприклад, зріст, вага, дохід, час, температура) і може прийняти будь-яке значення з деякого проміжку або, коли кількість різних значень дискретної ознаки занадто велике, що робить дискретний ряд громіздким і неінформативним.

У цьому випадку значення варіант групують за проміжками (інтервалами) (зазвичай однакової довжини), у першому рядку вказується проміжки, а у другому – число спостережень, що потрапили в даний проміжок. Може бути і третій рядок – частка інтервалу в загальному обсязі вибірки.

Структура інтервального варіаційного ряду:

Являє собою таблицю, що містить рядки:

1) Інтервал варіювання – це діапазон значень варіанти $[x_{\min};x_{\max}]$. Інтервали $[\alpha_1;\alpha_2)$, $[\alpha_2;\alpha_3)$,…, $[\alpha_{k-1};\alpha_k)$ мають бути суміжними та повністю охоплювати весь розмах варіювання. Це означає, що початок першого проміжку α_1 і кінець останнього α_k не обов'язково збігаються відповідно з $x_{\min}$ і $x_{\max}$. Варіанта, яка збігається з верхньою межею інтервалу, відноситься до наступного інтервалу. Якщо $\alpha_k=x_{\max}$, то останній інтервал є відрізком.

2) Частота (n_i) – це кількість значень із вибірки, що потрапили в інтервал. Сума всіх частот повинна дорівнювати обсягу вибірки.

3) Відносна частота (w_i) (опціонально) – це частка інтервалу в загальному обсязі вибірки.

Наприклад, вибірка розбита на k інтервалів, які не перетинаються і i-й інтервал містить варіанти вибірки у кількості n_i з відносною частотою $w_i = \frac{n_i}{n}$. Інтервальний варіаційний ряд матиме вигляд (табл. 3.4).

Таблиця 3.4

Інтервали	$\alpha_1 \le x < \alpha_2$	$\alpha_2 \le x < \alpha_3$	…	$\alpha_{k-1} \le x < \alpha_k$
Частота n_i	n_1	n_2	…	n_k
Відносна частота w_i	w_1	w_2	…	w_k

За величиною розрізняють проміжки (інтервали) рівні і нерівні. Рівні інтервали застосовують тоді, коли зміни кількісної ознаки всередині сукупності відбуваються рівномірно.

Означення 3.9. *Розмахом вибірки* називають відстань між найменшим та найбільшим значенням варіант цієї вибірки:

$$R = x_{\max} - x_{\min}. \tag{3.2}$$

При побудові інтервального ряду (угруповання даних) використовують різні підходи для визначення довжини інтервалів.

1. *Метод без "коригування"* – найпростіший підхід, коли інтервали вибираються довільно або інтуїтивно, без використання строгих формул.

Застосовується:

- коли дані вже мають чітко визначені категорії або дискретні значення;
- при невеликому обсязі даних, коли складний розрахунок інтервалів недоцільний;
- у випадку, коли необхідно швидко отримати перше уявлення про розподіл даних, не заглиблюючись у деталі;
- коли мета аналізу не вимагає високої точності у визначенні інтервалів або коли дані природно групуються.

Визначення довжини інтервалу без "коригування" (довільний чи інтуїтивний підхід)

Як сказано вище, "без коригування" означає, що немає суворої математичної формули, яка автоматично визначає оптимальну кількість інтервалів або їх довжину.

У цьому випадку довжина інтервалу може бути обрана на основі:

• *Досвід і здоровий глузд*: Якщо є відомі діапазони значень або категорії, які мають сенс для дослідження. Наприклад, під час аналізу віку можна вибрати інтервали 0–10, 10–20 років тощо.

• *Зручності подання*: Якщо потрібна певна кількість інтервалів для наочності (наприклад, 5–10 інтервалів для гістограми).

• *Вимоги до точності*: Іноді ширші інтервали можуть бути прийнятними для попереднього аналізу, а для більш детального – більш вузькі.

Зауваження 3.1. Формула для знаходження довжини інтервалу (загальна, після визначення числа інтервалів):

$$h = \frac{R}{k}, \tag{3.3}$$

де R – розмах вибірки, k– число інтервалів.

Приклад 2. Маємо дані про доходи 10 осіб на місяць (у тис. грн.): {8,12,15,18,20,22,25,30,35,40}. Побудувати інтервальний варіаційний ряд без "коригування".

Розв'язання. Знаходимо мінімальне та максимальне значення: мінімальне значення x_{min}=8, максимальне значення x_{max}=40.

Обчислюємо розмах R: $R=x_{max}-x_{min}=40-8=32$. Визначаємо бажану кількість інтервалів k. Наприклад, k=4 інтервали (довільний вибір).

Обчислюємо довжину інтервалу h:

$$h = \frac{R}{k} = \frac{32}{4} = 8.$$

Відповідь: інтервальний варіаційний ряд без "коригування" має вигляд [8;16), [16;24), [24;32), [32;40].

Зауваження 3.2. При формуванні інтервалів потрібно бути уважним до того, як включаються межі. Зазвичай ліва межа включається, а права ні (крім останнього інтервалу, який включає праву межу) для неперервних даних.

2. *Метод із "коригуванням"* (з використанням формул): Під "коригуванням" тут розуміється використання статистичних формул для визначення оптимальної довжини та кількості інтервалів, які найкраще відображають структуру даних та їх розподіл.

Найбільш відомі формули: Стерджесса, Скотта, Фрідмана-Дияконіса та інші. Самі ці формули вже по суті є "скоригованими" у тому сенсі, що вони враховують характеристики даних (розмах, обсяг

вибірки, стандартне відхилення) для визначення оптимальної кількості інтервалів та їх довжини.

1) *Формула Стерджесса*. Застосовується для даних, близьких до нормального розподілу, особливо при невеликих та середніх обсягах вибірки $(n < 200)$.

Кількість рівних інтервалів визначається за формулою Стерджесса:

$$k=1+3{,}322\cdot\lg(n), \tag{3.4}$$

де k – число інтервалів (береться найближче ціле до $1+3{,}322\cdot\lg(n)$), n – обсяг вибірки.

Величина (довжина) інтервалу обчислюється за формулою:

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}\ . \tag{3.5}$$

Зауваження 3.3. Якщо в результаті обчислення за формулою (3.5) довжина інтервалу h вийде дробовим числом, то його округляють до зручного для читання числа (наприклад, до цілого або до певної кількості знаків після коми). Також, початкова точка першого інтервалу може бути трохи меншою від мінімального значення, щоб усі дані точно потрапили в інтервали.

Приклад 3. Маємо дані про доходи 10 осіб на місяць (у тис. грн.): {8,12,15,18,20,22,25,30,35,40}. Побудувати інтервальний варіаційний ряд із застосуванням формули Стерджесса.

Розв'язання. n=10, R=32. Визначаємо кількість інтервалів k за формулою (3.4): $k=1+3{,}322\cdot\lg(n)=1+3{,}322.1=4{,}322$.

Округлюємо k. Зазвичай округляють до найближчого цілого числа. У даному випадку можна округлити до 4 або 5. Частіше округляють у більшу сторону, щоб уникнути занадто широких інтервалів, але іноді округляють і в меншу, щоб зберегти більш загальне уявлення. Візьмемо k=5.

Обчислюємо довжину інтервалу h за формулою (3.5): $h=R/k=32/5=6{,}4$.

Відповідь: інтервальний варіаційний ряд із застосуванням формули Стерджесса має вигляд [8;14,4), [14,4;20,8), [20,8;27,2), [27,2;33,6), [33,6;40].

2) *Формула Скотта*. Застосовується для даних, близьких до нормального розподілу, але чутливіша до викидів, ніж формула Фрідмана-Діаконіса.

Величина (довжина) інтервалу обчислюється за формулою:

$$h = 3{,}5 \cdot \frac{\sigma}{\sqrt[3]{n}} . \tag{3.6}$$

де n – обсяг вибірки, σ – стандартне відхилення вибірки $\left(\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$, $\bar{x}$ – середнє значення $\left(\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right)$.

Кількість інтервалів визначається за формулою:

$$k = \frac{R}{h}. \tag{3.7}$$

Приклад 4. Маємо дані про доходи 10 осіб на місяць (у тис. грн.): {8,12,15,18,20,22,25,30,35,40}. Побудувати інтервальний варіаційний ряд із застосуванням формули Скотта.

Розв'язання. n=10, R=32. Обчислюємо стандартне відхилення σ для даної вибірки. Знаходимо середнє значення

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i = \frac{1}{10} \left(8 + 12 + 15 + 18 + 20 + 22 + 25 + 30 + 35 + 40\right) = 22{,}5 .$$

Дисперсія $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{9} \Big((8 - 22{,}5)^2 + (12 - 22{,}5)^2 +$

$$+ (15 - 22{,}5)^2 + (18 - 22{,}5)^2 + (20 - 22{,}5)^2 + (22 - 22{,}5)^2 + (25 - 22{,}5)^2 +$$

$$+ (30 - 22{,}5)^2 + (35 - 22{,}5)^2 + (40 - 22{,}5)^2 \Big) = \frac{928{,}5}{9} \approx 103{,}1667 .$$

Стандартне відхилення

$$\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{103{,}1667} \approx 10{,}16 .$$

Обчислюємо довжину інтервалу h за формулою (3.6):

$$h = 3{,}5 \cdot \frac{\sigma}{\sqrt[3]{n}} = 3{,}5 \cdot \frac{10{,}16}{\sqrt[3]{10}} = 3{,}5 \cdot \frac{10{,}16}{2{,}1544} \approx 3{,}5 \cdot 4{,}7159 \approx 16{,}5 .$$

Обчислюємо кількість інтервалів k за формулою (3.7):

$$k = \frac{R}{h} = \frac{32}{16{,}5} \approx 1{,}94 .$$

Округляємо k: k=2.

Відповідь: інтервальний варіаційний ряд із застосуванням формули Скотта має вигляд [8;24,5), [24,5;41].

За формулою Скотта виходить невелика кількість широких інтервалів. Це часто відбувається з невеликими вибірками, де стандартне відхилення може бути досить великим.

3) *Формула Фрідмана-Діаконіса*. Найбільш робастний метод, тобто зберігає свою ефективність і достовірність результатів навіть за наявності "поганих" даних або при невеликих змінах у вихідних умовах, є стійким до викидів та асиметричних розподілів. Рекомендується, якщо є підозра на наявність викидів у даних.

Величина (довжина) інтервалу обчислюється за такою формулою:

$$h = 2 \cdot \frac{IQR}{\sqrt[3]{n}}, \qquad (3.8)$$

де n – обсяг вибірки, IQR – міжквартильний розмах: $IQR=Q_3-Q_1$, де Q_1, Q_3 – процентилі.

Зауваження 3.4. *Процентиль* або *перцентиль* – це показник, який вказує на місце певного значення у відсортованому за зростанням ряді даних. Він показує, який відсоток значень у наборі даних менший або дорівнює даному значенню.

Наприклад, Q_1 (25-й процентиль) – це значення, нижче якого знаходиться 25% чисел у списку; Q_3 (75-й процентиль) – це значення, нижче якого знаходиться 75% чисел у списку.

Зауваження 3.5. Набір даних можна розділити на 100 рівних частин за допомогою процентилів. Якщо ж ділимо дані на 4 рівні частини, то отримуємо *квартилі*.

Тобто, *квартилі* – це окремі випадки процентилів: *перший квартиль* (Q_1) – це 25-й процентиль, *другий квартиль* (Q_2) – це 50-й процентиль а *третій квартиль* (Q_3) – це 75-й процентиль. Це значення, нижче яких відповідно лежить 25%,50%,75% даних.

Щоб знайти процентиль для вибірки, потрібно спочатку впорядкувати дані за зростанням, а потім обчислити позицію, яка відповідає заданому процентилю. Формула для знаходження i-позиції k-го процентиля:

$$i = \frac{k}{100} \cdot (n+1),$$

де k – це номер процентиля (наприклад, 25 для 25-го процентиля), а n – загальна кількість значень у вибірці.

Якщо i – ціле число, то значення k-го процентиля – це значення даних у цій позиції.

Якщо i – не ціле число, то необхідно округлити його до найближчих цілих чисел (вгору – позиція (m+1) і вниз – позиція (m)) та застосувати метод інтерполяції щоб отримати наближене значення процентиля:

а) якщо дробова частина обчисленої позиції i процентилю дорівнює 0,5, то усереднюємо значення даних, що знаходяться на цих двох позиціях у вибірці: $(x_{m+1}+x_m)/2$;

б) якщо дробова частина обчисленої позиції процентилю відрізняється від 0,5, то застосовуємо формулу для лінійної інтерполяції:

$$Q_k = x_m + (x_{m+1} - x_m) \cdot f \ , \qquad (3.9)$$

де Q_k – значення процентиля, яке ми шукаємо, i – позиція, яку ми розрахували (може бути нецілим числом). [i] – ціла частина позиції i, f – дробова частина позиції i (тобто, i–[i]), x_m – значення даних позиції m (у відсортованому наборі даних), x_{m+1} – значення даних на позиції m+1 (у відсортованому наборі даних) при цьому $m<i<m+1$.

Процентилі корисні для аналізу розподілу даних та виявлення викидів. Їх часто використовують у статистиці, тестуванні та інших галузях, де потрібно порівнювати значення.

Кількість інтервалів визначається за формулою (3.7).

Приклад 5. Маємо дані про доходи 10 осіб на місяць (у тис. грн.): {8,12,15,18,20,22,25,31,35,40}. Побудувати інтервальний варіаційний ряд із застосуванням формули Фрідмана-Діаконіса.

Розв'язання. n=10, R=32. Знаходимо Q_1 та Q_3.

Для впорядкованого ряду {8,12,15,18,20,22,25,31,35,40} Q_1 (25-й процентиль, тобто k=25).

Визначаємо позицію за формулою $i = \frac{k}{100} \cdot (n+1) = \frac{25}{100} \cdot 11 = 2{,}75$.

Оскільки 2,75 не є цілим числом і дробова його частина відмінні від 0,5, то округлюємо його до m=2 та m+1=3. Значення у 2-й позиції – x_m=12, а у 3-й – x_{m+1}=15, f=2,75–2=0,75. Розраховуємо значення процентиля за формулою (3.9):

$$Q_1 = 12 + (15-12) \cdot 0{,}75 = 12 + 3 \cdot 0{,}75 = 12 + 2{,}25 = 14{,}25.$$

Аналогічно знаходимо Q_3 (75-й процентиль, тобто k=75). Визначаємо позицію за формулою $i = \frac{k}{100} \cdot (n+1) = \frac{75}{100} \cdot 11 = 8{,}25$.

Оскільки 8,25 не є цілим числом, то округлюємо його до m=8 та m+1=9. Значення у 8-й позиції – x_m=31, а у 9-й – x_{m+1}=35, f=8,25–8=0,25. Розраховуємо значення процентиля за формулою (3.9):

$$Q_3 = 31 + (35-31) \cdot 0{,}25 = 31 + 4 \cdot 0{,}25 = 31 + 1 = 32.$$

Знаходимо міжквартильний розмах за формулою:

$$IQR = Q_3 - Q_1 = 42 - 14{,}25 = 17{,}75.$$

Обчислюємо довжину інтервалу h за формулою (3.8):

$$h = 2 \cdot \frac{IQR}{\sqrt[3]{n}} = 2 \cdot \frac{17{,}75}{\sqrt[3]{10}} = 2 \cdot \frac{17{,}75}{2{,}1544} \approx 2 \cdot 8{,}24 \approx 16{,}48.$$

Обчислюємо кількість інтервалів k за формулою (3.7):

$$k = \frac{R}{h} = \frac{32}{16{,}48} \approx 1{,}94.$$

Округляємо k: k=2.

Відповідь: інтервальний варіаційний ряд із застосуванням формули Фрідмана-Діаконіса має вигляд [8;24,48), [24,48;40,96].

Як і у випадку з формулою Скотта, для невеликої вибірки формула Фрідмана-Діаконіса також може давати невелику кількість інтервалів.

Поряд із застосуванням формул Стерджесса, Скотта, Фрідмана-Дияконисса для визначення k і h при побудові варіаційного ряду можна застосовувати для меж інтервалів групування підхід *без коригування меж першого і останнього інтервалів* і *з коригуванням меж першого і останнього інтервалів*.

1) Побудова варіаційного ряду *без коригування меж першого й останнього інтервалів*, тобто перший інтервал $[x_{\min};x_{\min}+h)$, який починається з $x_{\min}$, в закінчується $x_{\min}+h$. Наступні інтервали виходять додаванням до кінця попереднього інтервалу довжини інтервалу h: $x_{i+1}=x_i+h$. Потім підраховуємо кількість варіант, що потрапили у кожний інтервал.

2) Побудова варіаційного ряду *з коригуванням меж першого та останнього інтервалів*, щоб x_1 і x_n потрапили всередину першого та останнього інтервалів групування. Для цього початок (нижню межу) першого інтервалу рекомендується брати як $x_1'=x_{\min}-h/2$, а кінець останнього інтервалу як $x_n'=x_{\max}+h/2$. Проміжні інтервали виходять додаванням до кінця попереднього інтервалу величини інтервалу h: $x_{i+1}=x_i+h$. Потім підраховуємо кількість варіантів, що потрапили у кожний інтервал.

Загалом, такий підхід припустимий і навіть рекомендується в деяких випадках, але не є універсальним та має свої нюанси.

Обґрунтування такого підходу:

Облік неперервності: Якщо ознака, яку ми групуємо, є неперервною (наприклад, зріст, вага, температура), її значення можуть бути будь-якими в межах певного діапазону. Використання $x_{\min}$ як точної нижньої межі першого інтервалу і $x_{\max}$ як точної верхньої межі останнього інтервалу може створити враження, що дані починаються і закінчуються саме в цих точках, що не відповідає реальності для неперервних даних. Додавання $h/2$ до верхньої та віднімання $h/2$ від нижньої межі допомагає "розтягнути" інтервали, щоб вони охоплювали весь діапазон можливих значень, а не тільки спостережуваних.

Візуалізація (гістограми): Коли будується гістограма, важливо, щоб стовпчики починалися та закінчувалися логічно. Якщо інтервали строго від x_{min} до x_{max}, крайні стовпчики можуть бути "урізаними" або несиметричними, особливо якщо x_{min} і x_{max} є викидами або крайніми значеннями, а більшість даних зосереджено ближче до центру. Невелике розширення меж може зробити гістограму більш збалансованою та естетичною.

Уникнення втрати даних: Якщо h обчислено таким чином, що R/k не є цілим або "зручним" числом, і ми округляємо h у меншу сторону, або якщо x_{max} точно збігається з верхньою межею, це може призвести до того, що x_{max} не потрапить в жодний інтервал (якщо застосовується правило "ліва межа" включно, права – ні). Розширення діапазону гарантує, що всі дані будуть включені.

Коли це особливо актуально: працюючи з неперервними даними, при побудові гістограм для візуального аналізу розподілу.

Коли може бути менш актуальним або не застосовуватися: При роботі з дискретними даними, де значення мають чіткі, розділені межі (наприклад, кількість дітей, кількість автомобілів). Тут "розширення" меж може бути безглуздим. Якщо мета аналізу вимагає строгого дотримання діапазону, що спостерігається без будь-яких припущень про значення за його межами.

Приклад 6. Знаючи для вибірки $x_{min}=10$, $x_{max}=50$ і розраховану довжину інтервалу $h=8$, отримано інтервальний варіаційний ряд без коригування меж першого й останнього інтервалів:

[10;18), [18;26), [26;34), [34;42), [42;50].

Записати інтервальний варіаційний ряд з коригуванням меж першого й останнього інтервалів.

Розв'язання. Застосуємо коригування меж першого та останнього інтервалів:

$$x_1' = x_{min} - h/2 = 10 - 4 = 6,\ x_n' = x_{max} + h/2 = 50 + 4 = 54.$$

Тоді шуканий інтервальний варіаційний ряд матиме вигляд:

[6;14), [14;22), [22;30), [30;38), [38;46), [46;54].

Кількість інтервалів збільшилася. Це сталося тому, що вихідний розмах R змінився на $R+h$.

Відповідь: інтервальний варіаційний ряд матиме вигляд [6;14), [14;22), [22;30), [30;38), [38;46), [46;54].

Розглянемо, як узгоджується побудова інтервального варіаційного ряду з коригуванням меж першого і останнього інтервалів із застосуванням формул Стерджесса, Скотта, Фрідмана-Дияконіса для визначення k і h.

Це два послідовні етапи, які добре узгоджуються:

1) Спочатку використовуємо формули (Стерджесса, Скотта, Фрідмана-Дияконіса) для визначення оптимальної кількості інтервалів k та їх довжини h.

Ці формули орієнтовані на внутрішню структуру даних (розмах, обсяг вибірки, стандартне відхилення, міжквартильний розмах) для вибору відповідного k і h.

Важливо. На цьому етапі h зазвичай розраховується як $h=R/k$, де $R=x_{max}-x_{min}$.

2) Після того, як k і h визначені, можна застосувати коригування меж першого та останнього інтервалів.

Це коригування не змінює h, але розширює загальний діапазон, який покривається інтервалами.

Новий розмах для інтервалів фактично стає

$$(x_{max}+h/2)-(x_{min}-h/2)=x_{max}-x_{min}+h=R+h.$$

Відповідно, якщо ми розширюємо діапазон на h, а довжина кожного інтервалу залишається h, то кількість інтервалів також збільшиться на 1. Якщо початкова кількість інтервалів була k, то нова ефективна кількість інтервалів для охоплення розширеного діапазону буде $k+1$.

Висновок. Підхід із коригуванням меж не суперечить формулам Стерджесса, Скотта, Фрідмана-Дияконіса. Навпаки, він доповнює їх, пропонуючи спосіб формування фінальних меж інтервалів після того, як оптимальні k та h були визначені. Формули дають "ідеальне" k та h на основі властивостей даних. Коригування меж (розширення на $h/2$ з кожного боку) – це практичний крок для забезпечення того, щоб гістограма або інтервальний ряд виглядали завершеними, охоплювали весь діапазон даних (і навіть трохи за його межами для візуалізації неперервності) і включали всі значення, що спостерігаються, особливо при роботі з неперервними даними.

Важливо пам'ятати. Якщо застосовується коригування меж $x_1'=x_{\min}-h/2$ та $x_n'=x_{\max}+h/2$, то фактично працюємо з новим розмахом $R'=R+h$. Відповідно, кількість інтервалів, які будуть сформовані в цьому новому розмаху за тієї ж довжини h, буде

$$k' = \frac{R'}{h} = \frac{R+h}{h} = \frac{R}{h} + 1 = k + 1.$$

Тобто фактично кількість інтервалів збільшиться на один. Це потрібно враховувати під час інтерпретації результату.

У деяких випадках замість строгого використання $x_{\min}-h/2$ і $x_{\max}+h/2$ просто вибирають зручні "красиві" числа, які трохи виходять за межі $x_{\min}$ і $x_{\max}$, щоб інтервали починалися і закінчувалися на круглих значеннях.

Таким чином, коригування меж є гнучким інструментом, який допомагає покращити подання інтервального ряду, особливо для неперервних даних та побудови гістограм, і застосовується після розрахунку k та h за допомогою статистичних формул.

Приклад 7. Для вибірки {10,15,20,25,30} побудувати інтервальний варіаційний ряд, застосовуючи формулу Стерджесса, а) без коригування меж першого та останнього інтервалів, б) з коригуванням меж першого та останнього інтервалів.

Розв'язання. Маємо $n=5$, $x_{\min}=10$, $x_{\max}=30$, $R=x_{\max}-x_{\min}=30-10=20$.

а) Визначаємо кількість інтервалів k за формулою Стерджесса: $k=1+3{,}322\cdot\lg(5)=1+3{,}322\cdot 0{,}69897=3{,}322$. Округлюємо k: $k=4$.

Обчислюємо довжину інтервалу h за формулою (3.5): $h=R/k=20/4=5$.

Інтервали без коригування меж (стандартний підхід):

[10;15), [15;20), [20;25), [25;30]

(або [25;30) та останній інтервал [30;35), якщо $x_{\max}$ потрапляє на межу, щоб $x_{\max}$ був включений).

б) Застосуємо коригування меж першого та останнього інтервалів. Знаходимо

$$x_1=x_{\min}-h/2=10-5/2=7{,}5,\ x_n=x_{\max}+h/2=30+5/2=32{,}5.$$

Маємо новий "розтягнутий" розмах: 32,5-7,5=25. При довжині інтервалу $h=5$ кількість інтервалів, необхідне для покриття цього нового розмаху, буде 25/5=5. Тобто k збільшилося з 4 до 5.

Інтервали з коригуванням меж:

[7,5;12,5), [12,5;17,5), [17,5;22,5), [22,5;27,5), [27,5;32,5].

Відповідь: а) інтервали без коригування меж [10;15), [15;20), [20;25), [25;30]; б) інтервали з коригування меж [7,5;12,5), [12,5;17,5), [17,5;22,5), [22,5;27,5), [27,5;32,5].

Якщо в інтервальнному варіаційному ряді в кожному інтервалі узяти середнє значення інтервалу: $x_i' = \frac{\alpha_{i-1} + \alpha_i}{2}$, $\left(i = \overline{1, k-1}\right)$, то *інтервальний варіаційний* ряд можна умовно подати *дискретним варіаційним* рядом.

Це зручно робити у тих випадках, коли:

1) Інформації про окремі значення варіанти достатньо для аналізу, і втрата точності при переході до дискретних значень не є критичною.

Наприклад. Вивчаємо розподіл зросту студентів у групі. Спочатку зібрані дані (у см) записали інтервальним рядом: [160;165), [165;170), [170;175) (табл. 3.5).

Однак, якщо для подальшого аналізу нам достатньо знати середній зріст в кожному інтервалі або частоту конкретних значень (наприклад, 163 см, 168 см, 173 см), то можемо перейти до дискретного ряду, використовуючи, наприклад, середини інтервалів (табл. 3.6).

Таблиця 3.5 Вихідний інтервальний ряд

Зріст (см)	[160;165)	[165;170)	[170;175)
Частота	10	15	12

Таблиця 3.6 Перетворений дискретний ряд (з використанням середин інтервалів)

Зріст (см)	162,5	167,5	172,5
Частота	10	15	12

Цей підхід зручний, коли для подальших розрахунків (наприклад, середнього значення, дисперсії) ми можемо оперувати середніми значеннями інтервалів.

2) Вихідні інтервали є досить вузькими, і значення варіанти всередині кожного інтервалу вважатимуться приблизно однаковими.

Наприклад. Вимірюємо температуру повітря щогодини та групуємо дані за інтервалами в 0,5°С: [20;20,5), [20,5;21) тощо (табл. 3.7). Якщо необхідно побудувати графік зміни температури щогодини, то для зручності візуалізації та аналізу можемо прийняти середину кожного інтервалу як представницьке значення. Різниця між серединою інтервалу та фактичними значеннями всередині нього буде незначною, і дискретний ряд добре відображатиме тенденції (табл. 3.8).

Таблиця 3.7 Вихідний інтервальний ряд

Температура (°С)	[20;20,5)	[20,5;21)	[21;21,5)
Частота	5	8	10

Таблиця 3.8 Перетворений дискретний ряд

Температура (°С)	20,25	20,75	21,25
Частота	5	8	10

3) Мета аналізу полягає у спрощенні представлення даних для побудови певних графіків (наприклад, полігону частот) або для проведення розрахунків, які потребують дискретних значень.

Наприклад. Для побудови полігону частот (ламаної лінії, що з'єднує точки, координати яких – середини інтервалів та частоти) потрібні дискретні значення. Інтервальний ряд для цього безпосередньо не підходить. У цьому випадку, необхідно перетворити інтервальний ряд на дискретний, використовуючи середини інтервалів, що дозволяє наочно уявити розподіл даних.

4) Вихідні інтервали мають фіксовану довжину і є природна середина або представницька точка для кожного інтервалу.

Наприклад. Якщо аналізувати вік людей, згрупований за 5-річними інтервалами: [0;5), [5;10), [10;15) і так далі, то для подальшого аналізу може бути зручно використовувати середини інтервалів: 2,5;7,5;12,5;… як дискретні значення. Це дозволяє працювати з даними як із дискретними, не втрачаючи при цьому загальну картину розподілу віку.

Зауваження 3.6. Перехід від інтервального до дискретного ряду завжди пов'язаний з деякою втратою інформації. Втрачаємо точність

щодо фактичних значень усередині кожного інтервалу. Тому таке перетворення виправдане лише тоді, коли ця втрата інформації не спотворює результати аналізу та не впливає на висновки. У випадках, коли потрібна максимальна точність або коли розкид значень усередині інтервалів великий, інтервальний ряд краще.

3.3 Побудова полігону та гістограми статистичного розподілу

Полігон і гістограма – це два основні графічні способи подання варіаційних рядів, які допомагають наочно зобразити розподіл частот або відносних частот ознаки.

Означення 3.10. *Полігон частот* – це графічне зображення статистичного розподілу у вигляді ламаної лінії. Її вершини розташовані в точках $(x_i; n_i)$, де x_i – варіанти вибірки, n_i – відповідні їм частоти.

Означення 3.11. *Полігон відносних частот* – аналогічна лінія, побудована на точках $(x_i; w_i)$, де x_i – варіанти вибірки, w_i – відповідні їм відносні частоти.

Полігон частот (або полігон відносних частот) застосовується як для дискретних, так неперервних даних. Графічне зображення:

• Для дискретних даних: Точки на графіку відповідають значенням ознаки по осі X та їх частотам (або відносним частотам) по осі Y. Ці точки потім з'єднуються відрізками ламаної лінії. Для дискретних даних полігон показує частоту кожного конкретного значення.

• Для неперервних даних: Точки на графіку відповідають серединам інтервалів угруповання по осі X і частотам (або щільності частот) цих інтервалів по осі Y. Потім ці точки також з'єднуються ламаною лінією. Для неперервних даних полігон апроксимує форму розподілу.

Полігон частот можна розглядати як згладжене уявлення гістограми, особливо для неперервних даних.

Приклад 8. Провели опитування 20 студентів про кількість пропущених ними занять з математики за останній місяць і отримали

такі дані: 0,1,2,1,0,3,2,0,1,1,2,0,1,0,3,1,0,2,2,1. Побудувати полігон частот і полігон відносних частот, склавши дискретний варіаційний ряд для випадкової величини X – кількість пропущених занять.

Розв'язання. Зробимо ранжування даних: упорядкуємо їх за зростанням та визначимо, скільки разів зустрічається кожне значення 0,1,2,3 серед отриманих, тобто зробимо підрахунок частот і відносних частот. Складемо таблицю 3.9:

Таблиця 3.9

Кількість пропусків занять x_i	Частота n_i	Відносна частота $w_i = \frac{n_i}{n}$
0	6	0,3
1	7	0,35
2	5	0,25
3	2	0,10
Разом	20	1,00

Будуємо полігон частот і полігон відносних частот: на горизонтальній осі відкладаються значення x_i, на вертикальній відповідні n_i або w_i. Точки з'єднуються відрізками: для полігона частот (0;6), (1;7), (2;5), (3;2); для полігона відносних частот (0;0,3), (1;0,35), (2;0,25), (3;0,1).

Графіки полігона частот і полігона відносних частот мають вигляд (рис. 3.1).

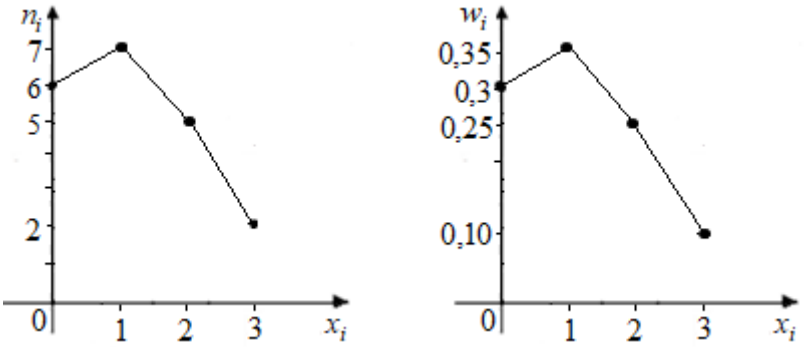


Рисунок 3.1

Гістограма служить тільки для зображення інтервальних варіаційних рядів

Означення 3.12. *Гістограма* – це стовпчаста діаграма, яка використовується для графічного представлення частотного розподілу неперервної (або дискретної з великим числом значень) змінної.

Призначення та інтерпретація:

• *Візуалізація форми розподілу*: дозволяє побачити, чи є розподіл симетричним, скошеним (асиметричним), унімодальним (з одним піком), бімодальним (з двома піками) тощо.

• *Виявлення викидів*: незвичайно розташовані стовпці можуть вказувати на викиди.

• *Оцінка щільності розподілу*: чим вище стовпець, тим більше даних у цьому інтервалі.

• Гістограма дає уявлення про емпіричну густину розподілу даних.

Вона показує, як часто значення потрапляють у певні інтервали. Площа гістограми пропорційна загальному числу спостережень.

Означення 3.13. *Гістограмою частот* називається сукупність прямокутників, основами яких служать часткові інтервали довжиною $\Delta x_i = x_{i+1} - x_i$ і висотами $h_i = \frac{n_i}{\Delta x_i}$.

Означення 3.14. *Гістограмою відносних частот* називається сукупність прямокутників, основами яких служать часткові інтервали довжиною $\Delta x_i = x_{i+1} - x_i$ і висотами $h_i = \frac{w_i}{\Delta x_i}$.

Площа всієї гістограми частот дорівнює n (обсягу вибірки), а площа всієї гістограми відносних частот дорівнює 1.

Графічне зображення гістограми частот (гістограми відносних частот): на горизонтальній осі X відкладаються інтервали угруповання, на вертикальній осі Y – частоти (або відносні частоти).

Прямокутники будуються так, що їх основи відповідають довжині інтервалу, а висота – частоті (або відносної частоти) попадання значень до цього інтервалу.

Гістограма і полігон можуть служити деяким наближенням графіка невідомої щільності розподілу $f(x)$ випадкової величини X. Точність наближення зростає зі зростанням обсягу вибірки та кількості часткових інтервалів.

Якщо з'єднати середини верхніх основ прямокутників відрізками прямої, можна отримати полігон того ж розподілу.

Приклад 9. Провели опитування серед 30 випадкових мешканців стосовно того, скільки коштів (у гривнях) на місяць вони витрачають на мобільний зв'язок. Отримали наступні дані: 311,285,235,275,190, 244,242,204,210,180,391,167,355,259,356,156,381,231,284,427,507,264, 453,246,402,222,346,311,235,361. Побудувати гістограму частот і гістограму відносних частот.

Розв'язання. Складемо інтервальний варіаційний ряд. Маємо n=30, x_{min}=156, x_{max}=507. Розмах вибірки: $R=x_{max}-x_{min}=507-156=351$. Визначаємо кількість інтервалів k за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg(30) = 1 + 3{,}322 \cdot 1{,}447 \approx 1 + 4{,}908 \approx 5{,}908 .$$

Округлюємо k: k=6. Обчислюємо довжину інтервалу h за формулою (3.5): $h = \frac{R}{k} = \frac{351}{6} \approx 58{,}5$. Визначаємо межі інтервалів: [156;214,5),[214,5;273),[273;331,5),[331,5;390),[390;448,5),[448,5;507]. Робимо підрахунок частот, відносних частот для кожного інтервалу. Складемо таблицю 3.10:

Таблиця 3.10

Інтервали часу (сек.)	Частота n_i	Відносна частота $w_i = \frac{n_i}{n}$
[156;214,5)	6	6/30≈0,2
[214,5;273)	7	7/30≈0,233
[273;331,5)	7	7/30≈0,233
[331,5;390)	5	5/30≈0,167
[390;448,5)	3	3/30≈0,1
[448,5;507]	2	2/30≈0,067
Разом	30	1,000

Графіки гістограми частот (рис.3.2) і гістограми відносних частот (рис. 3.3) мають вигляд:

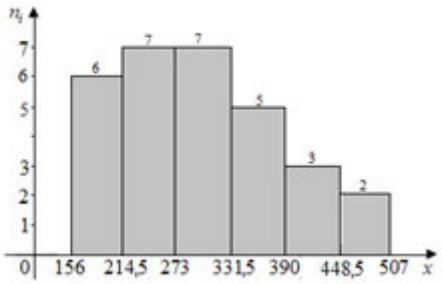


Рисунок 3.2

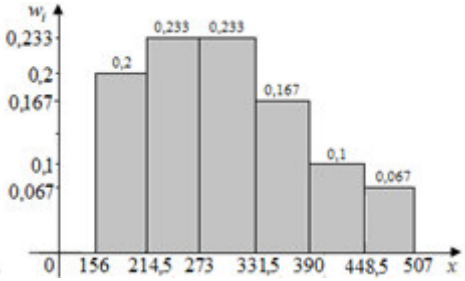


Рисунок 3.3

Означення 3.15. Щільність частоти або щільність ймовірності (емпірична) H_i – це відносна частота, яка припадає на одиницю довжини інтервалу. Обчислюється за формулою: $H_i = \frac{w_i}{h}$, $\left(i = \overline{1,k}\right)$.

Величина H_i відіграє ключову роль у створенні гістограми густини (або нормованої гістограми).

Зазначимо доцільність її використання:

1) Нормалізація площі гістограми до 1 (або 100%).

Основна причина використання $H_i = \frac{w_i}{h}$ полягає в тому, щоб площа кожного прямокутника гістограми дорівнювала відносної частоті відповідного інтервалу.

Якщо побудувати гістограму так, що висота прямокутника дорівнюватиме просто відносній частоті, а довжина – h, то площа буде $h \cdot w_i$. Це не завжди зручно, особливо при порівнянні розподілів із різною

довжиною інтервалів.

Якщо використовувати H_i як висоту, то площа прямокутника буде дорівнювати відносні частоті. Таким чином, сума площ усіх прямокутників гістограми дорівнюватиме 1 (якщо відносні частоти в сумі дають 1) або 100% (якщо використовуються відсотки). Це дозволяє інтерпретувати гістограму як графічне подання щільності ймовірності.

2) Порівнянність розподілів із різною довжиною інтервалів.

Маємо два набори даних, для яких треба побудувати гістограми. В одному випадку використані інтервали довжиною 5 одиниць, а в іншому – 10 одиниць.

Якщо використовувати відносні частоти як висоти, то у разі довших інтервалів (10 одиниць) стовпчики могли б здаватися "нижчими", навіть якщо щільність даних у них така ж або вище. Це позбавило можливості порівняння.

Використання H_i нівелює цю проблему. Висота H_i тепер відображає "концентрацію" даних в інтервалі незалежно від його довжини. Це дозволяє порівнювати форми розподілу, навіть якщо інтервали були обрані по-різному.

3) Відображення щільності розподілу.

Коли мова йде про щільність ймовірності для неперервних випадкових величин, то маємо на увазі функцію, яка описує відносну ймовірність того, що випадкова величина набуде значення в певному діапазоні. Гістограма щільності є емпіричним (заснованим на даних) наближенням цієї функції щільності.

Висота H_i в цьому випадку інтерпретується як емпірична щільність. Чим вище H_i, тим "щільнішими" є розташовані спостереження в даному інтервалі.

4) Підготовка до побудови кривої щільності.

Гістограма щільності часто є першим кроком до побудови кривої щільності. Крива щільності є більш гладким наближенням до справжнього розподілу, і вона будується на основі концепції щільності, яку відображає H_i.

Графічне зображення гістограми щільності: на горизонтальній осі X відкладаються інтервали угруповання, на вертикальній осі Y – відкладаємо щільність частоти H_i.

Будуємо прямокутники: довжина кожного прямокутника дорівнює довжині інтервалу h, а висота кожного прямокутника дорівнює щільності частоти H_i для відповідного інтервалу. Площа кожного прямокутника відповідає відносній частоті спостережень у цьому інтервалі. Загальна площа під гістограмою дорівнює 1, що дозволяє інтерпретувати її як емпіричний розподіл щільності ймовірності.

Для побудови гістограми щільності зручно використовувати таблицю 3.11:

Таблиця 3.11

Інтервали $(\alpha_i;\alpha_{i+1})$	Середини інтервалів $x_i' = \frac{\alpha_i + \alpha_{i+1}}{2}$	Інтервальні частоти n_i (кількість x_i в інтервалі)	Відносні частоти $w_i = \frac{n_i}{n}$	Щільність частоти $H_i = \frac{w_i}{h}$
$[\alpha_1;\alpha_2)$	x'_1	n_1	w_1	H_1
$[\alpha_2;\alpha_3)$	x'_2	n_2	w_2	H_2
...	…	…	…	…
$[\alpha_{k-1};\alpha_k]$	x'_{k-1}	n_{k-1}	w_{k-1}	H_{k-1}
k		$\sum_{i=1}^{k-1} n_i = n$	$\sum_{i=1}^{k-1} w_i = 1$	

Приклад 10. Припустимо, що є вибірка з 20 вимірювань часу реакції (у секундах) у досліджуваних: 12,5; 10,2; 11,8; 13,1; 10,5; 12,0; 11,5; 13,5; 10,8; 12,8; 11,2; 12,3; 11,0; 13,0; 11,7; 10,0; 12,1; 13,2; 11,4; 12,6. Необхідно побудувати гістограму щільності частоти.

Розв'язання. Визначимо діапазон даних та кількість інтервалів.

Маємо n=30, $x_{\min}$=10, $x_{\max}$=13,5.

Розмах вибірки: R=$x_{\max}$–$x_{\min}$=13,5–10=3,5.

Визначаємо кількість інтервалів k за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg(30) = 1 + 3{,}322 \cdot 1{,}301 \approx 1 + 4{,}322 \approx 5{,}322.$$

Округлюємо k: k=5. Обчислюємо довжину інтервалу h за формулою (3.5): $h = \frac{R}{k} = \frac{3{,}5}{5} = 0{,}7$.

Визначаємо межі інтервалів: [10;10,7), [10,7;11,4), [11,4;12,1), [12,1;12,8), [12,8;13,5]. Робимо підрахунок частот, відносних частот і щільності частоти для кожного інтервалу.

Для побудови гістограми щільності складаємо таблицю 3.12

Таблиця 3.12

Інтервал часу (сек.)	Інтервальні частоти n_i (кількість x_i в інтервалі)	Відносні частоти $w_i = \frac{n_i}{n}$	Щільність частоти $H_i = \frac{w_i}{h}$
[10;10,7)	3	3/20=0,15	0,15/0,7≈0,214
[10,7;11,4)	4	4/20=0,2	0,2/0,7≈0,286
[11,4;12,1)	5	5/20=0,25	0,25/0,7≈0,357
[12,1;12,8)	4	4/20=0,2	0,2/0,7≈0,286
[12,8;13,5]	4	4/20=0,2	0,2/0,7≈0,286
k=5	$\sum_{i=1}^{5} n_i = 20$	$\sum_{i=1}^{5} w_i = 1$	

Графік гістограми щільності має вигляд (рис. 3.4):

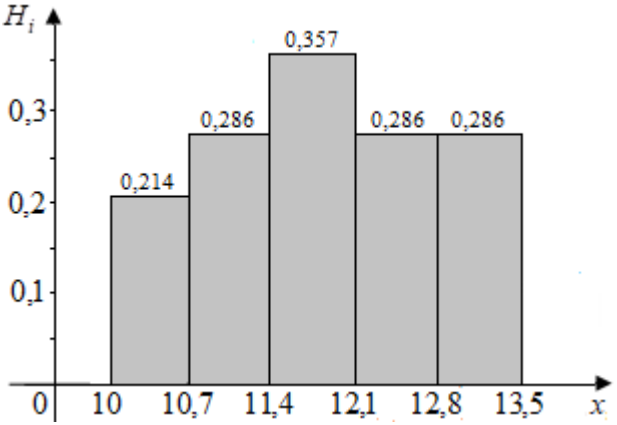


Рисунок 3.4

3.4 Побудова емпіричної функції розподілу та кумулятивної кривої

Функція розподілу (інтегральна функція розподілу) – це одна з ключових характеристик випадкової величини, яка описує ймовірність того, що випадкова величина набуде значення, яке менше або дорівнює заданому числу.

Означення 3.16. *Теоретична функція розподілу* $F(x)$ (*інтегральна функція розподілу* або *функція розподілу генеральної сукупності*) відображає накопичену ймовірність того, що випадкова величина X у результаті випробування виявиться меншою за фіксоване значення x: $F(x)=P(X<x)$.

Вона належить до генеральної сукупності (чи теоретичної моделі розподілу), а не до конкретної вибірки. Теоретична функція розподілу визначається для випадкової величини (дискретної чи неперервної) на основі її ймовірнісної моделі. Для дискретних вона ступінчаста, для неперервних – гладка. Її властивості та обчислення для дискретної і неперервної випадкової величина X розглянуті в теорії ймовірностей.

Означення 3.17. *Статистичною або емпіричною функцією розподілу* випадкової величини X за наявною вибіркою обсягу n називається функція $F^*(x)$, що дорівнює відносній частоті події $(X<x)$:

$$F^*(x)=\frac{n_x}{n}, \tag{3.10}$$

де n_x – число варіант у вибірці, менших x.

Ключова різниця між цими характеристиками полягає у формі оцінювання: якщо теоретична функція базується на апріорній імовірності появи значень, то емпірична (статистична) будується на фактично отриманій відносній частоті події $X<x$. Розглядається конкретна вибірка даних.

За теоремою Бернуллі відносна частота появи події A в n незалежних дослідах сходиться за ймовірністю до ймовірності $P(X<x)$ цієї події зі збільшенням n. Отже, за більших обсягів вибірки емпірична

функція розподілу $F^*(x)$ близька до теоретичної функції розподілу $F(x)$. Точніше, має місце теорема Глівенка–Кантеллі.

Теорема 3.1 (Глівенка–Кантеллі). Для будь-якого дійсного числа x та будь-якого $\varepsilon>0$

$$\lim_{n\to\infty} P\left(\left|F^*(x)-F(x)\right|>\varepsilon\right)=0.$$

Отже, по функції $F^*(x)$ можемо отримати приблизно функцію $F(x)$, тобто функція $F^*(x)$ є оцінкою $F(x)$.

Властивості емпіричної функції розподілу аналогічні властивостям теоретичної функції розподілу $F(x)$ дискретної випадкової величини X: $F^*(x)$ належить до конкретної вибірки даних; $F^*(x)$ – неспадна функція; $0\le F^*(x)\le 1$; якщо x_1 – найменша варіанта, то $F^*(x)=0$ при $x\le x_1$, якщо x_k – найбільша варіанта, то $F^*(x)=1$ при x> x_k.

Емпірична функція розподілу визначається за формулою:

$$F^*(x)=\begin{cases}0, & x\le x_1,\\ \dfrac{1}{n}\sum\limits_{j=1}^{i} n_j, & x_i<x\le x_{i+1},.\\ 1, & x>x_k,\end{cases} \qquad (3.11)$$

де x_i – варіанти варіаційного ряду.

Метод побудови емпіричної функції розподілу універсальний і однаковий як для дискретних, так неперервних даних.

У випадку дискретного варіаційного ряду емпірична функція розподілу $F^*(x)$ є ступінчастою функцією (її "стрибки" відбуваються в точках, що відповідають значенням варіанти, а величина стрибка дорівнює відносній частоті варіанти), оскільки дані дискретні за своєю природою (кінцева кількість спостережень) (рис. 3.5).

Для інтервального варіаційного ряду, як і у випадку дискретного варіаційного ряду, графік емпіричної функції розподілу $F^*(x)$ – це ступінчаста лінія із стрибками на межах інтервалів (потім її згладжують). Для цього емпірична функція розподілу $F^*(x)$ визначається тільки на кінцях інтервалу. Її зображують ламаною, що проходить через точки $(\alpha_i;F^*(\alpha_i))$, $i=\overline{1,k}$. При цьому $F^*(\alpha_1)=0$, $F^*(\alpha_k)=1$ (рис. 3.6).

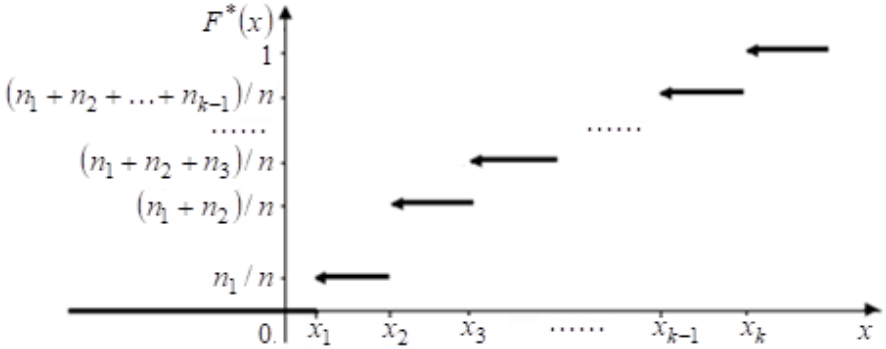


Рисунок 3.5

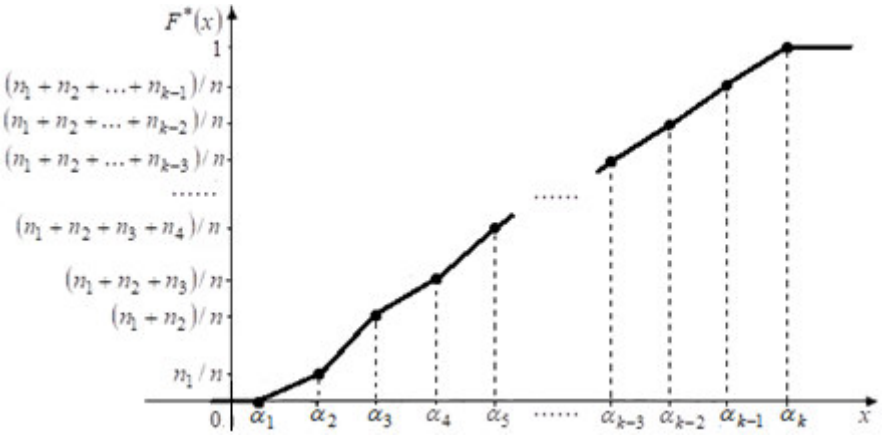


Рисунок 3.6

Емпіричним аналогом графіка інтегральної функції розподілу є *кумулятивна крива* (*кумулята*), яка відноситься до графічного представлення накопичених частот або частостей будь-якого параметра. Це один із способів візуалізації розподілу даних, який дозволяє зрозуміти, як значення ознаки накопичуються в міру їх збільшення.

Кумулятивна крива показує суму частот (або відсоткових частот) всіх значень, які менші або дорівнюють певному значенню x. Тобто, для кожного значення ознаки вона відображає частку спостережень, які потрапляють у цей діапазон або нижче за нього. Вона може будуватися як для дискретних, так неперервних даних.

Загалом, кумулятивна крива частіше використовується в описовій статистиці для графічного представлення емпіричних даних (вибірки) і є потужним інструментом для аналізу та візуалізації розподілу даних, дозволяючи швидко оцінити основні характеристики сукупності та зробити висновки про її структуру. Це ламана лінія, що візуально відображає накопичення частот (або відносних частот).

Кумулятивна крива використовується для таких цілей:

1) *Оцінка розподілу даних*. Вона дозволяє наочно побачити, як дані розподілені, і де зосереджена переважна більшість спостережень.

2) *Визначення медіани та квартилів*. За кумулятивною кривою легко знайти медіану (значення, яке поділяє дані на дві рівні половини, що відповідає 50% накопиченої частоти) та інші квартилі (25%, 75% і т.д.).

3) *Порівняння розподілів*. Можна порівнювати кумулятивні криві різних груп даних, щоб зрозуміти відмінності у розподілі.

4) *Візуалізація процентних часток*. Крива наочно показує, який відсоток даних знаходиться нижче за певний поріг.

5) *Аналіз процентилей*: Дозволяє визначити, якому значенню відповідає той чи інший процентиль. Наприклад, можна дізнатися, яке значення ознаки знаходиться у 90-му процентилі.

6) *Контроль якості*. У деяких областях, таких як контроль якості, кумулятивні криві використовуються для моніторингу процесів та виявлення відхилень.

Як будується кумулята для різних типів рядів:

1) Для дискретного ряду:

Кумулята будується шляхом з'єднання точок, де *X*-координата – це значення (варіанти) ознаки x_i, а *Y*-координата – накопичена частота (або накопичена частость) для цього значення x_i.

Наприклад, якщо є дискретний ряд (табл. 3.2), то кумулята буде з'єднувати точки: $(x_1;n_1')$, $(x_2;n_2')$, $(x_3;n_3')$ і т.д. (рис. 3.7) або $(x_1;w_1')$, $(x_2;w_2')$, $(x_3;w_3')$ і т.д. (рис. 3.8).

У випадку дискретного ряду, така кумулята (ламана лінія) також є візуальним поданням або апроксимацією, але строга емпірична функція розподілу для дискретного ряду все одно буде ступінчастою.

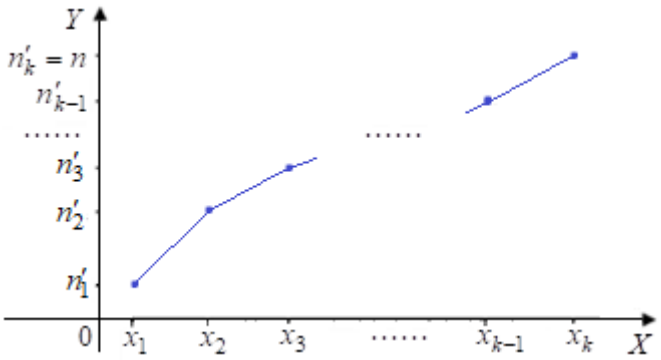


Рисунок 3.7

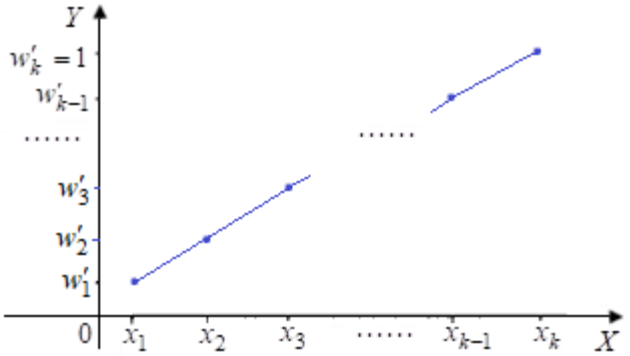


Рисунок 3.8

2) Для інтервального ряду:

При побудові кумуляти для інтервального ряду накопичені частоти (або частості) відносять до верхніх меж інтервалів.

Кумулята будується шляхом з'єднання точок, де *X*-координата – це верхня межа інтервалу, а *Y*-координата – накопичена частота (або накопичена частость) для цієї верхньої межі. Наприклад, якщо є інтервальний ряд (табл. 3.3), то кумулята буде з'єднувати точки: (α_2;n_1'), (α_3;n_2'), (α_4;n_3') і т.д. (рис. 3.9) або (α_2;w_1'), (α_3;w_2'), (α_4;w_3') і т.д. (рис. 3.10).

Часто додатково для початку кривої додають точку (α_1;0), де α_1 – "нижня межа першого інтервалу. У цьому випадку кумулята є апроксимацією істинної ступінчастої емпіричної функції розподілу для інтервального ряду.

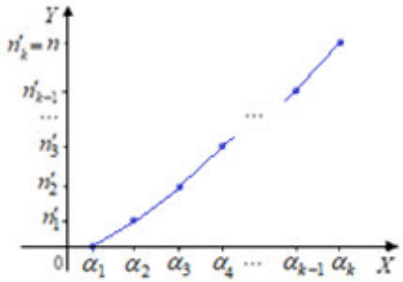


Рисунок 3.9

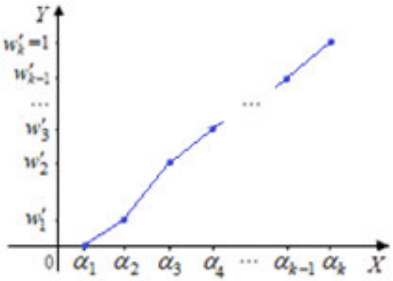


Рисунок 3.10

Приклад 11. Для аналізу надійності роботи телефонної станції було проведено серію спостережень за частотою виникнення помилкових з'єднань (X). Протягом години фіксувався стан системи з двохвилинним інтервалом, у результаті чого було сформовано вибірку з 30 значень: 3, 2, 1, 0, 1, 2, 3, 4, 3, 1, 5, 2, 4, 3, 1, 2, 4, 0, 1, 2, 2, 1, 0, 1, 2, 3, 1, 0, 2, 7. Побудувати емпіричну функцію розподілу за отриманими даними.

Розв'язання. Очевидно, що число X є випадковою дискретною величиною, а отримані дані є значення цієї випадкової величини. Обсяг вибірки n=30.

Впорядкуємо дані та підрахуємо частоту кожної варіанти. У підсумку маємо сім значень випадкової величини (варіанти): 0, 1, 2, 3, 4, 5, 7. При цьому значення 0 у цій групі зустрічається n_1=4 рази, значення 1 – n_2=8 разів, значення 2 – n_3=8 разів, значення 3 – n_4=5 разів, значення 4 – n_5=3 рази, значення 5 – n_6=1 раз, значення 7 – n_7=1 раз.

Складемо таблицю для дискретного варіаційного ряду з обчисленими значеннями частот і відносних частот (табл. 3.13):

Таблиця 3.13

Кількість неправильних з'єднань x_i	0	1	2	3	4	5	7	Разом
Частота n_i	4	8	8	5	3	1	1	30
Відносна частота $w_i = \frac{n_i}{n}$	4/30	8/30	8/30	5/30	3/30	1/30	1/30	1

Обчислимо емпіричну функцію розподілу за формулою (3.11), використовуючи складений дискретний варіаційний ряд:

$$F^*(x)=\begin{cases}0, & x\le 0\\ 0{,}13, & 0<x\le 1\\ 0{,}4, & 1<x\le 2\\ 0{,}67, & 2<x\le 3\\ 0{,}84, & 3<x\le 4\\ 0{,}94, & 4<x\le 5\\ 0{,}97, & 5<x\le 7\\ 1{,}0, & x>7\end{cases}$$

Графік емпіричної функції розподілу має вигляд (рис. 3.11):

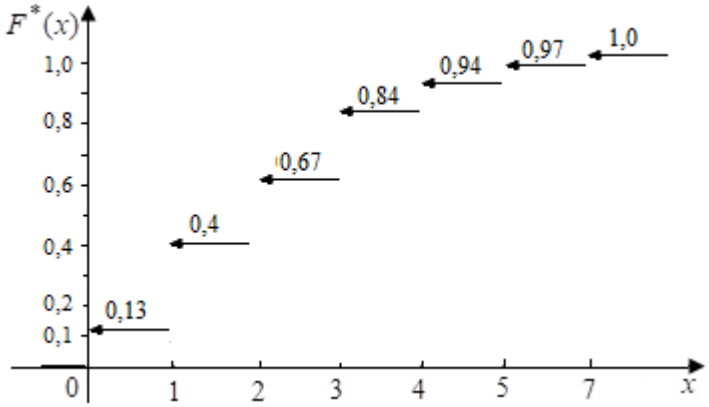


Рисунок 3.11

Приклад 12. Тест з математики оцінюється в 20 балів. Здавали його 20 студентів, які отримали наступні бали: 10, 12, 8, 15, 10, 12, 18, 15, 8, 10, 12, 20,15,8,18,12,10,15,20,12. Побудувати графік кумулятивної кривої (кумуляти).

Розв'язання. Впорядкуємо дані та підрахуємо частоту кожної варіанти. Для дискретного варіаційного ряду з обчисленими значеннями частот, накопичених частот і накопичених частостей

складемо таблицю 3.14:

Таблиця 3.14

№ варіанти	Бали x_i	Частота n_i	Накопичена частота n_i'	Накопичена частость $w_i' = \frac{n_i'}{n} \cdot 100\%$
1	8	3	3	(3/20)·100%=15%
2	10	4	3+4=7	(7/20)·100%=35%
3	12	5	7+5=12	(12/20)·100%=60%
4	15	4	12+4=16	(16/20)·100%=80%
5	18	2	16+2=18	(18/20).100%=90%
6	20	2	18+2=20	(20/20)·100%=100%

По осі абсцис (X) відкладаємо значення ознаки (бали x_i). По осі ординат (Y) відкладаємо накопичену частоту (або накопичену частину у відсотках). Для кожного значення балу відзначаємо точку, що відповідає його накопиченій частості. З'єднуємо отримані точки ламаною лінією.

Графік кумулятивної кривої має вигляд (рис. 3.12):

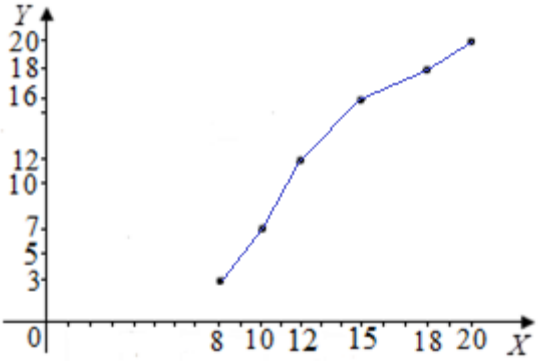


Рисунок 3.12

Приклад 13. Провели опитування серед 50 пацієнтів лікарні стосовно їх віку. За отриманими даними склали інтервальний ряд:

Інтервали	[15;25)	[25;35)	[35;45)	[45;55)	[55;65]
Частота n_i	7	12	18	8	5

Побудувати графік кумулятивної кривої (кумуляти).

Розв'язання. Визначимо накопичені частоти та накопичену частость. Складемо таблицю 3.15:

Таблиця 3.15

Інтервали	Частота n_i	Накопичена частота n_i'	Накопичена частость $w_i' = \frac{n_i'}{n} \cdot 100\%$
[15;25)	7	7	(7/50)·100%=14%
[25;35)	12	7+12=19	(19/50)·100%=38%
[35;45)	18	19+18=37	(37/50)·100%=74%
[45;55)	8	37+8=45	(45/50)·100%=90%
[55;65]	5	45+5=50	(50/50)·100%=100%

По осі абсцис (X) відкладаємо верхні межі інтервалів. По осі ординат (Y) відкладаємо накопичену частоту (або накопичену частость у відсотках). Для кожної верхньої межі інтервалу відзначаємо точку, що відповідає накопиченій частоті.

Графік кумулятивної кривої має вигляд (рис. 3.13):

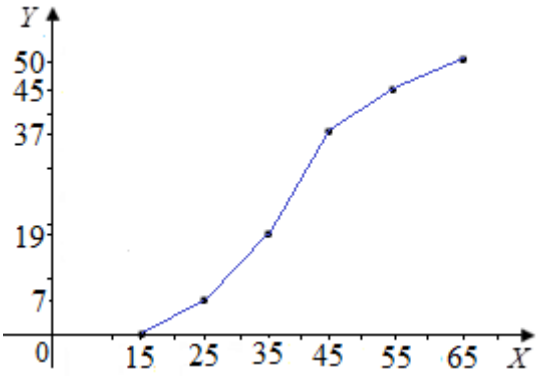


Рисунок 3.13

Приклад 14. Отримано дані про кількість побутової техніки, що продається щорічно в магазині протягом 25 днів: 14, 16, 12, 15, 15, 23, 9, 13, 15, 14, 14, 15, 21, 14, 17, 15, 27, 16,12, 16, 19, 14, 16, 17, 12. Скласти інтервальний статистичний ряд, емпіричну функцію для інтервального ряду та гістограму відносних частот.

Розв'язання. Визначимо кількість інтервалів.

Маємо n=25, x_{min}=9, x_{max}=27. Розмах вибірки: R=x_{max}–x_{min}=27–9=18.

Визначаємо кількість інтервалів k за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg(25) \approx 1 + 3{,}322 \cdot 1{,}398 \approx 1 + 4{,}644 \approx 5{,}644 .$$

Округлюємо k: k=6. Обчислюємо довжину інтервалу h за формулою (3.5): $h=R/k=18/6=3$.

Визначаємо межі інтервалів: [9;12), [12;15), [15;18), [18;21), [21;24), [24;27] та кількість значень продукції для кожного з них: 1,9,11,1,2,1.

Знайдемо значення накопиченої частоти n_i' та накопиченої частості w_i' для кожного інтервалу. Складемо таблицю 3.16:

Таблиця 3.16

Інтервали	[9;12)	[12;15)	[15;18)	[18;21)	[21;24)	[24;27]
Частота n_i	1	9	11	1	2	1
Відносна частота w_i	0,04	0,36	0,44	0,04	0,08	0,04
Накопичена частота n_i'	1	10	21	22	24	25
Накопичена частость w_i'	0,04	0,4	0,84	0,88	0,96	1

Графік емпіричної функції для інтервального ряду має вигляд (рис. 3.14):

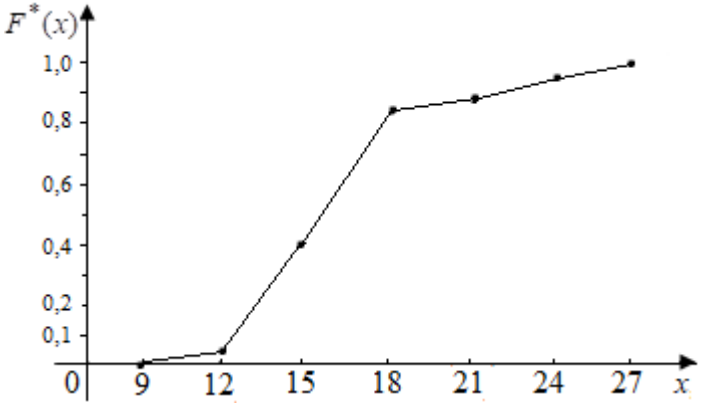


Рисунок 3.14

Оскільки емпірична функція розподілу $F^*(x)$ визначається за формулою (3.11), то відповідні суми – це значення відповідної

накопиченої частості для відповідного інтервалу. Тому при побудові емпіричної функції треба з'єднати точки графіка, що відповідають кінцям інтервалів та відповідної накопиченої частості, відрізками прямої.

Графік гістограми відносних частот має вигляд (рис. 3.15):

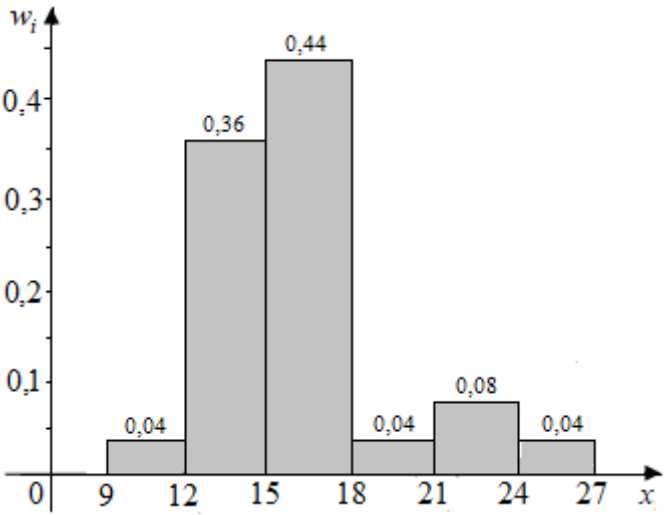
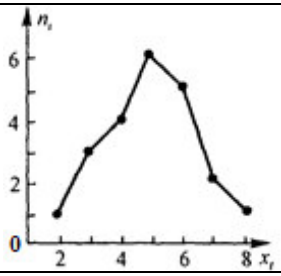


Рисунок 3.15

3.5 Завдання для самостійної роботи

Завдання	Відповідь
1. Утворивши дискретний варіаційний ряд для вибірки: {5,3,7, 10,5,5,2,10,7,2,7,7,4,2,4}, визначити відносні частоти варіант.	w_i : 0,2; 0,067; 0,133; 0,2; 0,267; 0,133.
2. Задано дискретний варіаційний ряд: x_i | 2 | 5 | 7 n_i | 5 | 4 | 3 Знайти накопичену частоту.	n_i' : 5, 9, 12 .
3. Знайти накопичені частості щодо даного розподілу вибірки: x_i | 15 | 20 | 25 | 30 | 35 n_i | 10 | 15 | 30 | 20 | 25	w_i' : 0,1; 0,25; 0,55; 0,75; 1 .

Завдання	Відповідь
4. Знайти емпіричну функцію щодо даного розподілу вибірки: x_i | 1 | 5 | 9 n_i | 15 | 10 | 25	$F^*(x)=\begin{cases}0, & x\le 1\\ 0{,}3, & 1<x\le 5\\ 0{,}5, & 5<x\le 9\\ 1, & x>9\end{cases}$
5. Для заданої вибірки: $\{5,8,10,12,14,15,16,18,19,20,21,23,25,27,30\}$ для побудови інтервального ряду визначити розмах, кількість інтервалів за формулою Стерджесса і довжину.	$R=25$, $k=5$, $h=5$.
6. Для заданої вибірки: $\{2,3,6,3,5,4,5,6,4,5,3,7,6,4,5,4,8,6,5,6,5,7\}$ побудувати полігон частот.	

4. Статистичні оцінки параметрів розподілу

В математичній статистиці буває ситуація, коли потрібно дізнатися деякі характеристики (параметри) генеральної сукупності (тобто всієї сукупності об'єктів чи даних, які нас цікавлять). Наприклад, середній зріст всіх чоловіків у країні, середній бал студентів університету, або відсоток бракованої продукції на заводі.

Проте, зазвичай, неможливо чи недоцільно досліджувати всю генеральну сукупність. Натомість можна взяти вибірку – підмножину генеральної сукупності (ГС), яку можемо спостерігати та вимірювати.

Означення 4.1. *Статистична оцінка* θ^* *параметра* θ – це функція від елементів вибірки, яку побудовану за тим самим законом, як і оцінюваний параметр ГС і яка використовується для оцінювання невідомих параметрів генеральної сукупності.

Наприклад, якщо треба дізнатися середній дохід населення (параметр генеральної сукупності), то беремо випадкову вибірку людей, розраховуємо їхній середній дохід (це і буде статистична оцінка) і використовуємо це значення як наближення до справжнього середнього доходу всього населення.

Важливо розуміти, що оцінка, будучи функцією від випадкових величин (спостережень вибірки), теж є випадковою величиною. Тому вона має свої властивості (наприклад, незміщеність, спроможність, ефективність), які показують, наскільки "якісна" ця оцінка.

Означення 4.2. ***Незміщеною*** називають таку статистичну оцінку θ^*, математичне сподівання якої точно збігається з істинним значенням досліджуваного параметра θ при будь-якому обсязі вибірки n. Це свідчить про відсутність систематичної помилки у процедурі оцінювання: $M\left(\theta^*\right)=\theta$.

Тобто, якщо з однієї ГС утворити багато вибірок і для кожної визначити значення оцінки, то у випадку її незміщеності в середньому значення цих оцінок будуть збігатися з справжнім (невідомим нам) значенням параметра ГС.

Незміщеність гарантує відсутність систематичної помилки. Якщо оцінку зміщена, вона постійно занижуватиме або завищуватиме справжнє значення параметра.

Означення 4.3. Оцінка θ^* параметра θ називається *спроможною*, якщо зі збільшенням обсягу вибірки n справедливо, що для будь-якого скільки завгодно малого $\varepsilon>0$ $\lim\limits_{n\to\infty} P\left(\left|\theta^*-\theta\right|<\varepsilon\right)=1$.

Таким чином, зі збільшенням обсягу вибірки практично достовірним є те, що $\theta^*\approx\theta$.

Теорема 4.1. Якщо оцінка θ^* параметра θ є незміщеною і $D\left(\theta^*\right)\to 0$ при $n\to\infty$, то оцінка θ^* є спроможною.

Якщо оцінка неспроможна, навіть збільшення обсягу вибірки не гарантує, що оцінка буде близька до справжнього значення.

Означення 4.4. Серед множини незміщених оцінок параметра θ *ефективною* вважається та оцінка θ^*, яка має мінімально можливу дисперсію, тобто $D\left(\theta^*\right)\to 0$ при $n\to\infty$.

Тобто, якщо маємо кілька незміщених оцінок для того самого параметра, то найефективніша з них буде "кращою" в тому сенсі, що її значення будуть найменш розкидані навколо справжнього значення параметра. Це означає, що вона дає найбільш "точні" чи "компактні" результати.

Ефективність дозволяє вибрати найкращу з кількох незміщених оцінок. Оцінка з меншою дисперсією вимагає меншого обсягу вибірки для досягнення тієї ж точності, чим менш ефективна оцінка.

Розрізняють два основні типи оцінок: точкова оцінка і інтервальна оцінка (довірчий інтервал).

Точкові та інтервальні оцінки відносяться до параметрів генеральної сукупності, але обчислюються за даними вибірки. Мета статистичної оцінки (як точкової, так і інтервальної) – отримати інформацію саме про ці невідомі параметри ГС.

Означення 4.5. *Точкова оцінка* – це одне число, яке є найкращим наближенням для невідомого параметра.

Точкові оцінки використовуються, коли:

1) Для швидкого і короткого уявлення. Коли потрібно швидко отримати уявлення про величину параметра без детального аналізу точності. Наприклад, середній зріст студентів у групі становить 175 см.

2) Для попереднього аналізу даних. На початкових етапах дослідження, коли формується загальне уявлення про дані, точкові оцінки можуть бути корисними. *Як вхідні дані для подальших розрахунків: Точкові оцінки часто використовуються як відправна точка для складніших статистичних моделей або розрахунків, де потрібно одне число*

3) Для порівняння: Коли хочемо порівняти середні значення або частки між різними групами, точкові оцінки зручні для безпосереднього порівняння.

4) У повсякденних звітах та зведеннях: Для простоти сприйняття та стислості.

5) При дуже великих вибірках: Як основу для інтервальних оцінок та інших статистичних процедур. Точкова оцінка зазвичай буде дуже близька до справжнього значення параметра і її точність буде високою.

Для точкових оцінок важливі такі вимоги: незміщеність, спроможність, ефективність, достатність, робастність.

Обмеженість точкових оцінок. Головний недолік точкових оцінок полягає в тому, що вони не дають жодної інформації щодо точності цієї оцінки. Точкова оцінка майже напевно не дорівнює справжньому значення параметра, і ми не знаємо, наскільки вона близька до нього.

4.1 Точкові оцінки параметрів генеральної сукупності

Означення 4.6. *Параметри генеральної сукупності* – це числові характеристики, що описують всю досліджувану генеральну сукупність.

Вони а) *описують всю сукупність*, тобто відносяться до всіх об'єктів або явищ, що становлять інтерес для дослідження; б) *фіксовані та унікальні*, тобто для цієї генеральної сукупності значення параметра є одним і незмінним; в) *часто невідомі*, тобто на практиці виміряти кожен елемент генеральної сукупності найчастіше неможливо або недоцільно (наприклад, через великий розмір, вартість або руйнівний характер вимірювання); г) є цільовою оцінкою, тобто саме ці параметри необхідно оцінити, використовуючи дані з вибірки.

Параметри ГС – це *постійні величини*, які не змінюються, якщо сама генеральна сукупність залишається незмінною.

Розглянемо найчастіше використовувані параметри ГС.

Означення 4.7. *Генеральною середньою* ($x_г$) називається середнє арифметичне значень ознаки генеральної сукупності (ГС), яка обчислюється за формулою:

якщо x_i різні

$$x_{\text{г}} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \,, \tag{4.1}$$

де N – обсяг ГС, x_i – i-е спостереження у ГС;

якщо дані згруповані (x_i мають частоти N_i, $i = \overline{1,k}$)

$$x_{\text{г}} = \frac{1}{N} \cdot \sum_{i=1}^{k} x_i \cdot N_i \,. \tag{4.2}$$

Означення 4.8. *Дисперсією генеральної сукупності* ($D_{\text{г}}$) називають статистичний показник, що обчислюється як середній квадрат відхилення значень досліджуваної ознаки від їхнього загального середнього значення. Математично цей параметр відображає ступінь неоднорідності генеральної сукупності і обчислюється за формулою:

якщо x_i різні

$$D_{\text{г}} = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(x_i - x_{\text{г}}\right)^2 \,, \tag{4.3}$$

де N – обсяг ГС, x_i – i-е спостереження у ГС;

якщо дані згруповані (x_i мають частоти N_i, $i = \overline{1,k}$)

$$D_{\text{г}} = \frac{1}{N} \cdot \sum_{i=1}^{k} \left(x_i - x_{\text{г}}\right)^2 \cdot N_i \,. \tag{4.4}$$

Додатковою мірою варіативності даних у генеральній сукупності є середнє квадратичне відхилення ($\sigma_{\text{г}}$). На відміну від дисперсії, цей параметр вимірюється в тих самих одиницях, що й сама досліджувана ознака.

Означення 4.9. *Генеральним середнє квадратичним відхиленням* ($\sigma_{\text{г}}$) називається корінь квадратний із генеральної дисперсії:

$$\sigma_{\text{г}} = \sqrt{D_{\text{г}}} \,. \tag{4.5}$$

4.2 Точкові оцінки параметрів вибірки

Означення 4.10. *Параметри вибірки* (або *вибіркові статистики*) – це числові характеристики, обчислені на основі даних, зібраних із конкретної вибірки, яка є підмножиною генеральної сукупності.

Вони а) *описують лише вибірку*, тобто належать лише до тих елементів, які були включені до конкретної вибірки; б) *варіюються від вибірки до вибірки*, тобто, якщо взяти іншу випадкову вибірку з тієї ж генеральної сукупності, значення вибіркових статистик, швидше за все, трохи відрізнятимуться; в) *завжди відомі* (після збору даних), тобто їх можна обчислити, маючи дані щодо вибірки; г) *використовуються для оцінки параметрів генеральної сукупності*, тобто мета обчислення вибіркових статистик – отримати уявлення про невідомі параметри генеральної сукупності.

Вибіркові статистики використовуються як точкові оцінки відповідних параметрів генеральної сукупності.

Розглянемо найчастіше використовувані параметри ГС і вибірки. Нехай маємо ГС обсягу N і для вивчення ГС вилучено з неї вибірку обсягу n.

Означення 4.11. *Вибірковою середньою* ($\bar{x}_{\text{в}}$) називається середнє арифметичне значень ознаки вибіркової сукупності (вибірки).

Вибіркова середня визначається за формулою:

а) для дискретного варіаційного ряду

якщо всі x_i різні

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \tag{4.6}$$

якщо дані згруповані (x_i мають частоти n_i, $i = \overline{1,k}$)

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i \ ; \tag{4.7}$$

б) для інтервального варіаційного ряду

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i \cdot n_i = \sum_{i=1}^{k} \bar{x}_i \cdot w_i \ , \qquad (4.8)$$

де $\bar{x}_i$ – середина i-го інтервалу, $i = \overline{1,k}$.

Зазначимо, що основні властивості вибіркової середньої, аналогічні до властивостей математичного сподівання випадкової величини.

Вибіркову середню приймають як оцінку для ГС, і ця оцінка є незміщеною, тобто $M\left(\bar{x}_{\text{в}}\right) = x_{\text{г}}$. Вона є також і спроможною. При збільшенні обсягу вибірки $\bar{x}_{\text{в}} \to x_{\text{г}}$, і для різних вибірок з однієї ГС значення $\bar{x}_{\text{в}}$ приблизно дорівнюють між собою.

Статистичний розкид кількісної ознаки *X* відносно центрального значення фіксується за допомогою показників дисперсії.

Оцінка варіативності даних проводиться на двох рівнях: для кількісного опису розкиду значень у межах усієї генеральної сукупності застосовують *генеральну дисперсію*, тоді як для аналізу мінливості ознаки безпосередньо у вибірці використовують її практичний аналог – *вибіркову дисперсію*.

Означення 4.12. *Вибіркова дисперсія* ($D_{\text{в}}$) – це статистичний показник, що характеризує ступінь розсіювання значень усередині вибірки. Вона обчислюється як середній квадрат відхилення кожної варіанти від вибіркового середнього значення $\bar{x}_{\text{в}}$.

Вибіркова дисперсія оцінює дисперсію генеральної сукупності та є зміщеною оцінкою. "Зміщена" вибіркова дисперсія $D_{\text{в}}$ використовується для опису розкиду у вибірці.

Вибіркова дисперсія визначається за формулою:

а) для дискретного варіаційного ряду

якщо всі x_i різні

$$D_{\text{в}} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_{\text{в}})^2 \ , \qquad (4.9)$$

якщо дані згруповані (x_i мають частоти n_i, $i=\overline{1,k}$)

$$D_{\text{в}} = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_{\text{в}})^2 \cdot n_i ; \tag{4.10}$$

б) для інтервального варіаційного ряду

$$D_{\text{в}} = \frac{1}{n} \cdot \sum_{i=1}^{k} (\bar{x}_i - \bar{x}_{\text{в}})^2 \cdot n_i , \tag{4.11}$$

де $\bar{x}_i$ – середина i-го інтервалу, $i=\overline{1,k}$.

Окрім дисперсії, для оцінки розкиду даних у межах вибірки використовують вибіркове середнє квадратичне відхилення. Цей показник є зручнішим для інтерпретації, оскільки має ті самі одиниці вимірювання, що й досліджувана ознака.

Означення 4.13. *Вибірковим середнім квадратичним відхиленням* ($\sigma_{\text{в}}$) називається арифметичний квадратний корінь із вибіркової дисперсії:

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} . \tag{4.12}$$

При обчисленні дисперсії генеральної або вибіркової зручно використовувати формулу:

$$D = \overline{x^2} - \left(\bar{x}\right)^2 . \tag{4.13}$$

Оскільки вибіркова дисперсія є зміщеною оцінкою, то її середнє значення (математичне сподівання) не збігається з істинної дисперсією генеральної сукупності. Виявляється, що вона систематично занижує істинну дисперсію.

Формула дисперсії містить суму квадратів відхилень від середнього. Коли працюємо з вибіркою, то використовуємо вибіркове середнє ($\bar{x}_{\text{в}}$), а не істинне середнє генеральної сукупності ($x_{\text{г}}$), яке, як правило, невідоме. Сума квадратів відхилень від вибіркового середнього завжди буде меншою або дорівнюватиме сумі квадратів

відхилень від будь-якого іншого значення (включаючи істинне середнє генеральної сукупності). Через це ділення на n призводить до заниженої оцінки.

Щоб отримати незміщену оцінку дисперсії генеральної сукупності, використовується виправлена вибіркова дисперсія (також відома як незміщена дисперсія). Вона відрізняється від звичайної тим, що у знаменнику використовується n–1 замість n.

Означення 4.14. *Виправлена вибіркова дисперсія* (s^2) – це незміщена та обґрунтована оцінка дисперсії випадкової величини в генеральній сукупності, яка найчастіше використовується у висновках, що стосуються генеральної сукупності.

Вона обчислюється за формулою:

$$s^2 = \frac{n}{n-1} D_{\text{в}}. \tag{4.14}$$

Означення 4.15. Середнє квадратичне відхилення, яке розраховане на основі незміщеної дисперсії, називається *виправленим середнім квадратичним відхиленням* (s).

Виправлене вибіркове значення (s) застосовують для отримання найбільш точної та незміщеної оцінки середнього квадратичного відхилення в генеральній сукупності. Це дозволяє компенсувати систематичну похибку, яка виникає при аналізі малих обсягів даних. Воно обчислюється за формулою:

$$s = \sqrt{s^2}. \tag{4.15}$$

Хоча виправлене середнє квадратичне відхилення технічно є зміщеною оцінкою, однак, на практиці його вважають найкращою оцінкою, особливо для невеликих вибірок, і різниця зменшується зі збільшенням розміру вибірки.

Розглянуті вище міри розсіювання є розмірними величинами. Тому можуть виникати труднощі при їх порівнянні для вибірок із різних генеральних сукупностей. Найбільш поширеною безрозмірною мірою розсіювання є коефіцієнт варіації V^*.

Означення 4.16. *Коефіцієнт варіації* (V^*) – це відносна міра мінливості або розсіювання даних, яка показує, наскільки значення у вибірці або генеральній сукупності відхиляються від їхнього середнього значення.

Він виражається у відсотках і дозволяє порівнювати ступінь розкиду даних у різних сукупностях, навіть якщо середні значення сильно відрізняються.

Формула для розрахунку коефіцієнта варіації:

$$V^* = \frac{s}{\bar{x}} \cdot 100\%, \tag{4.16}$$

де s – виправлене середнє квадратичне відхилення; $\bar{x}$ – середнє арифметичне вибірки. Ці параметри розраховуються в залежності від того, яким чином представлені дані.

Треба звернути увагу, щоб значення x_i були додатними.

Коефіцієнт варіації використовується з метою оцінки однорідності вибірки. Якщо коефіцієнт варіації менший або дорівнює 33%, то вибірка вважається однорідною, тобто розкид даних відносно невеликий, а середнє значення добре характеризує всю сукупність. Якщо коефіцієнт варіації перевищує 33%, то вибірка неоднорідна. У цьому випадку розкид даних значний, і середнє значення вже не є надійним показником, що характеризує всю сукупність.

Іноді використовують термін "компактна вибірка", який не є загальноприйнятим статистичним терміном і часто використовується як синонім "однорідної вибірки" у контексті коефіцієнта варіації – це вибірка, в якій дані щільно згруповані навколо середнього значення.

Значення коефіцієнта варіації, що не виходять за межі 10%, прийнято вважати компактними.

Приклад 15. Для заданого дискретного варіаційного ряду

x_i	2	3	4	5
n_i	4	8	5	3

Знайти вибіркову середню $\bar{x}_в$, вибіркову дисперсію $D_в$, вибіркове середнє квадратичне відхилення $\sigma_в$ та виправлену вибіркову дисперсію s^2.

Розв'язання. Обсяг вибірки $n = \sum_{i=1}^{4} n_i = 4+8+5+3 = 20$. Вибіркову середню $\bar{x}_{\text{в}}$ знаходимо за формулою (4.7):

$$\bar{x}_{\text{в}} = \frac{1}{20}\left(2\cdot 4 + 3\cdot 8 + 4\cdot 5 + 5\cdot 3\right) = \frac{1}{20}\cdot 67 = 3{,}35\,.$$

Вибіркову дисперсію $D_{\text{в}}$ знаходимо за формулою (4.10):

$$D_{\text{в}} = \frac{1}{20}\cdot\left((2-3{,}35)^2\cdot 4 + (3-3{,}35)^2\cdot 8 + (4-3{,}35)^2\cdot 5 + (5-3{,}35)^2\cdot 3\right) =$$

$$= \frac{1}{20}\cdot\left(1{,}8225\cdot 4 + 0{,}1225\cdot 8 + 0{,}4225\cdot 5 + 2{,}7225\cdot 3\right) = \frac{1}{20}\cdot\left(7{,}29 + 0{,}98 +\right.$$

$$\left. + 2{,}1125 + 8{,}1675\right) = \frac{1}{20}\cdot 18{,}55 = 0{,}9275$$

Вибіркове середнє квадратичне відхилення $\sigma_{\text{в}}$ знаходимо за формулою (4.12): $\sigma_{\text{в}} = \sqrt{0{,}9275} \approx 0{,}963$.

Виправлену вибіркову дисперсію s^2 знаходимо за формулою (4.14):

$$s^2 = \frac{20}{19}\cdot 0{,}9275 \approx 0{,}9763\,.$$

Відповідь: $\bar{x}_{\text{в}} = 3{,}35$; $D_{\text{в}} = 0{,}9275$; $\sigma_{\text{в}} \approx 0{,}963$; $s^2 \approx 0{,}9763$.

Приклад 16. Для заданого інтервального варіаційного ряду

Інтервали	[10;15)	[15;20)	[20;25)	[25;30]
n_i	6	7	8	2

Знайти вибіркову середню $\bar{x}_{\text{в}}$, вибіркову дисперсію $D_{\text{в}}$ та вибіркове середнє квадратичне відхилення $\sigma_{\text{в}}$.

Розв'язання. Обсяг вибірки n=6+7+5+2=20. Визначимо середини інтервалів за формулою $\bar{x}_i = \frac{\alpha_i + \alpha_{i+1}}{2}$. Матимемо

$\bar{x}_i$	12,5	17,5	22,5	27,5

Вибіркову середню $\bar{x}_{\text{в}}$ знаходимо за формулою (4.8):

$$\bar{x}_{\text{в}} = \frac{1}{20}(12{,}5 \cdot 6 + 17{,}5 \cdot 7 + 22{,}5 \cdot 5 + 27{,}5 \cdot 2) = \frac{1}{20} \cdot 365 = 18{,}25.$$

Вибіркову дисперсію $D_{\text{в}}$ знаходимо за формулою (4.13): $D_{\text{в}} = \overline{x_{\text{в}}^2} - (\overline{x_{\text{в}}})^2$. Знайдемо

$$\overline{x_{\text{в}}^2} = \frac{1}{20}\left(12{,}5^2 \cdot 6 + 17{,}5^2 \cdot 7 + 22{,}5^2 \cdot 5 + 27{,}5^2 \cdot 2\right) = \frac{1}{20} \cdot 7125 = 356{,}25.$$

Тоді $D_{\text{в}}$=356,25–18,25^2=356,25–333,0625=23,1875.

Вибіркове середнє квадратичне відхилення $\sigma_{\text{в}}$ знаходимо за формулою (4.12): $\sigma_{\text{в}} = \sqrt{23{,}1875} \approx 4{,}8153$.

Відповідь: $\bar{x}_{\text{в}} = 18{,}25$; $D_{\text{в}} = 23{,}1875$; $\sigma_{\text{в}} \approx 4{,}8153$.

Приклад 17. Маємо дані про вік співробітників невеликого кафетерія: 28,25,27,26,24. З'ясувати, чи є вік працюючих, хорошим показником для цього закладу.

Розв'язання. Обчислимо середнє арифметичне ($\bar{x}$) за формулою (4.6), оскільки маємо дискретний варіаційний ряд, у якому всі x_i різні (n=5): $\bar{x} = (28 + 25 + 27 + 26 + 24)/5 = 130/5 = 26$.

Обчислимо виправлене середнє квадратичне відхилення (s). Спочатку знайдемо суму квадратів відхилень:

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (28-26)^2 + (25-26)^2 + (27-26)^2 + (26-26)^2 +$$

$$+ (24-26)^2 = 2^2 + (-1)^2 + 1^2 + 0^2 + (-2)^2 = 4 + 1 + 1 + 0 + 4 = 10.$$

Виправлену дисперсію (s^2) знайдемо за формулою $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$: $s^2 = \frac{10}{5-1} = 2{,}5$.

Виправленим середнє квадратичним відхиленням (s):

$$s = \sqrt{2{,}5} \approx 1{,}5811.$$

Тоді коефіцієнт варіації V^* знаходимо за формулою (4.16):

$$V^* = \frac{1{,}5811}{26} \cdot 100\% \approx 6{,}08\%.$$

Відповідь: Оскільки 6,08%<33%, то вибірка вважається однорідною (компактною). Це означає, що вік співробітників дуже

близький до середнього значення 26 років.

Приклад 18. Маємо дані про вік співробітників великої корпорації, що включає як стажистів, так і керівників: $22, 23, 24, 55, 60$. З'ясувати, чи є вік працюючих, хорошим показником для цієї корпорації.

Розв'язання. Обчислимо середнє арифметичне ($\bar{x}$) за формулою (4.6), оскільки маємо дискретний варіаційний ряд, у якому всі x_i різні (n=5): $\bar{x} = (22 + 23 + 24 + 55 + 60)/5 = 184/5 = 36{,}8$.

Обчислимо виправлене середнє квадратичне відхилення (s). Спочатку знайдемо суму квадратів відхилень:

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (22-36{,}8)^2 + (23-36{,}8)^2 + (24-36{,}8)^2 + (55-36{,}8)^2 +$$

$$+ (60-36{,}8)^2 = (-14{,}8)^2 + (-13{,}8)^2 + (-12{,}8)^2 + (18{,}2)^2 + (23{,}2)^2 =$$

$$= 219{,}04 + 190{,}44 + 163{,}84 + 331{,}24 + 538{,}24 = 1442{,}8 \ .$$

Виправлену дисперсію (s^2) знайдемо за формулою $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$: $s^2 = \frac{1442{,}8}{5-1} = 360{,}7$.

Виправленим середнє квадратичним відхиленням (s):

$$s = \sqrt{360{,}7} \approx 18{,}9921.$$

Тоді коефіцієнт варіації V^* знаходимо за формулою (4.16):

$$V^* = \frac{18{,}9921}{36{,}8} \cdot 100\% \approx 51{,}61\% \ .$$

Відповідь: Оскільки 51,61%>33%, то вибірка вважається неоднорідною. Це означає, що середній вік у 36,8 років не є хорошим показником для всієї корпорації, тому що в ній є як дуже молоді, так і досить літні співробітники.

Частка генеральної сукупності та вибірки

Означення 4.17. *Частка* (*пропорція*) *генеральної сукупності* (p) – це відношення кількості об'єктів, що мають певну властивість, до загальної кількості об'єктів у генеральній сукупності.

Вона показує, яка частина сукупності володіє заданою ознакою. Позначається зазвичай літерою p (або π, коли йдеться про параметр сукупності). Застосовується у статистиці для опису та аналізу даних, а також для оцінки параметрів сукупності на основі вибіркових даних та обчислюється за формулою:

$$p = \frac{k}{N}, \tag{4.17}$$

де N – обсяг ГС, k – кількість об'єктів у генеральній сукупності, що мають певну ознаку.

Приклади використання (коли p відома або вважається відомою):

1. *Контроль якості на виробництві*. Наприклад. Виробнича лінія виробляє деталі, і визначено, що 5% всіх вироблених деталей дефектними. Тут p=0,05. Ця інформація використовується для встановлення стандартів якості, прогнозування кількості дефектних виробів та планування обслуговування обладнання. Тобто, якщо відома частка дефектних деталей в генеральній сукупності, можна розраховувати на певний рівень браку та приймати рішення про постачання, гарантійні зобов'язання та необхідність доопрацювання виробничого процесу.

2. *Епідеміологія* (поширеність захворювання). Наприклад. Відомо, що 10% населення є носіями певного вірусу. Тут p=0,1. Це знання допомагає планувати охорону здоров'я, розподіляти ресурси, організовувати кампанії з вакцинації. Тобто, знаючи справжню поширеність (частку) захворювання, можна оцінити потребу в медичному персоналі, кількості ліків та лікарняних ліжок.

3. *Демографія* (співвідношення статей, вікові групи). Наприклад. За даними перепису населення, частка жінок у місті становить 53%. Тут p=0,53. Ця інформація використовується для міського планування, розроблення соціальних програм, маркетингових досліджень. Тобто, знаючи точну частку жінок, місцева влада може планувати будівництво шкіл, дитячих садків, соціальних центрів, враховуючи потреби певної групи населення.

Означення 4.18. *Частка вибірки* ($\overline{p}$) – це пропорція елементів, що мають певну ознаку, у випадково відібраній вибірці з генеральної сукупності.

Вона використовується з метою оцінки невідомої частки генеральної сукупності та обчислюється за формулою:

$$\overline{p} = \frac{k}{n}, \tag{4.18}$$

де n – обсяг вибірки, k – кількість варіант у вибірці, що мають певну ознаку.

Приклади використання:

1. *Соціологічні опитування та політичні дослідження*. Наприклад. Перед виборами проводиться опитування 1000 виборців, із яких 450 висловлюють підтримку кандидату А. Частка вибірки $\overline{p} = 450/1000 = 0{,}45$ (або 45%). Тобто на основі цієї вибіркової частки створюються довірчі інтервали для оцінки справжньої частки підтримки кандидата А серед усього населення. Це допомагає кандидатам та їхнім штабам приймати стратегічні рішення.

2. *Маркетингові дослідження*. Наприклад. Компанія хоче дізнатися, який відсоток потенційних покупців зацікавлений у новому продукті. Проводять опитування 500 осіб із яких 150 висловили зацікавленість. Частка вибірки $\overline{p} = 150/500 = 0{,}30$ (або 30%). Тобто на основі цієї вибіркової частки можна оцінити ринковий потенціал продукту, передбачити обсяг продажу та вирішити, чи варто запускати продукт на ринок.

3. *Медичні дослідження* (ефективність нового препарату). Наприклад. Для оцінки ефективності нових ліків від високого тиску, їх дають 200 пацієнтам, із яких 160 повідомляють про поліпшення стану. Частка вибірки $\overline{p} = 160/200 = 0{,}80$ (або 80%). Тобто ця вибіркова частка дозволяє зробити висновки щодо ефективності препарату для всіх пацієнтів з високим тиском. Проводяться статистичні тести, щоб визначити, чи спостерігається поліпшення статистично значимим.

4. *Контроль якості* (вибірковий контроль). Наприклад. На підприємстві, що виробляє лампочки, з партії 10000 штук випадково відбирають 100 лампочок для перевірки. Виявляється, що 5 із них є браковані. Частка вибірки $\overline{p} = 5/100 = 0{,}05$ (або 5%). Тобто на основі цієї вибіркової частки робиться висновок про якість усієї партії. Якщо

5% дефектних ламп є прийнятним рівнем браку, то партія може бути відправлена. В іншому випадку вся партія може бути повернена на підприємство.

Основна відмінність та взаємозв'язок:

• Частка генеральної сукупності (p) – це справжнє значення для всієї групи, яке часто невідоме і є метою для оцінки.

• Частка вибірки ($\overline{p}$) – це оцінка p, отримана на основі даних, зібраних із підгрупи (вибірки) населення.

$\overline{p}$ використовуємо для того, щоб робити висновки про p, оскільки часто неможливо чи недоцільно досліджувати всю генеральну сукупність. Чим більша і репрезентативніша вибірка, тим точніше частка вибірки $\overline{p}$ наближатися до частки генеральної сукупності p.

Структурні середні

До вибіркових характеристик належать також мода та медіана, які називають ще структурними середніми.

Означення 4.19. *Медіаною* (M_{e}^{*}) варіаційного ряду називається серединна точка в варіаційному ряду, яка ділить варіаційний ряд на дві рівні за кількістю членів частини.

Для варіаційного ряду медіана визначається залежно від того, чи є обсяг вибірки n числом парним або непарним:

$$M_{\text{e}}^{*} = \begin{cases} \dfrac{x_m + x_{m+1}}{2}, & n = 2m, \\ x_{m+1}, & n = 2m+1. \end{cases} \tag{4.19}$$

Для інтервального варіаційного ряду медіану визначають так: а) знаходять інтервал угруповання, в якому міститься медіана шляхом підрахунку накопичених частот або накопичених відносних частот – це інтервал, в якому накопичена частота вперше виявиться більшою за $n/2$ або накопичена частость буде більше 0,5.

Усередині медіанного інтервалу $[x_{\text{m}};x_{\text{m+1}}]$ медіана визначається за такою формулою:

$$M_{\text{e}}^{*} = x_m + \frac{0{,}5\,n - n'_{m-1}}{n_m} \cdot \left(x_{m+1} - x_m\right), \qquad (4.20)$$

де n – обсяг вибірки; x_{m} – нижня межа медіанного інтервалу; $(x_{\text{m+1}}–x_{\text{m}})$ – довжина медіанного інтервалу; n'_{m-1} – накопичена частота інтервалу, попереднього медіанному; n_m – частота медіанного інтервалу.

Означення 4.20. *Мода* (M_{o}^{*}) – найбільш ймовірне значення у вибірці (варіанта з найбільшою частотою), тобто це таке значення варіанти, що попереднє і наступне за ним значення мають менші частоти, в саме $M_{\text{o}}^{*} = x_i$ при умові, що $n(x_i)$=max.

Для інтервального варіаційного ряду знаходять інтервал угруповання з максимальною частотою (модальний інтервал). Всередині модального інтервалу $[x_{\text{m}};x_{\text{m+1}}]$ моду (M_{o}^{*}) знаходять за формулою:

$$M_{\text{o}}^{*} = x_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} \cdot \left(x_{m+1} - x_m\right), \qquad (4.21)$$

де x_{m} – нижня межа модального інтервалу; $(x_{\text{m+1}}–x_{\text{m}})$ – довжина модального інтервалу; n_m – частота модального інтервалу; n_{m-1} та n_{m+1} – частоти відповідно попереднього та наступного за модальним інтервалів.

Приклад 19. Для заданого дискретного ряду

x_i	0	1	2	3	4
n_i	21	46	34	14	9

Знайти моду для даного ряду.

Розв'язання. За означенням мода дорівнює варіанті з найбільшою частотою. У даному прикладі $M_{\text{o}}^{*} = 1$, оскільки їй відповідає найбільша

частота n_i=46.

Відповідь: $M_{\text{o}}^{*}=1$.

Приклад 20. Маємо два ряди розподілу: а) 18,22,23,28,35,40,42 і б) 5,8,18,22,28,35,40,42. Знайти медіану для кожного ряду.

Розв'язання. а) Оскільки ряд містить непарну кількість членів (n=7), то за формулою (4.19) $M_{\text{e}}^{*}=x_3=28$. б) Заданий ряд має парну кількість членів (n=8), то за формулою (4.19) $M_{\text{e}}^{*}=\dfrac{x_4+x_5}{2}=\dfrac{22+28}{2}=25$.

Відповідь: а) $M_{\text{e}}^{*}=28$, б) $M_{\text{e}}^{*}=25$.

Приклад 21. Провели дослідження віку працівників невеликого підприємства, розподіливши на п'ять вікових груп. Отримали наступні результати:

Вікові групи (x_i)	[18;28)	[28;38)	[38;48)	[48;58)	[58;і старше)
Кількість працюючих n_i	14	22	34	28	12

Знайти моду і медіану для даного інтервального ряду.

Розв'язання. Обсяг вибірки n=110. Для знаходження моди застосовуємо формулу (4.21). Для цього знайдемо модальний інтервал. Це інтервал [38;48), для якого частота найбільша (n_i=34). Тоді у формулу підставимо: x_{m}=38, $x_{\text{m}+1}$=48, n_m=34, n_{m-1}=22, n_{m+1}=28,

$$M_{\text{o}}^{*}=38+\frac{34-22}{2\cdot 34-22-28}\cdot(48-38)=38+\frac{12}{18}\cdot 10\approx 44{,}67\ .$$

Для знаходження медіани застосовуємо формулу (4.20). Знайдемо накопичені частоти для кожного інтервалу. Складемо таблицю 4.1:

Таблиця 4.1

Вікові групи (x_i)	[18;28)	[28;38)	[38;48)	[48;58)	[58;і старше)
Кількість працюючих n_i	14	22	34	28	12
n_i'	14	36	70	98	110

Знаходимо медіанний інтервал. Оскільки $n/2=100/2=55$, то медіанним буде інтервал [38;48), для якого $n_i' > n/2$, тобто, 70>55. Тоді у формулу підставимо: $x_{\text{m}}=38$, $x_{\text{m}+1}=48$, $n_m=34$, $n'_{m-1}=36$.

$$M_{\text{e}}^{*}=38+\frac{0{,}5\cdot 110-36}{34}\cdot\left(48-38\right)=38+\frac{19}{34}\cdot 10\approx 43{,}59.$$

Відповідь: $M_{\text{o}}^{*}\approx 44{,}67$; $M_{\text{e}}^{*}\approx 43{,}59$.

Узагальнені числові характеристики варіаційного ряду. Асиметрія і ексцес

Вибіркова середня та вибіркова дисперсія є окремим випадком більш загального поняття – момент статистичного ряду. Моменти – це узагальнені числові характеристики, що описують розподіл даних. Враховуючи ці особливості, можна записати загальні формули для обчислення вибіркових початкових ν_k^* і центральних μ_k^* емпіричних моментів випадкової величини.

Означення 4.21. *Початковим вибірковим моментом порядку k* $\left(\nu_k^*\right)$ називається середнє арифметичне *k*-х степенів значень випадкової величини, що спостерігаються.

Початковий вибірковий момент порядку *k* визначається за формулою:

а) для дискретного варіаційного ряду

якщо всі x_i різні

$$\nu_k^{*}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i\right)^k, \tag{4.22}$$

якщо дані згруповані (x_i мають частоти n_i, $i=\overline{1,l}$)

$$\nu_k^{*}=\frac{1}{n}\sum_{i=1}^{l}\left(x_i\right)^k\cdot n_i; \tag{4.23}$$

б) для інтервального варіаційного ряду

$$\nu_k^* = \frac{1}{n}\sum_{i=1}^{l}\left(\overline{x}_i\right)^k \cdot n_i \,, \tag{4.24}$$

де $\overline{x}_i$ – середина *i*-го інтервалу, $i = \overline{1,l}$.

Означення 4.22. *Центральним вибірковим моментом порядку k* $\left(\mu_k^*\right)$ називається середнє арифметичне *k*-х степенів відхилень значень випадкової величини, що спостерігаються, від їх середнього арифметичного.

Центральний вибірковий момент порядку *k* визначається за формулою:

а) для дискретного варіаційного ряду

якщо всі x_i різні

$$\mu_k^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \nu_1^*\right)^k \cdot n_i \,, \tag{4.25}$$

де ν_1^* – початковий момент порядку k=1;

якщо дані згруповані (x_i мають частоти n_i, $i = \overline{1,l}$)

$$\mu_k^* = \frac{1}{n}\sum_{i=1}^{l}\left(x_i - \nu_1^*\right)^k \cdot n_i \,; \tag{4.26}$$

б) для інтервального варіаційного ряду

$$\mu_k^* = \frac{1}{n}\sum_{i=1}^{l}\left(\overline{x}_i - \nu_1^*\right)^k \cdot n_i \,, \tag{4.27}$$

де $\overline{x}_i$ – середина *i*-го інтервалу, $i = \overline{1,l}$.

Зв'язок між моментами:

$\nu_1^* = \overline{x}$ – середнє арифметичне, $\mu_1^* = 0$,

$\mu_2^* = \nu_2^* - \left(\nu_1^*\right)^2$ – дисперсія,

$$\mu_3^* = \nu_3^* - 3\nu_2^*\nu_1^* + 2\left(\nu_1^*\right)^3,$$

$$\mu_4^* = \nu_4^* - 4\nu_3^*\nu_1^* + 6\nu_2^*\left(\nu_1^*\right)^2 - 3\left(\nu_1^*\right)^4.$$

На основі центральних моментів обчислюються коефіцієнти, що описують форму розподілу даних.

Коефіцієнт асиметрії (*асиметричності*) A_s визначається за формулою:

$$A_s = \frac{\mu_3^*}{\sigma^3}, \tag{4.28}$$

де μ_3^* – центральний момент 3-го порядку, σ – середнє квадратичне відхилення $\left(\sigma = \sqrt{\mu_2^*}\right)$.

Він показує, як розподіл відхиляється від симетричного. Якщо A_s=0, то розподіл симетричний. Якщо A_s>0, то розподіл має правосторонню асиметрію (скошений вліво), тобто більшість значень зосереджена ліворуч від середнього. Якщо A_s<0, то розподіл має ліву асиметрію (скошений вправо), тобто більшість значень зосереджена праворуч від середнього.

Коефіцієнт ексцесу (*крутості*) E_k визначається за формулою:

$$E_k = \frac{\mu_4^*}{\sigma^4} - 3, \tag{4.29}$$

де μ_4^* – центральний момент 4-го порядку, σ – середнє квадратичне відхилення. Віднімання "3" обумовлено тим, що для нормального розподілу $\frac{\mu_4^*}{\sigma^4} = 3$.

Він показує, наскільки гостровершинним або плосковершинним є поточний розподіл у порівнянні з еталонним нормальним розподілом. Якщо E_k=0, то маємо мезокуртичний розподіл. Дані розподілені

симетрично і помірно концентруються навколо центрального значення. Якщо $E_k > 0$, то маємо лептокуртичний розподіл. Це вказує на сильну концентрацію даних навколо середнього значення, але при цьому супроводжується «важкими хвостами» – підвищеною ймовірністю появи рідкісних екстремальних відхилень (викидів). Якщо $E_k > 0$, то маємо платикуртичний розподіл. Пік виражений слабо, а дані розподілені по діапазону рівномірніше, ніж у нормальному законі. Для такого розподілу характерні «легкі хвости», тобто екстремальні значення трапляються дуже рідко.

Приклад 22. Знайти коефіцієнти асиметрії та ексцесу для заданого дискретного ряду:

x_i	2	4	5	7	9
n_i	4	6	7	5	3

Розв'язання. Кількість інтервалів l=5. Знайдемо обсяг вибірки:

$$n = \sum_{i=1}^{5} n_i = 4 + 6 + 7 + 5 + 3 = 25 .$$

Знайдемо середнє арифметичне за формулою (4.7):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{l} x_i \cdot n_i = \frac{1}{25}\cdot\left(2\cdot 4 + 4\cdot 6 + 5\cdot 7 + 7\cdot 5 + 9\cdot 3\right) = \frac{1}{25}\cdot 129 = 5{,}16 .$$

Знаходимо центральні моменти μ_2^*, μ_3^*, μ_4^* за формулою (4.26).

$$\mu_2^* = \frac{1}{n}\sum_{i=1}^{l}\left(x_i - \bar{x}\right)^2 \cdot n_i = \frac{1}{25}\cdot\left((2-5{,}16)^2\cdot 4 + (4-5{,}16)^2\cdot 6 + (5-5{,}16)^2\cdot 7 + \right.$$

$$\left. + (7-5{,}16)^2\cdot 5 + (9-5{,}16)^2\cdot 3\right) = \frac{1}{25}\left(9{,}9856\cdot 4 + 1{,}3456\cdot 6 + 0{,}0256\cdot 7 + \right.$$

$$\left. + 3{,}3856\cdot 5 + 14{,}7456\cdot 3\right) = \frac{1}{25}\cdot 109{,}3608 \approx 4{,}3744 ;$$

$$\sigma = \sqrt{\mu_2^*} = \sqrt{4{,}3744} \approx 2{,}0915 ;$$

$$\mu_3^* = \frac{1}{n}\sum_{i=1}^{l}\left(x_i - \bar{x}\right)^3 \cdot n_i = \frac{1}{25}\left((2-5{,}16)^3\cdot 4 + (4-5{,}16)^3\cdot 6 + (5-5{,}16)^3\cdot 7 + \right.$$

$$+(7-5{,}16)^3\cdot 5+(9-5{,}16)^3\cdot 3\Big)\approx\frac{1}{25}\big(-31{,}5545\cdot 4-1{,}5609\cdot 6-0{,}0041\cdot 7+$$

$$+6{,}1279\cdot 5+56{,}6231\cdot 3\big)=\frac{1}{25}\cdot 64{,}8967\approx 2{,}5959;$$

$$\mu_4^*=\frac{1}{n}\sum_{i=1}^{l}\left(x_i-\bar{x}\right)^4\cdot n_i=\frac{1}{25}\Big((2-5{,}16)^4\cdot 4+(4-5{,}16)^4\cdot 6+(5-5{,}16)^4\cdot 7+$$

$$+(7-5{,}16)^4\cdot 5+(9-5{,}16)^4\cdot 3\Big)=\frac{1}{25}\big(99{,}7122\cdot 4+1{,}8106\cdot 6+0{,}0006\cdot 7+$$

$$+11{,}2753\cdot 5+217{,}4327\cdot 3\big)=\frac{1}{25}\cdot 1118{,}3912\approx 44{,}7356.$$

Знайдемо коефіцієнти асиметрії та ексцесу за формулами (4.28) і (4.29) відповідно:

$$A_s=\frac{\mu_3^*}{\sigma^3}=\frac{2{,}5959}{(2{,}0915)^3}=\frac{2{,}5959}{9{,}149}\approx 0{,}2837;$$

$$E_k=\frac{\mu_4^*}{\sigma^4}-3=\frac{44{,}7356}{(2{,}0915)^4}-3\approx\frac{44{,}7356}{19{,}1351}-3\approx -0{,}6621.$$

Оскільки $A_s>0$, то розподіл має правосторонню асиметрію, тобто більшість значень зосереджена ліворуч від середнього. Оскільки $E_k<0$, то розподіл більш плосковершинний, ніж нормальний.

Відповідь: $A_s\approx 0{,}2837$; $E_k\approx -0{,}6621$.

Приклад 23. Нехай 50 випадково обраних стрільців беруть участь у змаганні. Кожен з них робить по 100 пострілів. У таблиці наведено кількість очок, отриманих кожним стрільцем:

75	54	67	62	88	55	65	46	68	73
62	46	64	55	68	72	58	73	88	62
58	73	54	62	65	68	55	62	77	79
55	83	64	53	72	68	79	46	64	73
62	46	54	72	55	62	65	79	72	88

Врахувавши, що випадкова величина X – кількість вибитих очок, необхідно: а) записати дискретний та інтервальний варіаційні ряди; б) побудувати гістограму, полігон частот; в) графік емпіричної функції розподілу групованої вибірки; г) знайти вибіркове середнє і вибіркове середнє квадратичне відхилення та коефіцієнт варіації V^*.

Розв'язання. а) Значення кількості очок у вибірці змінюється від 46 до 88, тобто випадкова величина X приймає 43 значення. Обсяг вибірки n=50 осіб.

Упорядкуємо дані вибірки за зростанням (ранжируємо вибірку) (табл. 4.2):

Таблиця 4.2

46	46	46	46	53	54	54	54	55	55
55	55	55	58	58	62	62	62	62	62
62	62	64	64	64	65	65	65	67	68
68	68	68	72	72	72	72	73	73	73
73	75	77	79	79	79	83	88	88	88

Записуємо дискретний варіаційний ряд (табл. 4.3):

Таблиця 4.3

x_i	46	53	54	55	58	62	64	65	67
n_i	4	1	3	5	2	7	3	3	1
x_i	68	72	73	75	77	79	83	88	
n_i	4	4	4	1	1	3	1	3	

Складемо інтервальний варіаційний ряд. Маємо $x_{\min}$=46, $x_{\max}$=88. Розмах вибірки: $R=x_{\max}-x_{\min}=88-46=42$.

Визначаємо кількість інтервалів k за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg(50) = 1 + 3{,}322 \cdot 1{,}699 \approx 1 + 5{,}644 \approx 6{,}644 .$$

Округлюємо k: k=7. Обчислюємо довжину інтервалу h за формулою (3.5): $h=R/k=42/7=6$.

Визначаємо межі інтервалів: [46;52), [52;58), [58;64), [64;70), [70;76), [76;82), [82;88].

Для кожного інтервалу, визначаємо: середини інтервалів x_i', частоти n_i, відносні частоти w_i, накопичені частоти n_i' та накопичені

частості w_i'. Складаємо таблицю 4.4.

Таблиця 4.4

Інтервали	Середини інтервалів $x_i' = \frac{x_i + x_{i+1}}{2}$	Частота n_i	Відносна частота $w_i = \frac{n_i}{n}$	Накопичена частота n_i'	Накопичена частость $w_i' = \frac{n_i'}{n}$
[46;52)	49	4	0,08	4	0,08
[52;58)	55	9	0,18	13	0,26
[58;64)	61	9	0,18	22	0,44
[64;70)	67	11	0,22	33	0,66
[70;76)	73	9	0,18	42	0,84
[76;82)	79	4	0,08	46	0,92
[82;88]	85	4	0,08	50	1

б) За результатами складеної таблиці будуємо гістограму частот. На горизонтальній осі X відкладаємо інтервали угруповання, на вертикальній осі Y – частоти n_i. Будуємо прямокутники, основи яких відповідають довжині інтервалу, а висота – частоті (рис. 4.1).

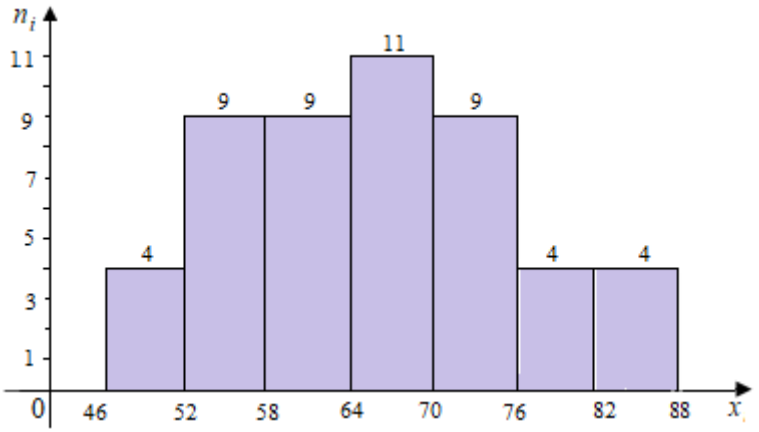


Рисунок 4.1

З'єднуючи відрізками ламаної середини верхніх основ прямокутників, з яких складається отримана гістограма, одержуємо полігон частот (рис. 4.2).

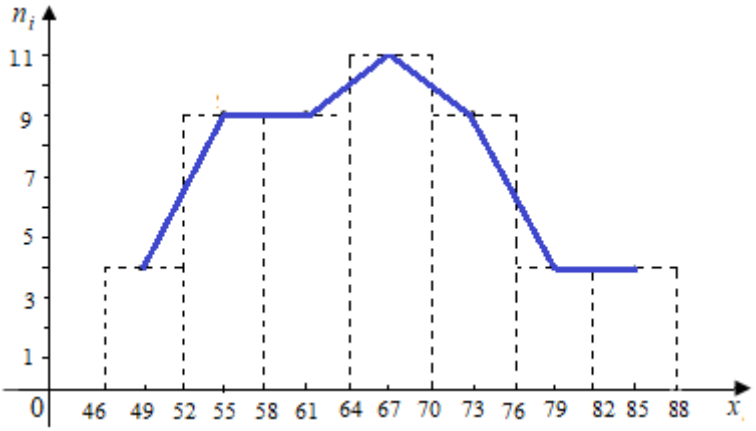


Рисунок 4.2

в) Знаючи середини інтервалів x_i' і накопичені частості w_i', знайдені для кожного інтервалу, емпірична функція розподілу визначається за формулою (3.11), як для дискретного варіаційного ряду. Емпірична функція розподілу матиме вигляд:

$$F^*(x)=\begin{cases} 0, & x\le 49,\\ 0{,}08, & 49<x\le 55,\\ 0{,}26, & 55<x\le 61,\\ 0{,}44, & 61<x\le 67,\\ 0{,}66, & 67<x\le 73,\\ 0{,}84, & 73<x\le 79,\\ 0{,}92, & 79<x\le 85,\\ 1{,}0, & x>85. \end{cases}$$

Графік емпіричної функції розподілу має вигляд (рис. 4.3):

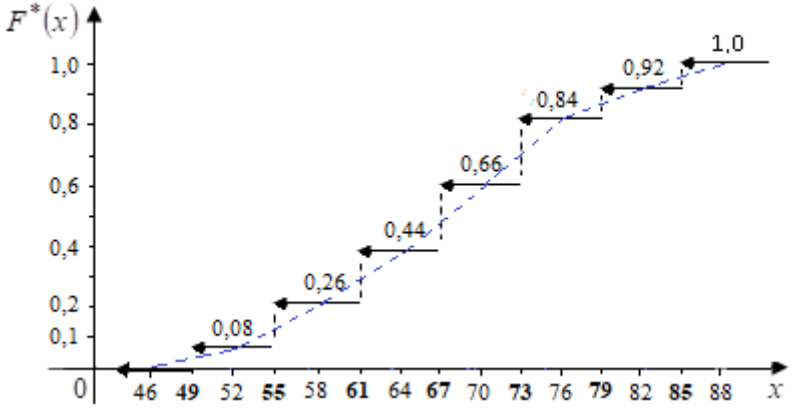


Рисунок 4.3

Оскільки інтервальний ряд дає дискретну інформацію про накопичені частоти лише у вузлових точках (межах інтервалів), для побудови графіку функцію $F^*(x)$ довизначають лінійно. Процес з'єднання точок $(x_i, F^*(x_i))$ відрізками перетворює східчасту функцію на ламану, яка утворює кумуляту – графічну модель накопичення ознаки (на графіку пунктирна лінія).

г) Оскільки маємо інтервальний варіаційний ряд, то застосовуємо для знаходження вибіркового середнього $\bar{x}_{\text{в}}$ формулу (4.8):

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i' \cdot n_i = \frac{1}{50}\cdot\left(49\cdot 4 + 55\cdot 9 + 61\cdot 9 + 67\cdot 11 + 73\cdot 9 + 79\cdot 4 + 85\cdot 4\right) =$$

$$= \frac{1}{50}\cdot\left(196 + 495 + 549 + 737 + 657 + 316 + 340\right) = \frac{3290}{50} = 65{,}8\,.$$

Для знаходження вибіркової дисперсії $D_{\text{в}}$ застосовуємо формулу (4.13):

$$D_{\text{в}} = \overline{x_{\text{в}}^{\,2}} - \left(\overline{x_{\text{в}}}\right)^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i'\right)^2 \cdot n_i - \left(\overline{x_{\text{в}}}\right)^2 = \frac{1}{50}\cdot\left(49^2\cdot 4 + 55^2\cdot 9 + 61^2\cdot 9 + \right.$$

$$\left. + 67^2\cdot 11 + 73^2\cdot 9 + 79^2\cdot 4 + 85^2\cdot 4\right) - \left(65{,}8\right)^2 = \frac{1}{50}\cdot\left(2401\cdot 4 + 3025\cdot 9 + \right.$$

$$\left. + 3721\cdot 9 + 4489\cdot 11 + 5329\cdot 9 + 6241\cdot 4 + 7225\cdot 4\right)\Big) - 4329{,}64 =$$

$$= \frac{1}{50}\cdot\left(9604 + 27225 + 33489 + 49379 + 47961 + 24964 + 28960\right) -$$

$$-4329{,}64 = \frac{221522}{50} - 4329{,}64 = 4430{,}44 - 4329{,}64 = 100{,}80\,.$$

Вибіркове середнє квадратичне відхилення $\sigma_{\text{в}}$ знаходимо за формулою (4.12):

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{100{,}80} \approx 10{,}0399\,.$$

Виправлену вибіркову дисперсію s^2 знаходимо за формулою (4.14):

$$s^2 = \frac{n}{n-1} D_{\text{в}} = \frac{50}{49} \cdot 100{,}80 \approx 102{,}8571\,.$$

Виправлене вибіркове середнє квадратичне відхилення s знаходимо за формулою (4.15):

$$s = \sqrt{s^2} = \sqrt{102{,}8571} \approx 10{,}1419\,.$$

Коефіцієнт варіації V^* знаходимо за формулою (4.16):

$$V^* = \frac{s}{\overline{x}_{\text{в}}} \cdot 100\,\% = \frac{10{,}1419}{65{,}8} \cdot 100\,\% \approx 15{,}41\,\%\,.$$

Оскільки коефіцієнт варіації V^* менший ніж 33%, то вибірка вважається однорідною, тобто розкид даних відносно невеликий, а середнє значення $\overline{x}_{\text{в}}$ добре характеризує всю сукупність.

Відповідь: $\overline{x}_{\text{в}} = 65{,}8$, $\sigma_{\text{в}} \approx 10{,}0399$, $V^* \approx 15{,}41\%$.

4.3 Спрощений спосіб розрахунку вибіркової середньої, дисперсії та вибіркових моментів

У тих випадках, коли експериментальні дані представлені великими числами, обчислення вибіркової середньої $\overline{x}$, дисперсії σ_x^2 та вибіркових моментів ускладнюється громіздкими операціями. Найбільш поширеним "спрощеним" способом розрахунку вибіркової середньої та дисперсії є метод моментів (або метод умовних варіант/відхилень) з використанням "умовного нуля" (або "робочої середньої"). Цей метод значно спрощує обчислення, особливо під час роботи з великими числами, оскільки при обчисленні вибіркової середньої $\overline{x}$ та дисперсії σ_x^2 варіаційного ряду використовуються не первісні варіанти x_i $\left(i = \overline{1,n}\right)$, а нові варіанти u_i.

Існує відмінність у розрахунку вибіркової середньої та дисперсії для дискретного варіаційного ряду та інтервального варіаційного ряду. Основна відмінність у тому, що для інтервального ряду перед розрахунком необхідно знайти середини інтервалів (x_i'), які у подальших розрахунках використовуються як варіанти.

Формула для розрахунку вибіркової середньої $\bar{x}$ за методом моментів має вигляд:

$$\bar{x} = c + \frac{\sum u_i \cdot n_i}{\sum n_i} \cdot h, \tag{4.30}$$

де $\bar{x}$ – шукана вибіркова середня; c – "умовний нуль" (або "робоча середня"). Вибирається як середина інтервалу (або варіанта x_i) з найбільшою частотою n_i (тобто моду); $u_i = \frac{x_i - c}{h}$ – умовні варіанти; x_i – варіанти (для дискретного варіаційного ряду) і середини інтервалів x_i' (для інтервального варіаційного ряду); n_i – частоти; h – величина (довжина) інтервалу (для інтервального ряду) або крок між варіантами (якщо він є постійним). Якщо кроку немає або ряд варіаційний, то h=1 і його можна не писати; $\sum n_i = n$ – загальний обсяг сукупності.

Зауваження 4.1. Якщо середніх інтервалів два (при парному числі інтервалів), то в якості c рекомендується взяти середину одного з цих інтервалів, наприклад, що має більшу частоту, а як h взяти найбільший дільник різниці (x_i–c) або таке число, яке дозволило позбутися дробів.

Розрахунок дисперсії σ_x^2 за методом моментів здійснюється за формулою:

$$\sigma_x^2 = h^2 \cdot \left(\frac{\sum u_i^2 \cdot n_i}{\sum n_i} - \left(\frac{\sum u_i \cdot n_i}{\sum n_i} \right)^2 \right), \tag{4.31}$$

де σ_x^2 – шукана дисперсія; u_i, n_i, h, $\sum n_i$ – ті самі значення, що і у формулі (4.30); $\sum u_i \cdot n_i$ – сума добутків умовних варіант на частоти

(використовується для розрахунку середнього); $\sum u_i^2 \cdot n_i$ – сума добутків квадратів умовних варіантів на частоти.

Відмінності у розрахунку для дискретного і інтервального варіаційних рядів наведені у таблиці 4.5.

Таблиця 4.5

Характеристика	Дискретний варіаційний ряд (з частотами)	Інтервальній варіаційний ряд
Вихідні дані	Варіанти x_i і частоти n_i	Інтервали $[x_i; x_{i+1})$ і частоти n_i
Підготовка даних	Не потрібно. Використовуються x_i безпосередньо	Необхідно знайти середини інтервалів $x_i' = \frac{x_i + x_{i+1}}{2}$
Крок h	h=1 (або крок між x_i, якщо він сталий)	Величина інтервалу: $h=x_{i+1}-x_i$ (при умові рівних інтервалів)
Вибір c	Варіанта x_i з найбільшою частотою n_i	Середина інтервалу x_i' з найбільшою частотою n_i

Також для дискретного варіаційного ряду можна застосовувати перехід до умовних варіант:

$$u_i=x_i-c. \tag{4.32}$$

Як c вибирають число, близьке до вибіркового середнього чи інше число на власний розсуд. Досить часто результати спостережень представлені у формі десяткових дробів із фіксованою кількістю знаків (k). Така розрядність даних вимагає відповідної точності при визначенні меж інтервалів та розрахунку кроку варіації. У цьому випадку зручно перейти до умовних варіантів виду:

$$u_i = \frac{x_i}{h}, \tag{4.33}$$

де $h = 10^{-k}$. Подібна заміна використовується і у випадку, коли дані мають наступний вигляд: 2000,4000,6000,8000 Прийнявши $h = 10^3$, отримаємо перетворені дані 2,4,6,8.

Для запропонованих перетворень значення вибіркової середньої та

дисперсії представлені в таблиці 4.6:

Таблиця 4.6

Заміна	Середнє, $\bar{x}$	Дисперсія, σ_x^2
$u_i = x_i - c$	$\bar{x} = \bar{u} + c$	$\sigma_x^2 = \sigma_u^2$
$u_i = \frac{x_i}{h}$	$\bar{x} = h \cdot \bar{u}$	$\sigma_x^2 = h^2 \cdot \sigma_u^2$
$u_i = \frac{x_i - c}{h}$	$\bar{x} = h \cdot \bar{u} + c$	$\sigma_x^2 = h^2 \cdot \sigma_u^2$

Значення $\bar{u}$ і σ_u^2 обчислюються за формулами:

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i , \qquad \sigma_u^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (u_i - \bar{u})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} u_i^2 - (\bar{u})^2 . \qquad (4.34)$$

Розрахунок коефіцієнтів асиметрії (A_s) і ексцесу (E_k) для дискретних та інтервальних рядів з використанням умовних варіантів (моментів), потребує обчислення центральних моментів до четвертого порядку. Цей метод спрощує обчислення, особливо при роботі з великими числами.

Для цього необхідно замінити вихідні варіанти x_i на умовні варіанти u_i. Вибір "умовного нуля" розглянутий вище.

1) Розрахунок умовних варіант u_i:

для дискретного варіаційного ряду: $u_i = x_i - c$;

для інтервального ряду (з рівними інтервалами h): $u_i = \frac{x_i - c}{h}$, де x_i – середина інтервалу x_i'. Це робить умовні варіанти цілими числами, що спрощує обчислення.

Умовний момент k-го порядку (μ_k') розраховується за формулою:

$$\mu_k' = \frac{\sum u_i^k \cdot n_i}{\sum n_i} , \qquad (4.35)$$

де u_i – умовні варіанти; n_i – частота варіанти x_i або частота інтервалу; $\sum n_i = n$ – загальний обсяг сукупності.

Для розрахунку асиметрії та ексцесу нам знадобляться моменти до

4-го порядку: μ_1', μ_2', μ_3', μ_4'.

Центральні момент *k*-го порядку μ_k^* – це моменти відносно вибіркової середньої $\bar{x}$. Вони розраховуються через умовні моменти μ_k':

центральний момент 1-го порядку $\mu_1^* = 0$ (за означенням),

центральний момент 2-го порядку $\mu_2^* = \mu_2' - (\mu_1')^2$ (дисперсія),

центральний момент 3-го порядку $\mu_3^* = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$,

центральний момент 4-го порядку $\mu_4^* = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu')^4$

Використовуючи розраховані центральні моменти, можна обчислити коефіцієнти асиметрії та ексцесу. У випадку використовування умовних варіант $u_i = \frac{x_i' - c}{h}$ отримані центральні моменти відносяться до u_i. Для отримання справжніх моментів (щодо x_i) необхідно домножити знайдені μ_k^* на h^k.

Коефіцієнт асиметрії (*асиметричності*) A_s, заснований на моментах, дорівнює:

$$A_s = \frac{\mu_3^*}{\sigma^3}, \tag{4.36}$$

де σ – середнє квадратичне відхилення, яке дорівнює $\sigma = \sqrt{\mu_2^*}$.

Якщо $A_s \approx 0$, то розподіл симетричний чи близький до нього. Якщо $A_s > 0$, то розподіл має правосторонню асиметрію. Якщо $A_s < 0$ то розподіл має лівосторонню асиметрію.

Коефіцієнт ексцесу E_k визначається за формулою:

$$E_k = \frac{\mu_4^*}{\sigma^4} - 3 = \frac{\mu_4^*}{(\mu_2^*)^2} - 3, \tag{4.37}$$

Коефіцієнт ексцесу показує гостроверхість або плосковершинність розподілу в порівнянні з нормальним: якщо $E_k > 0$, розподіл є гостроверхим; якщо $E_k < 0$, розподіл є плосковершинним; якщо $E_k \approx 0$, то

розподіл нормальний.

Зауваження 4.2. У випадку інтервального ряду для підвищення точності розрахунків коефіцієнтів дисперсії (μ_2^*) та ексцесу (μ_4^*) застосовуються поправки Шеппарда: $\mu_{2(\text{скор})}^* = \mu_2^* - \frac{h^2}{12}$, $\mu_{4(\text{скор})}^* = \mu_4^* - \frac{h^2}{2}\cdot\mu_2^* + \frac{7h^4}{240}$, де h – величина інтервалу. Ці поправки використовуються для подальшого розрахунку A_s та E_k, але практично часто опускаються, якщо інтервали досить малі.

Приклад 24. Протягом певного часу у взуттєвому магазині фіксувався розмір взуття, яке купували покупці, для визначення середнього розміру чоловічого взуття, на який потрібно орієнтуватися при оптових закупівлях. Поєднуючи однакові значення розміру взуття, отримали наступний згрупований варіаційний ряд:

x_i	38	39	40	41	42	43	44	45	46
n_i	1	4	8	12	13	12	6	3	1

Обчислити дисперсію розміру взуття, який має попит у населення.

Розв'язання. Скористайтеся формулою (4.32). Виберемо c=42 і розглянемо величину $U=X-c=X-42$. Запишимо для неї згрупований варіаційний ряд:

$u_i=x_i-c$	–4	–3	–2	–1	0	1	2	3	4
n_i	1	4	8	12	13	12	6	3	1

Знайдемо дисперсії σ_x^2 за формулою (4.31). Маємо h=1, n=60. Тоді

$$\sum_{i=1}^{9} u_i \cdot n_i = (-4)\cdot 1 + (-3)\cdot 4 + (-2)\cdot 8 + (-1)\cdot 12 + 0\cdot 13 + 1\cdot 12 + 2\cdot 6 +$$

$$+ 3\cdot 3 + 4\cdot 1 = -4 - 12 - 16 - 12 + 0 + 12 + 12 + 9 + 4 = -7,$$

$$\frac{1}{n}\cdot\sum_{i=1}^{9} u_i \cdot n_i = \frac{1}{60}\cdot(-7) = -\frac{7}{60}.$$

$$\sum_{i=1}^{9} u_i^2 \cdot n_i = (-4)^2\cdot 1 + (-3)^2\cdot 4 + (-2)^2\cdot 8 + (-1)^2\cdot 12 + 0\cdot 13 + 1^2\cdot 12 +$$

$$+2^2\cdot 6+3^2\cdot 3+4^2\cdot 1=16+36+32+12+0+12+24+27+4=163,$$

$$\frac{1}{n}\cdot\sum_{i=1}^{9}u_i^2\cdot n_i=\frac{1}{60}\cdot 163=\frac{163}{60}.$$

$$\sigma_x^2=\frac{163}{60}-\left(\frac{-7}{60}\right)^2=\frac{163}{60}-\frac{49}{3600}=\frac{9731}{3600}\approx 2{,}703.$$

Відповідь: $\sigma_x^2=2{,}703$.

Приклад 25. На підприємстві проаналізували дані про вік співробітників (у роках) та склали таблицю

Вікові групи (x_i)	[20;30)	[31;40)	[41;50)	[51;60)	[61;70]
Кількість працюючих n_i	4	10	15	7	4

Обчислити вибіркову середню, дисперсію, асиметрію та ексцес, використовуючи спрощений спосіб розрахунку.

Розв'язання. Загальний обсяг сукупності:

$$n=\sum_{i=1}^{5}n_i=4+10+15+7+4=40.$$

Величина інтервалу h=10. Знайдемо середини інтервалів за формулою $x_i'=\frac{x_i+x_{i+1}}{2}$: $x_1'=\frac{20+30}{2}=25$; $x_2'=\frac{31+40}{2}=35{,}5$;

$x_3'=\frac{41+50}{2}=45{,}5$; $x_4'=\frac{51+60}{2}=55{,}5$; $x_5'=\frac{61+70}{2}=65{,}5$.

Вибираємо "умовний нуль" c – середину інтервалу із найбільшою частотою (n_i=15), тобто c=45,5.

Розраховуємо умовні варіанти u_i за формулою $u_i=\frac{x_i'-c}{h}$:

$u_1=\frac{25-45{,}5}{10}=-2{,}05$; $u_2=\frac{35{,}5-45{,}5}{10}=-1{,}0$; $u_3=\frac{45{,}5-45{,}5}{10}=0$;

$u_4=\frac{55{,}5-45{,}5}{10}=1{,}0$; $u_1=\frac{65{,}5-45{,}5}{10}=2{,}0$.

Для зручності складемо розрахункову таблицю 4.7.

Таблиця 4.7

x_i'	n_i	u_i	$u_i n_i$	$u_i^2 \cdot n_i$	$u_i^3 \cdot n_i$	$u_i^4 \cdot n_i$
25	4	–2,05	–8,2	16,81	–34,4605	70,644
35,5	10	–1,0	–10	10	–10	10
45,5	15	0	0	0	0	0
55,5	7	1,0	7	7	7	7
65,5	4	2,0	8	16	32	64
Σ	40		–3,2	49,81	–5,4605	151,644

Знайдемо вибіркову середню $\bar{x}$ за формулою (4.30):

$$\bar{x} = 45{,}5 + \frac{-3{,}2}{40} \cdot 10 = 44{,}7 .$$

Дисперсію σ_x^2 знайдемо за формулою (4.31):

$$\sigma_x^2 = 100 \cdot \left(\frac{49{,}81}{40} - \left(\frac{-3{,}2}{40} \right)^2 \right) = 123{,}885 .$$

Знайдемо умовні моменти μ_1', μ_2', μ_3', μ_4' за формулою (4.35):

$$\mu_1' = \frac{-3{,}2}{40} = -0{,}08 ; \ \mu_2' = \frac{49{,}81}{40} = 1{,}24525 ; \ \mu_3' = \frac{-5{,}4605}{40} = 0{,}1365125 ;$$

$$\mu_4' = \frac{151{,}644}{40} = 3{,}7911 .$$

Для розрахунку центральних моментів μ_1^*, μ_2^*, μ_3^*, μ_4^* використовуємо формули перерахунку:

$$\mu_2^* = \mu_2' - (\mu_1')^2 = 1{,}23525 - (-0{,}08)^2 = 1{,}23525 - 0{,}0064 = 1{,}22885 ;$$

$$\mu_3^* = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0{,}1365125 - 3 \cdot 1{,}24525 \cdot (-0{,}08) + 2 \cdot (-0{,}08)^3 = 0{,}4343485 ;$$

$$\mu_4^* = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu')^4 = 3{,}7911 - 4 \cdot 0{,}1365125 \cdot (-0{,}08) + 6 \cdot 1{,}24525 \cdot (-0{,}08)^2 - 3 \cdot (-0{,}08)^4 = 3{,}88247872 .$$

Знайдемо коефіцієнти асиметрії A_s та ексцесу E_k.

Оскільки використовували умовні варіанти $u_i = \frac{x_i' - c}{h}$, то отримані центральні моменти відносяться до u_i. Для отримання справжніх моментів (щодо x_i) необхідно домножити μ_k^* на h^k:

Момент	Відносно u_i	Справжній μ_k^* (відносно x_i)
μ_2^*	1,22885	$1{,}22885 \cdot 10^2 = 122{,}885$
μ_3^*	0,4343485	$0{,}4343485 \cdot 10^3 = 434{,}3485$
μ_4^*	3,88247872	$3{,}88247872 \cdot 10^4 = 38824{,}7872$

Середнє квадратичне відхилення: $\sigma = \sqrt{\mu_2^*} = \sqrt{122{,}885} \approx 11{,}08535$.

Знайдемо коефіцієнти асиметрії та ексцесу

$$A_s = \frac{\mu_3^*}{\sigma^3} = \frac{434{,}3485}{(11{,}08535)^3} \approx \frac{434{,}3485}{1\ 362{,}2231} \approx 0{,}3189;$$

$$E_k = \frac{\mu_4^*}{\sigma^4} - 3 = \frac{38824{,}7872}{(11{,}08535)^4} - 3 \approx \frac{38824{,}7872}{15\ 100{,}7194} - 3 \approx -0{,}4289.$$

Оскільки $A_s > 0$, розподіл має невелику правосторонню асиметрію. Оскільки $E_k < 0$, розподіл є плосковершинним порівняно з нормальним.

Відповідь: $\bar{x} = 44{,}7$; $\sigma_x^2 = 123{,}885$; $A_s = 0{,}3189$; $E_k = -0{,}4289$.

4.4 Методи знаходження точкових оцінок

Точкові оцінки використовуються, коли ми повинні отримати одне конкретне значення для невідомого параметра. Це необхідно у таких випадках:

• *Прогнози та передбачення*: Якщо ви будуєте модель для прогнозування, наприклад майбутніх продажів, вам потрібно отримати точкове значення, щоб зробити конкретний прогноз.

• *Ухвалення рішень*: Коли рішення має ґрунтуватися на одному значенні, наприклад, "середня зарплата в компанії складає X".

• *Порівняння моделей*: Точкові оцінки параметрів використовуються для порівняння ефективності різних статистичних моделей.

Існують три основні методи отримання точкових оцінок параметрів генеральної сукупності: метод максимальної правдоподібності, метод моментів та метод найменших квадратів.

4.4.1 Метод максимальної правдоподібності

Метод максимальної правдоподібності (Maximum Likelihood Estimation, MLE) – це один із найпоширеніших підходів для оцінки параметрів статистичної моделі. Основна ідея полягає у пошуку таких значень параметрів, що максимізують функцію правдоподібності. Функція правдоподібності вимірює "ймовірність" спостереження наявних даних за певних значень параметрів.

Метод максимальної правдоподібності часто використовують для знаходження точкових оцінок параметрів заданого розподілу. Його алгоритм передбачає дослідження функції одного або кількох невідомих параметрів на максимум.

Розглянемо вибірку спостережень $x_1, x_2, \ldots, x_n$ випадкової величини X, закон розподілу якої містить невідомий параметр θ.

Означення. 4.23. *Функція правдоподібності* $L(x_1,x_2,\ldots,x_n;\theta)$ – це спільна щільність ймовірності (або ймовірність) отримання вибірки $x_1, x_2, \ldots, x_n$, що спостерігається, при даному параметрі θ.

Якщо випадкова величина X дискретна, то функція правдоподібності $L(x_1,x_2,\ldots,x_n;\theta)$ визначається формулою:

$$L(x_1, x_2, .., x_n; \theta) = p(x_1; \theta) \cdot p(x_2; \theta) \cdot \ldots \cdot p(x_n; \theta) = \\ = \prod_{i=1}^{n} p(x_i; \theta), \tag{4.38}$$

де $p(x_i, \theta)$ – ймовірність того, що в результаті випробування величина X прийме значення x_i.

Якщо випадкова величина X неперервна, то функція правдоподібності $L(x_1,x_2,\ldots,x_n;\theta)$ визначається формулою:

$$L(x_1, x_2, .., x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot .... \cdot f(x_n; \theta) =$$
$$= \prod_{i=1}^{n} f(x_i; \theta), \tag{4.39}$$

де $f(x_i, \theta)$ – щільності ймовірності кожного спостереження x_i.

За оцінку невідомого параметра θ приймають таке число $\theta^*=\theta^*(x_1,x_2,\ldots,x_n)$, при якому функція правдоподібності $L(x_1,x_2,\ldots,x_n;\theta)$ аргументу θ досягає максимуму.

Зауваження 4.3. Для спрощення математичних викладок замість прямого дослідження функції L на максимум аналізують її логарифм $\ln L$. Такий підхід є цілком коректним, адже обидві функції досягають екстремального значення за одного й того самого параметра θ, проте робота з логарифмом суттєво знижує складність обчислень.

Наприклад, логарифмуємо функцію (4.39):

$$\ln L(x_1, x_2, .., x_n; \theta) = \ln \left(f(x_1; \theta) \cdot f(x_2; \theta) \cdot .... \cdot f(x_n; \theta) \right) =$$
$$= \sum_{i=1}^{n} \ln f(x_i; \theta). \tag{4.40}$$

Щоб знайти значення θ^*, яке максимізує функцію (4.40), робимо наступне:

а) знаходимо похідну від $\ln L(x_1,x_2,\ldots,x_n;\theta)$ по θ:

$$\frac{d \ln L(x_1, x_2, .., x_n; \theta)}{d\theta};$$

б) прирівнюємо її до нуля:

$$\frac{d \ln L(x_1, x_2, .., x_n; \theta)}{d\theta} = 0 \tag{4.41}$$

і розв'язуємо отримане рівняння, корінь якого є θ^*. Рівняння (4.41) називають *рівнянням правдоподібності*;

в) знаходимо другу похідну $\frac{d^2 \ln L(x_1, x_2, .., x_n; \theta)}{d\theta^2}$. Якщо вона при

$\theta=\theta^*$ від'ємна, то θ^* – точка максимуму.

Приклад 26. Знайти методом максимальної правдоподібності оцінку параметра λ експоненціального розподілу.

Розв'язання. Нехай $x_1, x_2, \ldots, x_n$ – вибірка спостережень випадкової величини X, яка має експоненціальний розподіл з невідомим параметром λ, тобто $P(X = x) = \lambda \mathrm{e}^{-\lambda x}$, $x > 0$.

Складемо функцію правдоподібності, врахувавши, що $\theta=\lambda$,:

$$L(x_1, x_2, .., x_n; \lambda) = p(x_1; \lambda) \cdot p(x_2; \lambda) \cdot \ldots \cdot p(x_n; \lambda) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda x_i} =$$

$$= \lambda^n \cdot \mathrm{e}^{-\lambda \sum_{i=1}^{n} x_i}.$$

Знайдемо логарифмічну функцію правдоподібності:

$$\ln L(x_1, x_2, .., x_n; \lambda) = \ln\left(\lambda^n\right) + \ln \mathrm{e}^{-\lambda \sum_{i=1}^{n} x_i} = n \cdot \ln \lambda - \lambda \sum_{i=1}^{n} x_i .$$

Диференціюємо $\ln L(x_1, x_2, \ldots, x_n; \lambda)$ по λ і прирівнюємо до нуля:

$$\frac{d \ln L(x_1, x_2, .., x_n; \lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i , \qquad \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \frac{n}{\lambda} = \sum_{i=1}^{n} x_i .$$

Корінь рівняння $\lambda^* = \dfrac{n}{\sum\limits_{i=1}^{n} x_i} = \dfrac{1}{\bar{x}}$, де $\bar{x}$ – вибіркове середнє.

Знайдемо другу похідну $\dfrac{d^2 \ln L(x_1, x_2, .., x_n; \lambda)}{d\lambda^2}$ і перевіримо її знак при

$\lambda=\lambda^*$. Матимемо: $\dfrac{d^2 \ln L(x_1, x_2, .., x_n; \lambda)}{d\lambda^2} = -\dfrac{n}{\lambda^2}$.

При $\lambda^* = \dfrac{1}{\bar{x}}$ друга похідна від'ємна, отже, ця точка є точка максимуму і її треба прийняти як оцінку максимальної правдоподібності невідомого параметра λ експоненціального розподілу.

Відповідь: $\lambda^* = \dfrac{1}{\bar{x}}$.

4.4.2 Метод моментів

Метод моментів (Method of Moments) – це один із найстаріших і найбільш інтуїтивних підходів до оцінки параметрів. Він заснований на принципі, що теоретичні моменти розподілу дорівнюють вибірковим моментам, розрахованим за даними.

Нехай відомий закон розподілу випадкової величини X, що містить невідомі параметри θ_1, θ_2, …, θ_k. Маємо вибірку даних x_1, x_2, …, x_n випадкової величини X обсягу n. За методом моментів необхідно k теоретичних моментів прирівняти до k перших вибіркових моментів випадкової величини X, знаючи, що k-й теоретичний момент $M(X^k)$ – це математичне сподівання X^k, а k-й вибірковий момент – це $\nu_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k$. Матимемо систему

$$\begin{cases} M(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \\ M(X^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \quad . \quad . \quad . \\ M(X^k) = \frac{1}{n}\sum_{i=1}^{n} x_i^k \end{cases} \quad \text{або} \quad \begin{cases} M(X) = \bar{x}, \\ D(X) = D_{\text{в}}, \\ \quad . \quad . \quad . \\ M(X - (M(X))^k) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k. \end{cases} \tag{4.42}$$

З отриманої системи рівнянь знаходимо оцінки θ_1^*, θ_2^*, …, θ_k^* параметрів θ_1, θ_2, …, θ_k.

Згідно з методом моментів, для оцінки єдиного невідомого параметра розподілу достатньо прирівняти перший теоретичний момент (математичне сподівання $M(X)$) до відповідного вибіркового моменту першого порядку ν_1^*. Матимемо

$$M(X) = \bar{x}_{\text{в}}. \tag{4.43}$$

Оскільки математичне сподівання безпосередньо залежить від параметрів розподілу, розв'язання рівняння (4.43) дозволяє знайти шукану точкову оцінку.

У ситуації, коли закон розподілу визначається двома невідомими параметрами, метод моментів передбачає побудову системи з двох рівнянь. Для цього два теоретичні моменти (зазвичай математичне

сподівання $M(X)$ та дисперсію $D(X)$) прирівнюють до їхніх вибіркових аналогів відповідного порядку – вибіркового середнього ν_1^* ($\bar{x}_в$) та вибіркової дисперсії ν_2^* ($D_в$):

$$\begin{cases} M(X) = \bar{x}_в, \\ D(X) = D_в. \end{cases} \tag{4.44}$$

Оскільки ліві частини сформованих рівнянь є теоретичними функціями шуканих параметрів, розв'язання системи (4.44) дозволяє обчислити конкретні значення оцінок.

Приклад 27. Випадкова величина X рівномірно розподілена. Отримана вибірка даних x_i випадкової величини X обсягу n=100:

x_i	3	7	11	15	19
n_i	21	16	15	26	22

Знайти оцінку параметрів a і b рівномірного розподілу методом моментів.

Розв'язання. Враховуючи, що випадкова величина X підпорядковується закону рівномірного розподілу, імовірність її появи є сталою на всьому проміжку [a, b]. Це дозволяє використовувати числові характеристики визначаються за формулами: $M(X) = \dfrac{a+b}{2}$ – математичне сподівання, $D(X) = \dfrac{(b-a)^2}{12}$ – дисперсія. Згідно методу моментів необхідно розв'язати систему:

$$\begin{cases} M(X) = \bar{x}_в, \\ D(X) = D_в. \end{cases}$$

Знайдемо $\bar{x}_в$ і $D_в$ за формулами (4.4) та (4.10) відповідно.

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{1}{100} \cdot \left(3 \cdot 21 + 7 \cdot 16 + 11 \cdot 15 + 15 \cdot 26 + 19 \cdot 22\right) = \frac{1}{100} \cdot 1148 =$$

$$= 11{,}48.$$

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 \cdot n_i = \frac{1}{100} \cdot \Big((3 - 11{,}48)^2 \cdot 21 + (7 - 11{,}48)^2 \cdot 16 +$$

$$+(11-11{,}48)^2\cdot 15+(15-11{,}48)^2\cdot 26+(19-11{,}48)^2\cdot 22\Big)=\frac{1}{100}\cdot 3400{,}96=$$

$$=34{,}0096.$$

Підставляємо в систему та розв'язуємо її:

$$\begin{cases}\dfrac{a+b}{2}=11{,}48,\\ \dfrac{(b-a)^2}{12}=34{,}0096.\end{cases}\Rightarrow\begin{cases}a+b=22{,}96,\\ a\cdot b=29{,}7616.\end{cases}\Rightarrow\quad a=1{,}3791;\quad b=21{,}5809.$$

Відповідь: оцінки параметрів: a=1,3791; b=21,5809.

4.5 Завдання для самостійної роботи

Завдання	Відповідь
1. За вибіркою обсягу n=50 знайдено зміщену оцінку $D_{\text{в}}=9{,}8$ теоретичної дисперсії. Знайти виправлену оцінку дисперсії генеральної сукупності.	$s^2=10$.
2. За семестр студентами групи з предмета «Вища математика» було отримано такі бали: 46,49,61,63,74,76,83,86,91,93. Знайти середній бал за семестр для цієї групи студентів.	$\bar{x}_{\text{в}}=72{,}2$.
3. В результаті шести вимірів деякої фізичної величини одним приладом отримані такі результати: 25, 23, 21, 26, 22, 23. Знайти незміщені оцінки генеральної середньої та генеральної дисперсії вимірювань.	$\bar{x}_{\text{г}}=23{,}333$, $s^2=3{,}467$.
4. Знайти основні характеристики заданої вибірки: x_i: 2, 4, 5, 6 n_i: 8, 9, 10, 3	$\bar{x}_{\text{в}}=4$, $D_{\text{в}}=1{,}8$, $s^2=1{,}86$.
5. Знайти асиметрію та ексцес для заданого статистичного ряду: x_i: 2,6; 3; 3,4; 3,8; 4,2 n_i: 8; 20; 45; 15; 12	$A_s=0{,}1352$, $E_k=-0{,}337$.

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>6. Знайти точкові оцінки параметрів генеральної сукупності для інтервального ряду:
<table>
<tr><td>x_i</td><td>$[2;5)$</td><td>$[5;8)$</td><td>$[8;11)$</td><td>$[11;14]$</td></tr>
<tr><td>n_i</td><td>9</td><td>10</td><td>25</td><td>6</td></tr>
</table></td><td>$\bar{x}_{\text{в}} = 3{,}91$,
$D_{\text{в}} = 59{,}24$,
$D_{\text{г}} \approx s^2 = 60{,}45$,
$\sigma_{\text{г}} \approx 7{,}77$.</td></tr>
<tr><td>7. З генеральної сукупності вилучено вибірку, подану у вигляді дискретного ряду:
<table>
<tr><td>x_i</td><td>6</td><td>8</td><td>10</td><td>12</td><td>15</td><td>18</td><td>20</td></tr>
<tr><td>n_i</td><td>5</td><td>7</td><td>10</td><td>14</td><td>10</td><td>8</td><td>6</td></tr>
</table>
Знайти виправлене середньоквадратичне відхилення, медіану і моду.</td><td>$s \approx 4{,}23$, $M_{\text{e}}^{*} = 12$,
$M_{\text{o}}^{*} = 12$.</td></tr>
<tr><td>8. Знайти вибіркове середнє, незміщену вибіркову дисперсію та незміщене вибіркове середнє квадратичне відхилення для статистичного ряду:
<table>
<tr><td>x_i</td><td>2</td><td>7</td><td>9</td><td>10</td></tr>
<tr><td>n_i</td><td>8</td><td>14</td><td>10</td><td>18</td></tr>
</table></td><td>$\bar{x}_{\text{в}} = 7{,}68$,
$s^2 = 7{,}73$,
$s \approx 2{,}78$.</td></tr>
<tr><td>9. Знайти моду і медіану за даними інтервального ряду:
<table>
<tr><td>x_i</td><td>[14;16)</td><td>[16;18)</td><td>[18;20)</td><td>[20;22)</td><td>[22;24]</td></tr>
<tr><td>n_i</td><td>2</td><td>6</td><td>10</td><td>4</td><td>3</td></tr>
</table></td><td>$M_{\text{o}}^{*} = 18{,}33$,
$M_{\text{e}}^{*} = 18{,}9$</td></tr>
<tr><td>10. Знайти методом моментів оцінку параметра p (імовірності) геометричного розподілу $P(X = x_i) = (1-p)^{x_i - 1}\, p$, де $x_1, x_2, \ldots, x_n$ – кількість випробувань до появи події; p – імовірність появи події в одному випробуванні.</td><td>$p^{*} = \frac{1}{\bar{x}}$</td></tr>
<tr><td>11. За даними про кількість проданих товарів за день (x_i) та відповідну кількість днів (n_i):
<table>
<tr><td>x_i</td><td>10</td><td>12</td><td>14</td><td>16</td><td>18</td></tr>
<tr><td>n_i</td><td>2</td><td>5</td><td>12</td><td>7</td><td>4</td></tr>
</table>
з використанням умовних варіант знайти коефіцієнти асиметрії та ексцесу.</td><td>$A_s = -0{,}083$,
$E_k = -0{,}471$.</td></tr>
</table>

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>12. Ряд розподілу заробітної плати (у тис.грн.) робітників деякого цеху підприємства наведено у таблиці:
<table>
<tr><td>x_i</td><td>[21,2;21,4)</td><td>[21,4;21,6)</td><td>[21,6;21,8)</td></tr>
<tr><td>n_i</td><td>7</td><td>12</td><td>12</td></tr>
</table>
<table>
<tr><td>x_i</td><td>[21,8;22,0)</td><td>[22,0;22,2]</td></tr>
<tr><td>n_i</td><td>9</td><td>5</td></tr>
</table>
Потрібно обчислити коефіцієнт варіації V^*.</td><td>$V^* \approx 1{,}15\%$.</td></tr>
<tr><td>13. Знайти методом максимальної правдоподібності оцінку параметра λ розподілу Пуассона: $P(X = x_i) = \frac{\lambda^{x_i}}{x_i!}\mathrm{e}^{-\lambda}$, де x_1, x_2, …, x_n – вибірка спостережень випадкової величини X.</td><td>$\lambda^* = \frac{1}{n}\sum_{i=1}^{n} x_i$</td></tr>
<tr><td>14. Розподіл працівників за стажем (у роках) задано інтервальним рядом:
<table>
<tr><td>Інтервал стажу</td><td>1–3</td><td>3–5</td><td>5–7</td><td>7–9</td><td>9–11</td></tr>
<tr><td>n_i</td><td>5</td><td>12</td><td>18</td><td>10</td><td>5</td></tr>
</table>
Застосовуючи спрощений спосіб розрахунку знайти вибіркову середню та дисперсію.</td><td>$\bar{x} = 5{,}92$;
$\sigma_{\text{x}}^2 = 4{,}9536$.</td></tr>
<tr><td>15. Застосовуючи спрощений спосіб розрахунку знайти середній бал та дисперсію успішності студентів за даними у таблиці:
<table>
<tr><td>Бал, x_i</td><td>12</td><td>13</td><td>14</td><td>15</td><td>16</td></tr>
<tr><td>Кількість студентів, n_i</td><td>3</td><td>7</td><td>15</td><td>8</td><td>2</td></tr>
</table></td><td>$\bar{x} = 13{,}97$;
$\sigma_{\text{x}}^2 = 0{,}999$.</td></tr>
</table>

5 Інтервальні оцінки параметрів. Довірчі інтервали

Точкові оцінки невідомого параметра θ зручні як початкові результати обробки спостережень. Точкова оцінка дає нам одне число, яке, швидше за все, не співпаде із справжнім значенням параметра генеральної сукупності через випадковість вибірки. Вона не дає нам

уявлення про те, наскільки ця оцінка точна. Для вирішення цієї проблеми використовують інтервальні оцінки.

Інтервальні оцінки використовуються, коли:

1) Необхідно оцінити точність оцінки. Це головна перевага інтервальних оцінок. Вони дають уявлення про те, наскільки надійна точкова оцінка.

2) Коли важлива точність та надійність: У наукових дослідженнях, медичних випробуваннях, контролі якості, де помилки можуть мати серйозні наслідки. Наприклад, оцінка ефективності нових ліків, де важливо знати як середній ефект, так і діапазон, у якому цей ефект, швидше за все, перебуває.

3) При малих та середніх вибірках: Якщо обсяг вибірки невеликий, точкова оцінка може сильно відрізнятися від справжнього значення параметра. Інтервальна оцінка у цьому випадку дає більш реалістичне уявлення про можливий діапазон істинного значення.

4) Для розуміння діапазону можливих значень параметра. Інтервальні оцінки показують, у яких межах, імовірніше, є справжнє значення, що особливо важливо, якщо розкид даних великий.

5) Під час публікації наукових результатів. У наукових статтях та звітах, як правило, потрібні інтервальні оцінки, щоб показати надійність та статистичну значущість отриманих результатів.

Для інтервальних оцінок важливі такі вимоги: заданий рівень довіри, мінімальна ширина інтервалу при заданому рівні довіри (ширина інтервалу пропорційна точності), спроможність (як і для точкових оцінок), правильне покриття.

Переваги інтервальних оцінок: Інтервальні оцінки більш інформативні, оскільки вони надають діапазон значень та асоційовану довірчу ймовірність того, що справжній параметр знаходиться в цьому діапазоні. Це дозволяє оцінити ризик та невизначеність.

Нехай статистична характеристика $\theta^*(x_1,x_2,\ldots,x_n)$, знайдена за даними вибірки $(x_1,x_2,\ldots,x_n)$, служить оцінкою невідомого параметра θ. Вважатимемо θ постійним числом (θ може бути і випадковою величиною). Чим менше $|\theta-\theta^*|$, тим точніше θ^* визначає параметр θ, тобто $|\theta-\theta^*|<\delta$, де $\delta>0$. Чим менше δ, тим точніше оцінка.

Означення 5.1. Величину δ називають *точністю оцінки*.

Означення 5.2. *Інтервальна оцінка* – це інтервал, який з певною ймовірністю (довірчою ймовірністю) містить невідомий параметр θ генеральної сукупності.

Оскільки $|\theta-\theta^*|<\delta$, то $-\delta<\theta-\theta^*<\delta$. Звідси маємо $\theta^*-\delta<\theta<\theta^*+\delta$ або $\theta \in \left(\theta^* - \delta; \theta^* + \delta\right)$.

Означення 5.3. *Надійністю оцінки* (*довірчою ймовірністю*, *коефіцієнтом довіри*) називається ймовірність γ, з якою виконується нерівність $|\theta-\theta^*|<\delta$, тобто $P(|\theta-\theta^*|<\delta)=\gamma$ або

$$P\left(\theta^* - \delta < \theta < \theta^* + \delta\right) = \gamma \text{ , } 0<\gamma<1. \qquad (5.1)$$

Зазвичай надійність γ вибирають заздалегідь. Типові значення: 0,9 (90%), 0,95 (95%), 0,99 (99%). Чим ближче γ до одиниці, тим надійніше оцінка.

Означення 5.4. *Довірчим інтервалом* є такий числовий проміжок $(\theta^*-\delta;\theta^*+\delta)$, у межі якого з заданою надійністю (ймовірністю) γ потрапляє справжнє значення невідомого параметра θ.

Означення 5.5. Величини $\theta^*-\delta$ та $\theta^*+\delta$ називаються *довірчими межами*: $\theta^*-\delta$ – нижня межа, $\theta^*+\delta$ – верхня межа.

Треба мати на увазі, що для різних вибірок однієї і тієї ж генеральної сукупності можуть виходити різні довірчі інтервали.

Значення меж довірчого інтервалу залежить від закону розподілу параметра θ^*, тобто межі різні для різних розподілів. Довірчий інтервал має випадковий характер і за розташуванням (відносно θ^*), і за шириною (оскільки межі залежать від даних вибірки), отже його межі є випадковими величинами. Тому прийнято говорити не про ймовірність попадання параметра θ в деякий побудований довірчий інтервал, а про те, що побудований довірчий інтервал покриє параметр θ з надійністю γ.

Означення 5.6. *Рівнем значущості* α називають ймовірність, з якою значення параметра не потрапляє у довірчий інтервал і визначається $\alpha=1-\gamma$.

Зазвичай α задається як 0,1; 0,05 або 0,01.
Наприклад, для надійності 95% рівень значущості α=0,05.

5.1 Деякі статистичні розподіли

Генеральні сукупності часто мають нормальний закон розподілу. У цьому випадку багато з вибіркових характеристик, у тому числі $\overline{x}_{\text{в}}$, $D_{\text{в}}$, s^2, виражаються через невелику кількість спеціальних розподілів. Як правило, у математичній статистиці використовуються не щільність цих розподілів, а деякі числові характеристики, представлені таблицями. Найчастіше в якості такої характеристики виступає або квантиль розподілу, або критична точка розподілу.

Означення 5.7. *Квантилем рівня p* ($0<p<1$) або *p-квантилем випадкової величини X* називається таке число d_p, що ймовірність $p=P(X<d_p)$ дорівнює заданій величині *p*.

Означення 5.8. *Критичною точкою розподілу випадкової величини X* (для правосторонньої критичної області) називається таке число $x_{\text{кр}}(\alpha)$, що ймовірність $\alpha=P(X<x_{\text{кр}}(\alpha))$ дорівнює заданій величині α-рівню значущості.

Розглянемо кілька розподілів, яким підпорядковуються вибіркові характеристики та які використовують для побудови інтервальних оцінок.

Розподіл Пірсона (χ^2–хі квадрат)

Нехай X_i $\left(i=\overline{1,n}\right)$ незалежні, нормовані, нормально розподілені випадкові величини з параметрами $N(0;1)$, тобто $M(X_i)=0$, $D(X_i)=1$ для $i=\overline{1,n}$.

Означення 5.9. *Розподілом* χ^2 (*хі-квадрат*) *Пірсона* із числом степенів свободи *n* називається закон розподілу випадкової величини, що являє собою суму квадратів суми *n* незалежних стандартних нормальних величин, тобто

$$\chi^2 = \sum_{i=1}^{n} X_i{}^2 , \tag{5.2}$$

де $X_i \in N(0;1)$, $i = \overline{1,n}$.

Зауваження 5.1. Кількість степенів свободи n є єдиним параметром χ^2-розподілу і значення χ^2 невід'ємні, тобто $P(\chi^2<0)=0$.

χ^2**-**розподіл з n степенями свободи має числові характеристики $M(\chi^2)=n$, $D(\chi^2)=2n$. Для великих n розподіл випадкової величини χ^2 близький до нормального розподілу з параметрами $a=n$, $\sigma^2=2n$. Якщо $\chi_1{}^2$ і $\chi_2{}^2$ – незалежні випадкові величини з степенями свободи n і m відповідно, то їх сума $\chi_1{}^2+\chi_2{}^2$ також має χ^2-розподіл з $(n+m)$ степенями свободи.

Розподіл Стьюдента (*t*-розподіл)

Нехай $X \in N(0;1)$ – випадкова величина, що має стандартний нормальний розподіл з параметрами $a=0$, $\sigma=1$, а V – незалежна від $N(0;1)$ випадкова величина, яка має χ^2-розподіл з n степенями свободи.

Означення 5.10. *Розподілом Стьюдента* (*t*-розподілом) називається розподіл випадкової величини

$$T_n = \frac{X}{\sqrt{V/k}} \tag{5.3}$$

з $k=n-1$ степенями свободи.

При малих значеннях n розподіл Стьюдента помітно відрізняється від стандартного нормального розподілу, однак при $n>30$ ці розподіли близькі.

Розподіл Стьюдента залежить від одного параметра – числа степенів свободи k.

Розподіл Стьюдента відіграє важливу роль в математичній статистиці, а саме, якщо випадкові величини X_i $\left(i = \overline{1,n}\right)$ незалежні і однаково розподілені за нормальним законом $N(a;\sigma)$, то величина

$$T_{n-1} = \frac{\left(\overline{X} - a\right)\sqrt{n}}{s}, \tag{5.4}$$

де $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ підпорядковується розподілу Стьюдента з $k=n-1$ степенями свободи.

Розподіл Фішера-Снедекора (*F*-розподіл)

Припустимо, що $\chi^2(k_1)$ і $\chi^2(k_2)$ – це незалежні випадкові величини, з відповідними числами степенів свободи k_1 та k_2.

Означення 5.11. *Розподілом Фішера-Снедекору* (*F*-розподілом) називають закон розподілу величини, що визначається як відношення цих хі-квадрат величин, кожна з яких поділена на своє число степенів свободи:

$$F_{k_1,k_2} = \frac{\frac{1}{k_1}\cdot\chi^2(k_1)}{\frac{1}{k_2}\cdot\chi^2(k_2)} \tag{5.5}$$

Оскільки випадкові величини $\chi^2(k_1)\geq 0$ і $\chi^2(k_2)\geq 0$, то $F_{k1,k2}\geq 0$.

Для кожного з наведених вище розподілів існують спеціальні таблиці, наведені у Додатку А.

5.2 Побудова довірчих інтервалів для різних параметрів

Формули для довірчих інтервалів різняться залежно від того, який параметр генеральної сукупності оцінюємо, чи відоме його стандартне відхилення, і який обсяг вибірки.

Припустимо, що спостерігається випадкова величина X, яка розподілена за нормальним законом $N(a;\sigma)$ з параметрами $a=M(X)$ – математичне сподівання і $\sigma=\sigma(X)$ – середнє квадратичне відхилення. Найліпшою незміщеною точковою оцінкою для математичного

сподівання a є вибіркова середня $\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, знайдена за вибіркою $(x_1, x_2, \ldots, x_n)$ обсягу n. Розглянемо, як будуються довірчі інтервали для нормальної сукупності (на прикладах).

Довірчий інтервал для математичного сподівання *a* при відомій дисперсії

Розглянемо випадкову величину $\frac{(\bar{x}_{\text{в}} - a)\sqrt{n}}{\sigma}$, яка розподілена за законом $N(0;1)$. Виберемо число t_γ, щоб

$$P\left(-t_\gamma < \frac{(\bar{x}_{\text{в}} - a)\sqrt{n}}{\sigma} < t_\gamma\right) = \gamma, \tag{5.6}$$

де n – обсяг вибірки, γ – задана надійність.

Значення t_γ знаходиться з використанням інтегральної функції Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}} dz$, оскільки

$$P(-t_\gamma < N(0;1) < t_\gamma) = \Phi(t_\gamma) - \Phi(-t_\gamma) = 2\Phi(t_\gamma) = \gamma, \tag{5.7}$$

і задовольняє нелінійному рівнянню $\Phi(t_\gamma)=\gamma/2$ та визначається по заданій надійності γ за таблицею А.2 (див. Додаток А).

Оскільки $\sigma>0$, то $-t_\gamma < \frac{(\bar{x}_{\text{в}} - a)\sqrt{n}}{\sigma} < t_\gamma$ і довірчий інтервал має вигляд:

$$\bar{x}_{\text{в}} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x}_{\text{в}} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}}. \tag{5.8}$$

Отриманий довірчий інтервал симетричний щодо $\bar{x}_{\text{в}}$.

Формула (5.8) не відрізняється для довірчих інтервалів при $n\geq 30$ і $n<30$.

Означення 5.12. *Точністю оцінки* називається величина

$$\delta_\gamma = t_\gamma \cdot \frac{\sigma}{\sqrt{n}}. \qquad (5.9)$$

Інтервальна оцінка залежить від трьох параметрів: надійності γ, точності оцінки δ_γ і обсягу вибірки n.

З аналізу формули (5.8) можна зробити такі висновки:

а) при збільшенні обсягу вибірки n точність інтервальної оцінки збільшується, оскільки величина δ_γ зменшується. При великих n найліпшою оцінкою для a стає $\bar{x}_\text{в}$, тобто точкова оцінка;

б) в силу того, що функція $\Phi(x)$ є неспадною, при збільшенні надійності γ зростає величина δ_γ, тобто зменшується точність (інтервал стає ширшим);

в) для фіксованих значень надійності $\gamma=1-\alpha$ та точності δ_γ з формули (5.9) можна визначити необхідний обсяг вибірки, що забезпечує задане значення $\gamma=1-\alpha$ та δ_γ.

Слід пам'ятати, що при незмінному обсягу вибірки одночасно збільшувати точність і надійність оцінки не можна.

Означення 5.13. *Мінімальний обсяг вибірки n*, що гарантує оцінку математичного сподівання a із заданою надійністю γ і точністю оцінки δ_γ, визначається з нерівності $\delta_\gamma \geq t_\gamma \cdot \frac{\sigma}{\sqrt{n}}$:

$$n \geq \left(\sigma \cdot \frac{t_\gamma}{\delta_\gamma} \right)^2. \qquad (5.10)$$

Довірчий інтервал для математичного сподівання *a* при невідомій дисперсії

Для обсягу вибірки $n<30$ розглянемо випадкову величину $\frac{\left(\bar{x}_\text{в} - a\right)\sqrt{n-1}}{\sqrt{D_\text{в}}}$, яка розподілена за законом Стьюдента T_{n-1}. При заданому значенні надійності γ, користуючись таблицею А.3 (див. Додаток А), обчислимо значення $t(\gamma, n-1)$ з умови

$$P\left(-t(\gamma,n-1)<\frac{(\bar{x}_{\text{в}}-a)\sqrt{n-1}}{\sqrt{D_{\text{в}}}}<t(\gamma,n-1)\right)=\gamma\,. \qquad (5.11)$$

Зауважимо, що k у таблиці А.3 означає число степенів свободи, тобто $k=n-1$.

З умови (5.11) отримуємо інтервальну оцінку надійності γ для невідомої генеральної середньої a:

$$\bar{x}_{\text{в}}-\frac{t(\gamma,n-1)\sqrt{D_{\text{в}}}}{\sqrt{n-1}}<a<\bar{x}_{\text{в}}+\frac{t(\gamma,n-1)\sqrt{D_{\text{в}}}}{\sqrt{n-1}}\,.$$

Виразимо межі інтервалу через виправлену дисперсію s^2. Оскільки $s^2=\frac{n}{n-1}D_{\text{в}}$, то $\frac{\sqrt{D_{\text{в}}}}{\sqrt{n-1}}=\frac{s}{\sqrt{n}}$.

Тоді довірчий інтервал для математичного сподівання a при невідомій дисперсії матиме вигляд:

$$\bar{x}_{\text{в}}-t(\gamma,n-1)\cdot\frac{s}{\sqrt{n}}<a<\bar{x}_{\text{в}}+t(\gamma,n-1)\cdot\frac{s}{\sqrt{n}}\,, \qquad (5.12)$$

а точність оцінки визначається співвідношенням

$$\delta=t(\gamma,n-1)\cdot\frac{s}{\sqrt{n}}\,. \qquad (5.13)$$

Також при побудові довірчого інтервалу для математичного сподівання a нормально розподіленої сукупності з невідомою дисперсією використовують t-розподіл Стьюдента. Цей розподіл залежить від одного параметра числа степенів свободи $k=n-1$. У цьому випадку формула (5.12) матиме вигляд:

$$\bar{x}_{\text{в}}-t_{\alpha/2;n-1}\cdot\frac{s}{\sqrt{n}}<a<\bar{x}_{\text{в}}+t_{\alpha/2;n-1}\cdot\frac{s}{\sqrt{n}}\,, \qquad (5.14)$$

де $t_{\alpha/2;n-1}$ – квантиль t-розподілу Стьюдента з $k=n-1$ степенями свободи, який відповідає значенню $\alpha/2$ (де $\alpha=1-\gamma$ – рівень значущості, а γ – надійність). Значення $t_{\alpha/2;n-1}$ знаходимо за таблицею А.4 (див. Додаток А).

Зауваження 5.2. Оскільки в таблиці А.4 – Критичні точки розподілу Стьюдента є позначки "Рівень значущості α (двостороння критична область)" та "Рівень значущості α (одностороння критична область)", то знаходимо $t_{\alpha/2;k}$ для односторонньої критичної області.

Для обсягу вибірки $n\geq 30$ розподіл Стьюдента несуттєво відрізняється від нормального, близький до $N(0;1)$, тому для оцінки математичного сподівання застосовують формулу (5.8), де замість σ використовують s і довірчий інтервал має вигляд:

$$\bar{x}_{\text{в}} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \bar{x}_{\text{в}} + t_\gamma \cdot \frac{s}{\sqrt{n}}, \qquad (5.15)$$

де n – обсяг вибірки, $s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_{\text{в}})^2}$ – виправлене середнє квадратичне відхилення, t_γ – значення аргументу функції Лапласа $\Phi(t_\gamma)$, при якому $\Phi(t_\gamma)=\gamma/2$, і визначається по заданій надійності γ за таблицею А.2 (див. Додаток А).

Довірчий інтервал для дисперсії σ^2 при відомому математичному сподіванні

Нехай випадкова величина X розподілена за нормальним законом $N(a;\sigma)$ з відомим параметром $a=M(X)$ – математичне сподівання, а дисперсія σ^2 невідома.

Для оцінки σ^2 вилучено вибірку $(x_1,x_2,\ldots,x_n)$ обсягу n. За точкову оцінку дисперсії $D(X)$ використовуємо вибіркову дисперсію $\sigma_{\text{в}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^2$.

При побудові довірчого інтервалу для дисперсії розглянемо статистику $\chi^2 = \frac{n \cdot \sigma_{\text{в}}^2}{\sigma^2}$, що має χ^2-розподіл із числом степенів свободи

n незалежно від значення параметра σ^2. Задамо необхідний рівень значущості $\alpha=1-\gamma$ (γ – надійність). Тоді, використовуючи таблицю критичних точок χ^2-розподілу, знаходимо критичні точки $\chi^2_{1-\alpha/2;n}$ і $\chi^2_{\alpha/2;n}$, для яких виконуватиметься рівність:

$$P\left(\chi^2_{1-\alpha/2;n} < \chi^2 < \chi^2_{\alpha/2;n}\right) = 1-\alpha\,, \tag{5.16}$$

Підставимо в (5.15) замість χ^2 значення $\dfrac{n\cdot\sigma_в^2}{\sigma^2}$. Матимемо

$$P\left(\chi^2_{1-\alpha/2;n} < \frac{n\cdot\sigma_в^2}{\sigma^2} < \chi^2_{\alpha/2;n}\right) = 1-\alpha\,.$$

Звідки довірчий інтервал для дисперсії σ^2

$$\frac{n\cdot\sigma_в^2}{\chi^2_{\alpha/2;n}} < \sigma^2 < \frac{n\cdot\sigma_в^2}{\chi^2_{1-\alpha/2;n}}\,, \tag{5.17}$$

де n – обсяг вибірки, $\chi^2_{1-\alpha/2;n}$ і $\chi^2_{\alpha/2;n}$ – критичні точки χ^2-розподілу при заданому рівні значущості α з n степенями свободи, які знаходимо за таблицею А.5 (див. Додаток А).

Довірчий інтервал для дисперсії σ^2 при невідомому математичному сподіванні

При побудові довірчого інтервалу для дисперсії розглянемо величину $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$, що має χ^2-розподіл із числом степенів свободи $k=n-1$ незалежно від значення параметра σ^2. Задамо необхідний рівень значущості $\alpha=1-\gamma$ (γ – надійність). Тоді, використовуючи таблицю критичних точок χ^2-розподілу, неважко вказати критичні точки $\chi^2_{1-\alpha/2;n-1}$ і $\chi^2_{\alpha/2;n-1}$, для яких виконуватиметься така рівність:

$$P\left(\chi^2_{1-\alpha/2;n-1} < \chi^2 < \chi^2_{\alpha/2;n-1}\right) = 1-\alpha\,, \tag{5.18}$$

Підставивши в (5.17) замість χ^2 значення $\frac{(n-1)s^2}{\sigma^2}$, отримаємо:

$$P\left(\chi^2_{1-\alpha/2;n-1} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2;n-1}\right) = 1-\alpha .$$

Звідки довірчий інтервал для дисперсії σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}, \qquad (5.19)$$

де n – обсяг вибірки, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2$ – виправлена дисперсія, $\chi^2_{1-\alpha/2;n-1}$ і $\chi^2_{\alpha/2;n-1}$ – критичні точки χ^2-розподілу при заданому рівні значущості α з $k=n-1$ степенями свободи, які знаходимо за таблицею А.5 (див. Додаток А).

Довірчий інтервал для середнього квадратичного відхилення σ

При обсягу вибірки $n \le 30$ добувши корінь з нерівності (5.19), отримаємо довірчий інтервал для середнього квадратичного відхилення σ:

$$s \cdot \sqrt{\frac{n-1}{\chi^2_{\alpha/2;n-1}}} < \sigma < s \cdot \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;n-1}}}, \qquad (5.20)$$

де n – обсяг вибірки, $s = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \bar{x}_в)^2}$ – виправлене середнє квадратичне відхилення, $\chi^2_{1-\alpha/2;n-1}$ і $\chi^2_{\alpha/2;n-1}$ – критичні точки χ^2-розподілу при заданому рівні значущості $\alpha=1-\gamma$ (γ – надійність) з $k=n-1$ степенями свободи, які знаходимо за таблицею А.5 (див. Додаток А).

Можна спростити формулу (5.20), якщо позначити:

$$\gamma_1 = \sqrt{\frac{n-1}{\chi^2_{\alpha/2;n-1}}} \text{ і } \gamma_2 = \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2;n-1}}} . \qquad (5.21)$$

Тоді γ_1 і γ_2 знаходимо за таблицею А.6 (див. Додаток А).

Довірчий інтервал для середнього квадратичного відхилення σ матиме вигляд:

$$\gamma_1 \cdot s < \sigma < \gamma_2 \cdot s\,. \tag{5.22}$$

У випадку великих вибірок (n>30) довірчий інтервал для середнього квадратичного відхилення σ можна визначити за формулою

$$\frac{s \cdot \sqrt{2n}}{\sqrt{2n-3}+t_\gamma} < \sigma < \frac{s \cdot \sqrt{2n}}{\sqrt{2n-3}-t_\gamma}, \tag{5.23}$$

де значення t_γ задовольняє нелінійному рівнянню $\Phi(t_\gamma)=\gamma/2$ та визначається по заданій надійності γ за таблицею А.2 (див. Додаток А).

Існує і інший спосіб визначення меж довірчого інтервалу середнього квадратичного відхилення σ, в основі якого лежить вибір довірчого інтервалу, симетричного щодо σ^2. Довірчий інтервал для середнього квадратичного відхилення σ у такому разі можна знайти, знаючи надійність γ і число степенів свободи n, за формулою

$$s \cdot (1-q) < \sigma < s \cdot (1+q), \tag{5.24}$$

де $q = q(\gamma, n) = \dfrac{\delta}{s}$ – деяке число, яке знаходимо за таблицею А.7 (див. Додаток А). Причому, якщо $1-q<0$, то інтервал має вигляд

$$0 < \sigma < s \cdot (1+q). \tag{5.25}$$

Довірчий інтервал для ймовірності успіху у схемі Бернуллі

Побудова довірчого інтервалу для ймовірності (частки, пропорції) – це важливе завдання в математичній статистиці, що використовується, коли потрібно оцінити справжню ймовірність p (частку ознаки в генеральній сукупності) на основі вибіркової частки $\overline{p}$.

Вибіркову частку $\overline{p}$ у формулі довірчого інтервалу для ймовірності (за методом нормального наближення) можна вважати відносною частотою *w* за формулою:

$$\overline{p} = w = \frac{m}{n}, \tag{5.26}$$

де $\overline{p}$ (вибіркова частка) – це оцінка істинної ймовірності *p* події з урахуванням наявної вибірки (результатів випробувань), *w* (відносна частота) – це відношення числа *m* настання події до загальної кількості *n* випробувань).

При побудові довірчого інтервалу, терміни (вибіркова частка та відносна частота) фактично позначають одне й те саме значення, розраховане за даними вибірки.

Нехай *n* – число незалежних випробувань, *m* – число настання події *A* в цих випробуваннях, *p* – ймовірність настання події *A* в кожному окремому випробуванні, причому ця ймовірність невідома.

Недолік точкової оцінки полягає в тому, що вона може виявитися далекою від істини (особливо, при малому *n*) і тому ймовірність *p* вигідно оцінити інтервалом:

$$\frac{m}{n} - \delta < p < \frac{m}{n} + \delta, \tag{5.27}$$

який із заздалегідь обраною надійністю γ накриє справжнє значення *p*, δ – точність оцінки.

Вищезазначене можна записати:

$$P\left(\left|\frac{m}{n} - p\right| < \delta\right) = \gamma.$$

Це ймовірність того, що відносна частота $w = \frac{m}{n}$ відхилиться від ймовірності *p* менш ніж на δ.

Якщо кількість випробувань *n* досить велика (близько сотні і більше) і значення *p* не надто мале (не дуже близько до нуля), то для побудови довірчого інтервалу для ймовірності застосовують *метод нормального наближення*. Цей метод заснований на наближенні

біномного розподілу нормальним розподілом при великому обсязі вибірки n.

Умови застосування нормального наближення.

Нормальний розподіл можна використовувати, якщо виконані умови, які забезпечують достатню схожість біноміального розподілу з нормальним:

1) Великий обсяг вибірки (n): чим більше n, то краще апроксимація.

2) Достатня кількість "успіхів" та "невдач": Часто використовують правило, що кількість "успіхів" $\left(n\cdot\overline{p}\right)$ та кількість "невдач" $\left(n\cdot\left(1-\overline{p}\right)\right)$ має бути не менше 10 (іноді 5).

Якщо ці умови виконуються, то вибіркова частка $\overline{p}$ приблизно розподілена за нормальним законом: $\overline{p}\approx N\left(p;np\left(1-p\right)\right)$. Саме це наближення дозволяє використовувати нормальний закон та стандартні критичні значення $t_{\gamma/2}$ для побудови інтервалу.

Довірчий інтервал для ймовірності p матиме вигляд

$$w-t_{\gamma/2}\cdot\sqrt{\frac{w\cdot\left(1-w\right)}{n}}<p<w+t_{\gamma/2}\cdot\sqrt{\frac{w\cdot\left(1-w\right)}{n}}\,, \tag{5.28}$$

де $w=\dfrac{m}{n}$ – відносна частота (вибіркова частка), n – обсяг вибірки, m – кількість "успіхів" (кількість варіант у вибірці, що мають певну ознаку); $1–w$ – вибіркова частка "невдач"; $t_{\gamma/2}$ – значення аргументу функції Лапласа $\Phi(t_{\gamma/2})$, при якому $\Phi(t_{\gamma/2})=\gamma/2$, і визначається по заданій надійності γ за таблицею А.2 (див. Додаток А).

Точність оцінки визначається співвідношенням

$$\delta=t_{\gamma/2}\cdot\sqrt{\frac{w\cdot\left(1-w\right)}{n}}\,. \tag{5.29}$$

Зауваження 5.3. Оскільки вибіркову частку $\overline{p}$ можна вважати відносною частотою w, то для використання формули (5.28) необхідно виконання умови: $\min\left(n\cdot w;n\cdot\left(1-w\right)\right)\geq 10$.

У невеликих вибірках нормальне наближення може бути не дуже точним. Іноді застосовують *поправку неперервності Йейтса* для покращення точності, особливо коли кількість "успіхів" або "невдач" мала (менше 10).

Поправка $\frac{1}{2n}$ додається до граничної помилки (точність оцінки) $\delta = t_{\gamma/2} \cdot \sqrt{\frac{w \cdot (1-w)}{n}}$. Наприклад, n=100, $\frac{1}{2n} = \frac{1}{200} = 0{,}005$.

Довірчий інтервал для ймовірності p із поправкою неперервності Йейтса матиме вигляд:

$$w - \left(t_{\gamma/2} \cdot \sqrt{\frac{w \cdot (1-w)}{n}} + \frac{1}{2n} \right) < p < w + \left(t_{\gamma/2} \cdot \sqrt{\frac{w \cdot (1-w)}{n}} + \frac{1}{2n} \right), \quad (5.30)$$

Метод нормального наближення може давати неточні результати, особливо коли w близько до 0 або 1, або коли кількість випробувань n мала. У цьому випадку для побудови довірчого інтервалу для ймовірності застосовують *інтервал Агрісті-Коулл*, який є модифікацією, що дає більш надійні результати у цих випадках.

Ідея полягає в тому, щоб додати фіктивну кількість "успіхів" та "невдач" до даних. Зазвичай для надійності 95% при побудові довірчого інтервалу додають 2 "успіхи" і 2 "невдачі" (всього 4 спостереження).

Спочатку розраховують "скоригований" обсяг вибірки $\tilde{n}$, "скориговану" кількість "успіхів" $\tilde{m}$ та "скориговану" вибіркову частку $\tilde{w}$. Нехай $t_{\gamma/2}$ – критичне значення (для надійності 95% це 1,96).

Скориговані значення:

$$\tilde{n} = n + \left(t_{\gamma/2}\right)^2, \qquad \tilde{m} = m + 0{,}5 \cdot \left(t_{\gamma/2}\right)^2, \qquad \tilde{w} = \frac{\tilde{m}}{\tilde{n}}. \quad (5.31)$$

Довірчий інтервал для ймовірності p матиме вигляд

$$\tilde{w} - t_{\gamma/2} \cdot \sqrt{\frac{\tilde{w} \cdot (1-\tilde{w})}{\tilde{n}}} < p < \tilde{w} + t_{\gamma/2} \cdot \sqrt{\frac{\tilde{w} \cdot (1-\tilde{w})}{\tilde{n}}}, \quad (5.32)$$

Приклад 28. Випадкова величина X розподілена нормально з невідомим математичним сподіванням a і відомою дисперсією σ^2=25.

За вибіркою ($x_1,x_2,\ldots,x_{100}$) обсягу n=100 обчислено вибіркове середнє $\bar{x}_{\text{в}} = \frac{1}{100}\cdot\sum_{i=1}^{100} x_i = 142{,}3$. Визначити довірчий інтервал для математичного сподівання a з надійністю γ=0,95.

Розв'язання. За умовою прикладу дисперсія σ^2 відома. Знайдемо довірчий інтервал для математичного сподівання a за формулою (5.8), в яку підставимо $\sigma = \sqrt{25} = 5$, n=100, а також $t_\gamma = t_{0,95} = 1{,}96$, знайдене за таблицею А.2 (див. Додаток А). Тоді

$$142{,}3 - 1{,}96\cdot\frac{5}{\sqrt{100}} < a < 142{,}3 + 1{,}96\cdot\frac{5}{\sqrt{100}},$$

тобто $141{,}32 < a < 143{,}28$.

Відповідь: з надійністю γ=0,95 невідоме математичне сподівання $a \in (141{,}32;\ 143{,}28)$.

Приклад 29. Випадкова величина X розподілена нормально з невідомим математичним сподіванням a і невідомою дисперсією σ^2. За вибіркою ($x_1,x_2,\ldots,x_{20}$) знайдені оцінки: $\bar{x}_{\text{в}} = \frac{1}{20}\sum_{i=1}^{20} x_i = 56{,}85$ і $s^2 = \frac{1}{19}\sum_{i=1}^{20}\left(x_i - \bar{x}_{\text{в}}\right)^2 = 26{,}1$. Знайти довірчі інтервали для математичного сподівання a і дисперсії σ^2 при надійності γ=0,95.

Розв'язання. Оскільки a і σ^2 невідомі і n=20<30, знайдемо довірчий інтервал для математичного сподівання a за формулою (5.12), в яку підставимо $\bar{x}_{\text{в}} = 56{,}85$, n=20, $s = \sqrt{26{,}1} \approx 5{,}109$, а також для надійності γ=0,95, користуючись таблицею А.3 (див. Додаток А), обчислимо значення $t(0{,}95;19)$=2,093, оскільки $n-1$=19. Тоді

$$56{,}85 - 2{,}093\cdot\frac{5{,}109}{\sqrt{20}} < a < 56{,}85 + 2{,}093\cdot\frac{5{,}109}{\sqrt{20}},$$

тобто $54{,}459 < a < 59{,}241$.

Довірчий інтервал для σ^2 знайдемо за формулою (5.19), в яку підставимо n=20, s^2=26,1, а також знайдені за таблицею А.5 (див. Додаток А) значення, $\chi^2_{\alpha/2;n-1}$ та $\chi^2_{1-\alpha/2;n-1}$. Маємо $\alpha = 1-\gamma = 1-0{,}95 = 0{,}05$;

$\alpha/2=0{,}05/2=0{,}025$; $1-\alpha/2=1-0{,}025=0{,}975$; $k=n-1=20-1=19$. Отже, $\chi^2_{0{,}025;19}=32{,}852$, $\chi^2_{0{,}975;19}=8{,}907$.

Таким чином, маємо

$$\frac{20-1}{32{,}862}\cdot 26{,}1<\sigma^2<\frac{20-1}{8{,}907}\cdot 26{,}1,$$

тобто $15{,}09<\sigma^2<55{,}675$.

Відповідь: з надійністю $\gamma=0{,}95$ довірчі інтервали для математичного сподівання $a\in(54{,}459;\ 59{,}241)$ і дисперсії $\sigma^2\in(15{,}09;\ 55{,}675)$.

Приклад 30. За вибіркою обсягу $n=18$ нормально розподіленої генеральної сукупності обчислено значення вибіркової дисперсії $\sigma^2_{\text{в}}=1{,}4$. Побудувати довірчий інтервал для параметра σ^2 з надійністю $\gamma=0{,}99$.

Розв'язання. Довірчий інтервал для дисперсії σ^2 знайдемо за формулою (5.17). Рівень значущості $\alpha=1-\gamma=1-0{,}99=0{,}01$. Тоді $\alpha/2=0{,}005$; $1-\alpha/2=0{,}995$. Критичні точки χ^2-розподілу $\chi^2_{0{,}005;18}$ і $\chi^2_{0{,}995;18}$ знаходимо за таблицею А.5 (див. Додаток А). Матимемо $\chi^2_{0{,}005;18}=37{,}156$ і $\chi^2_{0{,}995;18}=6{,}265$. Тоді довірчий інтервал для дисперсії σ^2 можемо записати у вигляді:

$$\frac{18\cdot 1{,}4}{37{,}156}<\sigma^2<\frac{18\cdot 1{,}4}{6{,}265}\ \text{або}\ 0{,}6782<\sigma^2<4{,}0223.$$

Відповідь: $\sigma^2\in(0{,}6782;\ 4{,}0223)$.

Приклад 31. Серед 35 новонароджених виявилося 18 хлопчиків. Знайти довірчий інтервал для ймовірності p народження хлопчика при надійності $\gamma=0{,}99$.

Розв'язання. За умовою прикладу $n=35$, $m=18$, $\gamma=0{,}99$. Тоді $w=\frac{m}{n}=\frac{18}{35}\approx 0{,}5143$, $1-w=0{,}4857$. Перевіримо виконання умови: $\min\left(n\cdot w;n\cdot(1-w)\right)\geq 10$. Маємо $\min(18;17)\geq 10$,тобто $17>10$. Довірчий інтервал для p знайдемо за формулою (5.28), в яку підставимо $n=35$,

m=18, γ=0,99, w=0,5143, 1−w=0,4857, а з таблиці А.2 (див. Додаток А) матимемо $t_{0,495} \approx 2{,}575$. Отримаємо

$$0{,}5143 - 2{,}575 \cdot \sqrt{\frac{0{,}5143 \cdot 0{,}4857}{35}} < p < 0{,}5143 + 2{,}575 \cdot \sqrt{\frac{0{,}5143 \cdot 0{,}4857}{35}},$$

тобто 0,2368<p<0,7318.

Відповідь: з надійністю γ=0,99 довірчий інтервал для ймовірності $p \in (0{,}2368\,;\, 0{,}7318)$.

Приклад 32. Правління ОСББ бажає на основі вибірки оцінити середні внески на управління будинком для трикімнатних квартир з надійністю щонайменше 99% і похибкою, меншою 120 грн. Передбачається, що внески на управління будинком мають нормальний розподіл із середнім квадратичним відхиленням, що не перевищує 400 грн. Необхідно знайти мінімальний обсяг вибірки.

Розв'язання. Для знаходження мінімального обсягу вибірки n, застосуємо формулу (5.10). За умовою задачі надійність γ=0,99, похибка δ_γ=120, середнє квадратичне відхилення σ=400. З таблиці А.2 (див. Додаток А) матимемо $t_{0,495}$≈2,575.Тоді

$$n = \left(\sigma \cdot \frac{t_\gamma}{\delta_\gamma}\right)^2 = \left(400 \cdot \frac{2{,}575}{120}\right)^2 \approx (8{,}583)^2 \approx 63{,}67\,.$$

Оскільки зі зростанням γ та зменшенням δ_γ зростає n, то n≥63,67 і $n_{\min}$=64 (зауважимо, що при зменшенні верхньої межі буде зменшуватися і $n_{\min}$).

Відповідь: $n_{\min}$=64 квартири.

Приклад 33. Під час перевірки 120 деталей з великої партії виявлено 15 бракованих деталей. З надійністю 95% знайти довірчий інтервал частки бракованих деталей у всій партії. Який мінімальний обсяг вибірки слід взяти для того, щоб із надійністю 95% стверджувати, що частка бракованих деталей у всій партії відрізняється від частоти появи бракованих деталей у вибірці трохи більше, ніж 1%.

Розв'язання. За умовою прикладу обсяг вибірки n=120, бракованих деталей m=15, надійність γ=0,95, а γ/2=0,475. Довірчий інтервал частки бракованих деталей знайдемо за формулою (5.28), оскільки кількість випробувань n досить велика.

Відносна частота $w=\frac{m}{n}=\frac{15}{120}=0{,}125$, а $1-w=0{,}875$. За таблицею А.2 (див. Додаток А) матимемо $t_{0{,}475}=1{,}96$. Тоді

$$0{,}125-1{,}96\cdot\sqrt{\frac{0{,}125\cdot 0{,}875}{120}}<p<0{,}125+1{,}96\cdot\sqrt{\frac{0{,}125\cdot 0{,}875}{120}},$$

тобто $0{,}0658<p<0{,}1842$.

Для знаходження мінімального обсягу n вибірки подаємо довірчий інтервал (5.28) у вигляді нерівності $|w-p|<\delta$, тобто $|w-p|<t_{\gamma/2}\cdot\sqrt{\frac{w\cdot(1-w)}{n}}$, яка виконується з ймовірністю $\gamma=0{,}95$. Оскільки за умовою задачі $|w-p|\le 0{,}01$, то для визначення n отримаємо нерівність $t_{\gamma/2}\cdot\sqrt{\frac{w\cdot(1-w)}{n}}\le 0{,}01$. Звідси випливає, що $1{,}96\cdot\sqrt{\frac{0{,}125\cdot 0{,}875}{n}}\le 0{,}01$. З нерівності матимемо $n\ge 4201{,}75$.

Отже, мінімальний обсяг вибірки $n_{\min}=4202$.

Відповідь: $p\in(0{,}0658;\,0{,}1842)$, $n_{\min}=4202$.

Приклад 34. При опитуванні 100 осіб виявили, що 30 з них користуються певним мобільним додатком. Побудувати 95% довірчий інтервал для справжньої частки користувачів цим мобільним додатком та перевірити застосуванням поправки неперервності Йейтса.

Розв'язання. За умовою прикладу обсяг вибірки $n=100$, користувачів мобільним додатком $m=30$, надійність $\gamma=0{,}95$. Довірчий інтервал частки користувачів знайдемо за формулою (5.28), оскільки кількість випробувань n досить велика.

Вибіркова частка (відносна частота) $w=\frac{m}{n}=\frac{30}{100}=0{,}3$, а $1-w=0{,}7$.

За таблицею А.2 (див. Додаток А) критичне значення $t_{0{,}475}=1{,}96$. Гранична помилка (точність оцінки) вибіркової частки

$$\delta=t_{\alpha/2}\cdot\sqrt{\frac{w\cdot(1-w)}{n}}=1{,}96\cdot\sqrt{\frac{0{,}3\cdot 0{,}7}{100}}=1{,}96\cdot\sqrt{\frac{0{,}21}{100}}=1{,}96\cdot\sqrt{0{,}0021}\approx$$

$$\approx 1{,}96\cdot 0{,}0458\approx 0{,}0898.$$

Тоді довірчий інтервал $w-\delta<p<w+\delta$ матиме вигляд:

$0{,}3-0{,}0898<p<0{,}3+0{,}0898$ або $0{,}2102<p<0{,}3898$.

Знайдемо довірчий інтервал, застосовуючи поправку неперервності Йейтса. Оскільки $n=100$, то поправка $\frac{1}{2n}=\frac{1}{200}=0{,}005$.

Додаємо поправку до граничної помилки (точність оцінки) δ: $\delta+\frac{1}{2n}=0{,}0898+0{,}005=0{,}0948$. Застосуємо формулу (5.30) для знаходження довірчого інтервалу, яку записуємо у вигляді:

$$w-\left(\delta+\frac{1}{2n}\right)<p<w+\left(\delta+\frac{1}{2n}\right).$$

Матимемо $0{,}3-0{,}0948<p<0{,}3+0{,}0948$ або $0{,}2052<p<0{,}3948$.

Відповідь: $p\in(0{,}2102;0{,}3898)$, тобто з 95%-ю впевненістю можна стверджувати, що справжня частка користувачів цим мобільним додатком у генеральній сукупності знаходиться в діапазоні від 21,02% до 38,98%. Застосовуючи поправку неперервності Йейтса отримали ширший інтервал $p\in(0{,}2052;0{,}3948)$.

Приклад 35. У ході вибіркового контролю перевірено 9 деталей. Вважаючи, що помилки виготовлення підпорядковуються нормальному закону розподілу, при відомому вибірковому середньому квадратичному відхиленні $\sigma_в=5$, знайти 95% довірчий інтервал для невідомого параметра σ.

Розв'язання. *1 спосіб*. За умовою обсяг вибірки $n=9$, надійність $\gamma=0{,}95$. Тоді рівень значущості $\alpha=1-\gamma=0{,}05$. Оскільки $n<30$ і відомо вибіркове середнє квадратичне відхилення $\sigma_в=5$, то застосуємо формулу (5.20), замінивши в ній s на $\sigma_в$:

$$\sigma_в\cdot\sqrt{\frac{n}{\chi^2_{\alpha/2;n-1}}}<\sigma<\sigma_в\cdot\sqrt{\frac{n}{\chi^2_{1-\alpha/2;n-1}}}.$$

Число степенів свободи $k=n-1=8$. Знайдемо значення ймовірностей для $\alpha=0{,}05$ $P\left(\chi^2>\chi^2_{\alpha/2;n-1}\right)=\frac{\alpha}{2}=\frac{0{,}05}{2}=0{,}025$ і $P\left(\chi^2>\chi^2_{1-\alpha/2;n-1}\right)=1-\frac{\alpha}{2}=1-\frac{0{,}05}{2}=0{,}975$. За таблицею А.5 (див. Додаток А) значення критичних точок χ^2-розподілу

$\chi^2_{1-\alpha/2;n-1} = \chi^2_{0,975;8} = 2{,}18$ і $\chi^2_{\alpha/2;n-1} = \chi^2_{0,025;8} = 17{,}535$.

Довірчий інтервал для невідомого параметра σ матиме вигляд

$$5\cdot\sqrt{\frac{9}{17{,}535}} < \sigma < 5\cdot\sqrt{\frac{9}{2{,}18}} \text{ або } 3{,}5821<\sigma<10{,}1593.$$

2 спосіб. Знайдемо виправлену дисперсію $s^2 = \frac{n}{n-1}\cdot\sigma_в^2 = \frac{9}{8}\cdot 5^2 = 28{,}125$, звідси $s\approx 5{,}3033$.

Для знаходження довірчого інтервалу для середнього квадратичного відхилення σ застосуємо формулу (5.22). За таблицею А.6 (див. Додаток А) знайдемо γ_1 і γ_2 за надійністю γ=0,95 та числу степенів свободи k=8: нижня межа γ_1=0,675 і верхня межа γ_2=1,916. Помноживши отримані значення γ_1 і γ_2 на s, знайдемо довірчий інтервал для σ:

$0{,}675\cdot 5{,}3033 < \sigma < 1{,}916\cdot 5{,}3033$ або $3{,}5797 < \sigma < 10{,}1611$.

3 спосіб. Для знаходження довірчого інтервалу для середнього квадратичного відхилення σ застосуємо формулу (5.24). Виправлене середнє квадратичне відхилення s≈5,3033. За таблицею А.7 (див. Додаток А) знайдемо значення $q=q(\gamma,n)$ за надійністю γ=0,95 та числу степенів свободи n=9: q=0,71. Підставимо знайдені значення q=0,71 та s=5,3033 у формулу (5.24). Довірчий інтервал для σ матиме вигляд:

$5{,}3033\cdot(1-0{,}71) < \sigma < 5{,}3033\cdot(1+0{,}71)$ або $1{,}538 < \sigma < 9{,}069$.

Як видно з обчислень, величина довірчого інтервалу залежить від способу його побудови та дає близькі між собою, але неоднакові результати.

Відповідь: 1) $\sigma\in(3{,}5821; 10{,}1593)$, 2) $\sigma\in(3{,}5797; 10{,}1611)$, 3) $\sigma\in(1{,}538; 9{,}069)$.

Приклад 36. На підставі вибіркових спостережень продуктивності праці 15 робітниць кондитерської фабрики було встановлено, що середнє квадратичне відхилення добового виробітку становить 9 коробок цукерок на годину. Припускаючи, що продуктивність праці робітниці має нормальний розподіл, знайти з надійністю 0,9 довірчі інтервали для генеральної дисперсії та середнє квадратичного відхилення добового виробітку робітниць.

Розв'язання. За умовою обсяг вибірки n=15, середнє квадратичне відхилення s=9, надійність γ=0,9. Тоді число степенів свободи $k=n-1=14$, а рівень значущості α=0,1.

Довірчий інтервал для дисперсії σ^2 знайдемо за формулою (5.19). Критичні точки χ^2-розподілу при заданому рівні значущості α=0,1 з k=14 степенями свободи знаходимо за таблицею А.5 (див. Додаток А). Матимемо $\chi^2_{0,05;14} = 23{,}685$ і $\chi^2_{0,95;14} = 6{,}57$. Тоді довірчий інтервал для дисперсії σ^2 можемо записати у вигляді:

$$\frac{14 \cdot 9^2}{23{,}685} < \sigma^2 < \frac{14 \cdot 9^2}{6{,}57} \text{ або } 47{,}8784 < \sigma^2 < 172{,}6027.$$

Довірчий інтервал для середнього квадратичного відхилення σ матиме вигляд:

$$\sqrt{47{,}8784} < \sigma < \sqrt{172{,}6027} \text{ або } 6{,}9194 < \sigma < 13{,}1378.$$

Відповідь: $\sigma^2 \in (47{,}8784;\ 172{,}6027)$, $\sigma \in (6{,}9194;\ 13{,}1378)$.

Приклад 37. При обробці експериментальних даних обсягу n=100 отримане вибіркове середнє квадратичне відхилення $\sigma_в$=2,3237. У припущенні про нормальний розподіл генеральної сукупності, з надійністю γ=0,9 визначити довірчий інтервал для оцінки генерального середнього квадратичного відхилення σ.

Розв'язання. Знайдемо виправлене середнє квадратичне відхилення $s = \sqrt{\frac{n}{n-1}}\,\sigma_в = \sqrt{\frac{100}{99}} \cdot 2{,}3237 \approx 2{,}3354$.

Оскільки n=100>30, то довірчий інтервал для оцінки генерального середнього квадратичного відхилення σ знайдемо за формулою (5.23). Значення t_γ визначається по заданій надійності γ за таблицею А.2 (див. Додаток А). Матимемо t_γ=1,64. Тоді

$$\frac{2{,}3354 \cdot \sqrt{200}}{\sqrt{197} + 1{,}64} < \sigma < \frac{2{,}3354 \cdot \sqrt{200}}{\sqrt{197} - 1{,}64}.$$

Шуканий інтервал 2,1069<σ<2,6644.

Відповідь: $\sigma \in (2{,}1069;\ 2{,}6644)$.

Приклад 38. Перевіривши 20 виробів, виявили, що 15 виробів першого сорту. Припускаючи, що має місце біноміальний розподіл,

побудувати інтервальну оцінку для ймовірності p першосортних виробів з надійністю 0,95.

Розв'язання. За умовою обсяг вибірки n=20, кількість виробів першого сорту в вибірці m=15, надійність γ=0,95. Оскільки обсяг вибірки n малий, то для знаходження довірчого інтервалу для ймовірності p першосортних виробів застосуємо інтервал Агрісті-Коулл.

Визначимо скориговані значення для обсягу n, кількості виробів першого сорту m та відносної частоти появи першосортного виробу $w = \frac{m}{n}$. Для надійності γ=0,95 критичне значення $t_{\gamma/2}$=1,96.

Скориговані значення за формулами (5.31):

$$\tilde{n} = n + \left(t_{\gamma/2}\right)^2 = 20 + (1{,}96)^2 = 23{,}8416,$$

$$\tilde{m} = m + 0{,}5 \cdot \left(t_{\gamma/2}\right)^2 = 15 + 0{,}5 \cdot (1{,}96)^2 = 16{,}9208,$$

$$\tilde{w} = \frac{\tilde{m}}{\tilde{n}} = \frac{16{,}9208}{23{,}8416} \approx 0{,}7097.$$

Довірчий інтервал для ймовірності p за формулою (5.31) матиме вигляд:

$$0{,}7097 - 1{,}96 \cdot \sqrt{\frac{0{,}7097 \cdot (1 - 0{,}7097)}{23{,}8416}} < p < 0{,}7097 + 1{,}96 \cdot \sqrt{\frac{0{,}7097 \cdot (1 - 0{,}7097)}{23{,}8416}}$$

або 0,5275<p<0,8919.

Відповідь: $p \in (0{,}5275;\ 0{,}8919)$.

Приклад 39. Виробник морозива хоче оцінити середній обсяг порції, що продається у кіосках. Він взяв випадкову вибірку з n=15 порцій і отримав такі обсяги (мл): 205, 198, 202, 200, 199, 201, 203, 197, 204, 200, 201, 196, 200, 202, 199. Побудувати 90% довірчий інтервал для справжнього середнього обсягу порції a.

Розв'язання. За умовою обсяг вибірки n=15, надійність γ=0,90.

Обчислюємо вибіркове середнє

$$\bar{x}_{\text{в}} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{15} \cdot (205 + 198 + 202 + 200 + 199 + 201 + 203 + 197 +$$

$$+ 204 + 200 + 201 + 196 + 200 + 202 + 199) = \frac{1}{15} \cdot 3007 \approx 200{,}47 \text{ (мл)}.$$

Знайдемо незміщену вибіркову дисперсію s^2 та стандартне відхилення s:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2 = \frac{1}{15-1}\cdot\Big(\left(205-200{,}47\right)^2 + \left(198-200{,}47\right)^2 +$$
$$+\left(202-200{,}47\right)^2 + \left(200-200{,}47\right)^2 + \left(199-200{,}47\right)^2 + \left(201-200{,}47\right)^2 +$$
$$+\left(203-200{,}47\right)^2 + \left(197-200{,}47\right)^2 + \left(204-200{,}47\right)^2 + \left(200-200{,}47\right)^2 +$$
$$+\left(201-200{,}47\right)^2 + \left(196-200{,}47\right)^2 + \left(200-200{,}47\right)^2 + \left(202-200{,}47\right)^2 +$$
$$+\left(199-200{,}47\right)^2\Big) \approx \frac{1}{14}\cdot 87{,}733 \approx 6{,}267.$$

Тоді $s = \sqrt{6{,}267} \approx 2{,}503\,(\text{мл})$.

Рівень значущості $\alpha = 1-\gamma = 0{,}1$ і $\alpha/2 = 0{,}05$, а число степенів свободи $k = n-1 = 14$. Користуючись таблицею А.4 – Критичні точки t-розподілу Стьюдента (див. Додаток А), обчислимо значення $t_{0{,}05;14} = 1{,}761$, застосовуючи "Рівень значущості α (одностороння критична область)". Підставимо знайдені значення в формулу (5.14). Матимемо

$$200{,}47 - 1{,}761\cdot\frac{2{,}503}{\sqrt{15}} < a < 200{,}47 + 1{,}761\cdot\frac{2{,}503}{\sqrt{15}}$$

або $199{,}332 < a < 201{,}608$.

Відповідь: з надійністю $\gamma = 0{,}90$ можемо стверджувати, що справжній середній розмір порції морозива коливається від 199,332мл до 201,608 мл.

5.3 Завдання для самостійної роботи

Завдання	Відповідь
1. Випадкова величина X розподілена нормально з невідомим математичним сподіванням a і відомою дисперсією $\sigma^2 = 1{,}88$. За вибіркою обсягу $n = 20$ обчислено вибіркове середнє $\bar{x}_в = 4{,}25$. Визначити довірчий інтервал для математичного сподівання a з довірчою ймовірністю $\gamma = 0{,}95$.	$a \in \left(3{,}65;\, 4{,}85\right)$.
Завдання	Відповідь

2. Зроблено 20 дослідів над випадковою величиною X, розподіленою за нормальним законом. Побудувати довірчі інтервали для математичного сподівання та дисперсії, що відповідають надійності 0,98, якщо знайдені оцінки $\bar{x}_{\text{в}}=10,78$ і $s^2=0,064$.	$a \in (10,637; 10,924)$, $\sigma^2 \in (0,0336; 0,1594)$.
3. Знайти довірчий інтервал для математичного сподівання (середнього) діаметра валу з надійністю 99%, якщо $n=9$, $\bar{x}_{\text{в}}=30$ мм і $s^2=9$ мм^2.	$a \in (26,64; 33,36)$.
4. По $n=10$ рівноточним вимірам знайдено виправлене середнє квадратичне відхилення $s=0,76$. Припускаючи, що результати вимірювань розподілені нормально, побудувати довірчий інтервал для оцінки справжнього значення σ (генерального стандартного відхилення) з надійністю $\gamma=0,95$.	$\sigma \in (0,52; 1,39)$.
5. Зроблено вимірювання зросту 20 студентів першого курсу університету та обчислено незміщену оцінку генеральної дисперсії $s^2=0,002$. Побудувати довірчий інтервал для середнього квадратичного відхилення зросту всіх студентів першого курсу університету з надійністю 0,95, вважаючи, що зростання має нормальний розподіл.	$\sigma \in (0,0342; 0,0657)$.
6. Для контролю терміну служби електроламп із великої партії було відібрано 17 електроламп. В результаті випробувань виявилося, що середній термін служби відібраних ламп дорівнює 980 год, а середнє квадратичне відхилення їх терміну служби – 18 год. Необхідно визначити межі, в яких із ймовірністю 0,95 укладено середній термін служби ламп у всій партії.	$a \in (970,5; 989,5)$.
7. Припустимо, відомо, що час відповіді сервера нормально розподілений з відомим стандартним відхиленням $\sigma=50$ мс. За випадковою вибіркою обсягу $n=100$ запитів виявили, що середній час відповіді складає $\bar{x}=320$ мс. Побудувати 95% довірчий інтервал для справжнього середнього a часу відповіді.	$a \in (310,2; 329,8)$.

Завдання	Відповідь

8. Визначити мінімальний обсяг вибірки для заданої точності оцінки δ_γ=0,2 математичного сподівання нормально розподіленої сукупності та довірчої ймовірності 0,925, якщо відоме середнє квадратичне відхилення генеральної сукупності σ=1,5.	n=179
9. Для обстеження великої партії виробів відібрано навмання 900 штук. Перевірка показала, що у тому числі 810 стандартні. Побудувати довірчий інтервал з надійністю 0,95 для частки стандартних виробів у партії.	$p \in (0{,}8004;\ 0{,}9186)$.
10. За вибіркою обсягу n=20 із нормально розподіленої генеральної сукупності обчислено значення вибіркової дисперсії $\sigma_{\text{в}}^2 = 1{,}5$. Побудувати довірчий інтервал для параметра σ^2 з надійністю 0,96.	$\sigma^2 \in (0{,}89;3{,}488)$
11. За вибіркою обсягу n=16 із нормально розподіленої генеральної сукупності знайдені вибіркове середнє $\bar{x}_{\text{в}} = 50$ і вибіркове виправлене середнє квадратичне відхилення s=4. Побудувати 90% довірчий інтервал для справжнього середнього a з невідомою дисперсією.	$a \in (48{,}247;\ 51{,}753)$

6 Перевірка статистичних гіпотез

6.1 Поняття статистичної гіпотези. Основні етапи перевірки гіпотези

При вирішенні практичних завдань, пов'язаних із застосуванням методів математичної статистики, часто виникає питання: чи може на підставі даних деякої вибірки бути прийняте або відкинуте певне припущення (гіпотеза) щодо генеральної сукупності. Наприклад, випробувано нову методику навчання. Чи можна за результатами випробування зробити обґрунтований висновок про те, що нова методика порівняно з попередньою ефективніша (має кращі оціночні показники). Аналогічний питання виникає і за апробації нових ліків, за впровадження нових технологій тощо.

Означення 6.1. *Перевіркою гіпотез* називається процедура порівняння висловленого припущення (гіпотези) з даними вибірки.

Ця процедура полягає у наступному. Щодо генеральної сукупності висловлюється певна гіпотеза (чи кілька гіпотез). З генеральної сукупності витягується вибірка. Потрібно вказати правило, яке давало б відповідь на запитання: чи слід відхилити гіпотезу (деякі гіпотези) або прийняти її (одну з них). Зазначимо, що статистичними методами довести гіпотезу не можна. Є лише можливість спростувати чи не спростувати її.

Означення 6.2. *Статистичною гіпотезою* називається будь-яке твердження, висловлене щодо невідомого закону генеральної сукупності або щодо числових характеристик цього закону (якщо відомий закон розподілу).

Приклади статистичних гіпотез: генеральна сукупність розподілена згідно із законом Пуассона, генеральна сукупність розподілена за нормальним законом, дисперсії двох нормальних законів рівні та інші.

Означення 6.3. Процес ухвалення рішення щодо статистичної гіпотези називається *статистичною перевіркою статистичних гіпотез*.

Означення 6.4. *Параметричною* називається статистична гіпотеза, якщо в ній міститься деяке твердження щодо одного чи кількох параметрів генеральної сукупності, розподіл якої вважається відомим (або близьким до відомого), найчастіше нормальним розподілом.

Ключові особливості.

Вимоги до даних: дані мають бути виміряні у кількісних шкалах (інтервальна або шкала відношень).

Припущення: вимагають виконання строгих умов, таких як: дані незалежні; дані у генеральній сукупності мають нормальний розподіл; для двох і більше вибірок: рівність дисперсій.

При виконанні всіх умов параметричні тести мають вищу ймовірність відхилити хибне твердження порівняно з непараметричними.

Об'єкти параметричних гіпотез – це середні значення, дисперсії, коефіцієнти кореляції Пірсона тощо.

Означення 6.5. *Непараметричні гіпотези* – це твердження, яке не робить жорстких припущень про вигляд розподілу генеральної сукупності (тобто, вони вільні від розподілу).

Непараметричні гіпотези замість параметрів генеральної сукупності нерідко зосереджуються на медіанах, рангах або частотах розподілу.

Ключові особливості.

Вимоги до даних: можуть використовуватися з даними, виміряними у порядковій (ранговій) чи номінальній шкалах, а також з кількісними даними, які не мають нормального розподілу.

Припущення: вимагають набагато менше припущень (наприклад, незалежність спостережень).

Застосовуються, коли обсяг вибірки занадто малий чи коли порушуються умови параметричних тестів (наприклад, сильна асиметрія даних).

Об'єкти гіпотез – це медіани, рангові суми, розподіли, зв'язок між категоріальними змінними.

Гіпотези позначаються великими латинськими літерами H_0, H_1, …

Означення 6.6. *Нульовою* (*основною*) *гіпотезою* H_0 називається припущення, якого ми дотримуємося спочатку, поки спостереження не змусять нас визнати протилежне.

Означення 6.7. *Альтернативною* (*конкуруючою*) *гіпотезою* H_1 називається гіпотеза, яка суперечить H_0 і яку ми приймаємо, якщо відхиляємо основну гіпотезу.

Нульова гіпотеза завжди одна, а альтернативних може бути декілька.

Означення 6.8. *Простою гіпотезою* називається статистична гіпотеза, якщо вона містить лише одне припущення.

Наприклад, гіпотеза "Математичне сподівання генеральної сукупності дорівнює 6" – проста гіпотеза.

Означення 6.9. *Складною гіпотезою* називається статистична гіпотеза, що складається з скінченного чи нескінченного числа простих гіпотез.

Наприклад, гіпотеза "Математичне сподівання генеральної сукупності менше 6" включає в собі нескінченну кількість припущень про те, чому ж дорівнює математичне сподівання генеральної сукупності. Це складна гіпотеза.

Гіпотезу перевіряють виходячи з вибірки, отриманої з генеральної сукупності. Через випадковість вибірки в результаті перевірки можуть виникнути помилки та приймаються неправильні рішення.

Помилки під час перевірки гіпотез

Оскільки будь-яке припущення має ймовірнісний характер, то можливі такі ситуації:

1) гіпотеза H_0 правильна, і під час перевірки вона приймається;
2) гіпотеза H_0 правильна, але під час перевірки вона відхиляється;
3) гіпотеза H_0 неправильна, і під час перевірки вона відхиляється;
4) гіпотеза H_0 неправильна, але під час перевірки вона приймається.

Розрізняють помилки першого та другого роду.

Означення 6.10. *Помилкою першого роду* називають помилку, яка припускається у випадку, коли відхилена правильна основна гіпотеза H_0 і прийнято конкуруючу гіпотезу H_1.

Означення 6.11. Імовірність помилки першого роду називають *рівнем значущості* та позначають α: $P(H_1/H_0)= \alpha$.

Означення 6.12. Величину $\gamma=1-\alpha$, тобто ймовірність прийняти правильну гіпотезу, називають *рівнем довіри* (*довірчим рівнем*).

Означення 6.13. *Помилкою другого роду* називають помилку, яка припускається в разі прийняття неправильної основної гіпотези H_0.

Імовірність помилки другого роду позначають β.

Означення 6.14. *Статистичним критерієм* називають правило ухвалення рішення, згідно з яким приймається або відхиляється основна гіпотеза H_0.

Класифікація статистичних критеріїв за видами:

- *Критерії згоди*. Перевірка припущення у тому, що досліджувана випадкова величина підпорядковується передбачуваному закону розподілу.
- *Критерії значущості*. Дозволяють підтвердити або спростувати припущення про параметри розподілу.
- *Критерії на однорідність*. Дослідження кількох вибірок на предмет їхньої належності до одного й того самого закону розподілу. Критерії однорідності є фундаментом дисперсійного аналізу, оскільки вони дозволяють встановити, чи впливає певний фактор на досліджувану величину, чи виявлені розбіжності є суто випадковими.

Цей поділ умовний, і найчастіше один і той самий критерій може бути використаний у різних якостях.

Часто для спрощення як нульова гіпотеза розглядається конкретне припущення, а як альтернативна, решта простору можливих варіантів.

Означення 6.15. *Потужність статистичного критерію* – це можливість відхилити помилкову альтернативну гіпотезу (потужність критерію дорівнює $1-\beta$).

Зауваження 6.1. Чим більша потужність критерію, тим ймовірніше, що він виявить помилковість альтернативної гіпотези. Зі зменшенням рівня значущості падає потужність критерію.

Означення 6.16. Випадкова величина K, побудована за спостереженнями для перевірки нульової гіпотези H_0, називається *статистикою критерію*.

При цьому передбачається, що розподіл статистики критерію відомий за справедливості гіпотези H_0.

Схема побудови критерію така: весь вибірковий простір ділиться на дві взаємодоповнюючі області: область відхилення основної

гіпотези H_0 – *критична область* S і область прийняття цієї гіпотези – *допустима область* $\overline{S}$.

Означення 6.17. *Критичними точками* $k_{кр}$, є граничні значення статистичного критерію, які розбивають множину всіх його можливих значень на дві неперетинні області: область прийняття гіпотези (де відхилення вважаються випадковими) та критичну область (де відхилення визнаються статистично значущими). Саме ці точки визначають поріг, за яким нульова гіпотеза H_0 має бути відхилена.

Існують три види критичних областей:

лівостороння критична область визначається з рівності:

$$P\left(K < k_{кр}\right) = \alpha, \tag{6.1}$$

тобто, при $K<k_{кр}$ нульова гіпотеза відхиляється (рис. 6.1):

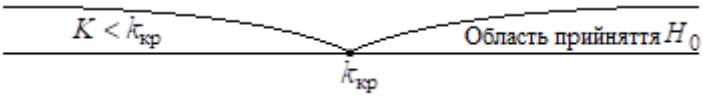


Рисунок 6.1

правостороння критична область визначається з рівності:

$$P\left(K > k_{кр}\right) = \alpha, \tag{6.2}$$

тобто, при $K>k_{кр}$ нульова гіпотеза відхиляється (рис. 6.2):

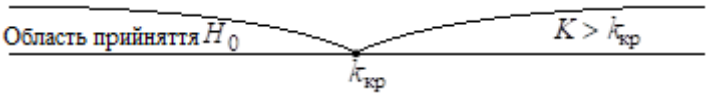


Рисунок 6.2

двостороння критична область визначається з рівності:

$$P\left(K < k_{кр1}\right) + P\left(K > k_{кр2}\right) = \alpha, \tag{6.3}$$

де $k_{кр1}$, $k_{кр2}$ – відповідно ліва і права критичні точки, тобто, при $K<k_{кр1}$ і $K>k_{кр2}$ нульова гіпотеза відхиляється (рис. 6.3):

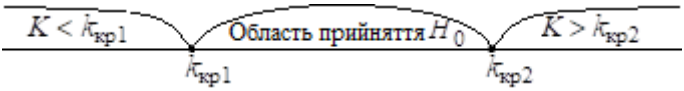


Рисунок 6.3

Найчастіше двосторонню критичну область будують як симетричну, визначаючи $k_{кр1}$ і $k_{кр2}$ відповідно з рівнянь

$$P(K<k_{кр1})=\alpha/2 \text{ і } P(K>k_{кр2})=\alpha/2. \quad (6.4)$$

На практиці при перевірці гіпотез задається рівень значущості α={0,1; 0,05; 0,01; 0,025; 0,005; 0,001}. Для кожного критерію, тобто відповідного розподілу, зазвичай складені таблиці, за якими знаходять критичні точки $k_{кр}$.

Означення 6.18. *Спостережуваним* (*емпіричним*) *значенням* $K_{спост}$ називають значення критерію, обчислене за вибіркою.

Основний принцип перевірки статистичних гіпотез полягає в наступному: якщо значення $K_{спост}$ потрапляє в критичну область S, то основну гіпотезу H_0 відхиляють і приймають альтернативну гіпотезу H_1, якщо $K_{спост}$ належить області прийняття гіпотези $\overline{S}$, то гіпотезу H_0 приймають, а гіпотезу H_1 відхиляють.

Розглянемо параметричні гіпотези. Якщо параметр є скаляром, то йдеться про однопараметричні гіпотези, якщо вектором, то про багатопараметричні гіпотези.

Основні етапи перевірки гіпотези

Перевірка параметричної статистичної гіпотези за допомогою критерію значущості може бути розбита на наступні етапи:

1) формулювання перевіряємої H_0 та альтернативної H_1 гіпотез;

2) обрання за змістом гіпотези рівня значущості α – ймовірності помилкового відхилення нульової гіпотези H_0. Цю величину називають також розміром критерію (тесту). Вибір величини рівня значущості залежить від розміру втрат, які понесемо у разі помилкового рішення. Найбільш поширеною є величина рівня значущості α=0,05 (5%), α=0,01 (1%) (якщо α=0,05, то у середньому в п'яти випадках зі 100 буде помилково відхилено висловлену гіпотезу);

3) вибір статистики K (наприклад, χ^2-розподіл Пірсона, t-розподіл Стьюдента, F-розподіл Фішера-Снедекора та інші), розподіл якої

відомий та залежить від параметра, який перевіряється, за умови, що правильна гіпотеза H_0. Вибір критерію залежить від: типу гіпотези (про середнє, про дисперсію тощо), виду розподілу (передбачається нормальний), обсягу вибірки;

4) за даними вибірки розраховується фактичне (спостережуване) значення обраного статистичного критерію $K_{спост}$;

5) за таблицею, відповідною розподілу обраного критерію, необхідно знайти критичну точку $k_{кр}$ для прийнятих рівня значущості α і обсягу вибірки n або кількості степенів свободи;

6) залежно від формулювання альтернативної гіпотези H_1 визначити критичну область S однією з нерівностей: $K<k_{кр}$, $K>k_{кр}$ або сукупністю нерівностей $K<k_{кр1}$ і $K>k_{кр2}$;

7) порівняти $K_{спост}$ з $k_{кр}$ та прийняти рішення, згідно з умовами (6.1) – (6.3), наприклад, для правосторонньої і двосторонньої симетричної областей виконання умови $K_{спост}\geq k_{кр}$ є підставою для відхилення гіпотези H_0, що не узгоджується з результатами спостережень, а не виконання – для прийняття гіпотези H_0, тобто вважати, що гіпотеза H_0 не суперечить результатам спостережень.

6.2 Перевірка непараметричних статистичних гіпотез

Фундаментальним завданням математичної статистики є ідентифікація теоретичного закону розподілу випадкової величини за її емпіричними даними. Якщо вигляд генерального розподілу невідомий, то формулюється припущення про його математичну модель. Зокрема, базовою гіпотезою у багатьох дослідженнях є припущення про нормальність розподілу досліджуваної ознаки.

Зауваження 6.2. При перевірці гіпотез про закон розподілу на заданому рівні значущості контролюється лише помилка першого роду, але не можна зробити висновок про рівень ризику, пов'язаного з прийняттям неправильної гіпотези, тобто з можливістю здійснення помилки другого роду.

Означення 6.19. Критерії, що встановлюють закон розподілу, називаються *критеріями згоди*.

Отже, критерій згоди – це статистичний інструмент, призначений для перевірки відповідності емпіричних даних теоретично передбачуваному закону розподілу. У математичній статистиці

найбільш поширеними є такі критерії згоди: χ^2 (хі-квадрат) Пірсона (універсальний критерій для дискретних та неперервних величин), критерій Колмогорова (базується на порівнянні емпіричної та теоретичної функцій розподілу), Критерій Мізеса-Смирнова (використовує квадрати відхилень для оцінки згоди).

6.2.1 Критерій згоди χ^2-Пірсона

Критерій згоди χ^2-Пірсона є одним з найбільш поширених критеріїв перевірки гіпотез про вид закону розподілу випадкової величини, що вивчається. Цей критерій дозволяє перевірити значущість розбіжності емпіричних (що спостерігаються) і теоретичних (очікуваних) частот. Таким чином, за допомогою даного критерію можна перевірити гіпотезу про належність вибірки, що спостерігається, деякому теоретичному закону розподілу.

Перевірка гіпотези за допомогою критерію χ^2-Пірсона здійснюється за наступною схемою.

1) Нехай з генеральної сукупності утворюється випадкова вибірка $x_1,x_2,\ldots,x_n$ спостережень випадкової величини X обсягом n. На її основі робиться припущення про нормальний закон розподілу. Висувається гіпотеза H_0: "генеральна сукупність розподілена нормально".

2) Область можливих значень випадкової величини X, що спостерігалися, розбивається на q інтервалів, що складаються з окремих значень для дискретної випадкової величини X.

Їх кількість визначається за формулою Стерджесса: q=1+3,322·lg(n), де n – обсяг вибірки (q – це найближче ціле до 1+3,322·lg(n)).

Довжина інтервалу обчислюється за формулою:

$$h = \frac{x_{\max} - x_{\min}}{q}, \qquad (6.5)$$

Нехай n_i – число елементів вибірки, що належать i-му інтервалу (i=1,2,…,q). Тоді $w_i = \frac{n_i}{n}$ – значення відносних частот для кожного

інтервалу (i=1,2,…,q). Очевидно, що $\sum_{i=1}^{q} n_i = n$, а $\sum_{i=1}^{q} w_i = 1$.

Складемо таблицю 6.1:

Таблиця 6.1

Інтервали	$[x_1,x_2)$	$[x_2,x_3)$	…	$[x_q,x_{q+1}]$
Середина інтервалу $x_i' = \frac{x_i + x_{i+1}}{2}$	x_1'	x_2'	…	x_q'
Частоти n_i	n_1	n_2	…	n_q
Відносна частота $w_i = \frac{n_i}{n}$	w_1	w_2	…	w_q

Емпірична функція розподілу $F^*(x) = \frac{n_x}{n}$, де n_x – число варіант, менших x; n – обсяг вибірки, $F^*(x)$=0 при $x \le x_{\min}$ і $F^*(x)$=1 при $x > x_{\max}$.

3) За вибіркою спостережень знаходять оцінки невідомих параметрів (наприклад, $\bar{x}_в$, $\sigma_в$) передбачуваного закону розподілу випадкової величини X.

Числові характеристики вибірки знаходимо за формулами:

а) *Вибіркове середнє*:

для дискретного ряду

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{q} x_i \cdot n_i = \sum_{i=1}^{q} x_i \cdot w_i \; ; \qquad (6.6)$$

для інтервального ряду

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{q} x_i' \cdot n_i = \sum_{i=1}^{q} x_i' \cdot w_i \; . \qquad (6.7)$$

б) *Вибіркова дисперсія*:

для дискретного ряду

$$D_в = \frac{1}{n}\sum_{i=1}^{q} (x_i - \bar{x}_в)^2 \cdot n_i = \sum_{i=1}^{q} (x_i - \bar{x}_в)^2 \cdot w_i \; ; \qquad (6.8)$$

для інтервального ряду

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{q}(x_i)^2 \cdot n_i - \left(\bar{x}_{\text{в}}\right)^2 = \sum_{i=1}^{q}\left(x_i'\right)^2 w_i - \left(\bar{x}_{\text{в}}\right)^2 . \qquad (6.9)$$

в) *Вибіркове середнє квадратичне відхилення*:

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} . \qquad (6.10)$$

Відповідно до гіпотези, щільність розподілу ймовірностей має вигляд:

$$f(x) = \frac{1}{\sigma_{\text{в}}\sqrt{2\pi}} \text{e}^{-\frac{\left(x-\bar{x}_{\text{в}}\right)^2}{2\sigma_{\text{в}}^2}} . \qquad (6.11)$$

Графік цієї функції називається *вирівнювальною кривою*.

4) Обчислюються теоретичні частоти.

а) Для дискретного ряду

$$n_i' = \frac{n \cdot h}{\sigma_{\text{в}}} \cdot \varphi(u_i), \qquad (6.12)$$

де n – обсяг вибірки, h – різниця між двома сусідніми варіантами, $u_i = \frac{x_i - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}$, значення $\varphi(u) = \frac{1}{\sqrt{2\pi}} \text{e}^{-\frac{u^2}{2}}$ знаходимо за таблицею А.1 (див. Додаток А).

б) Для інтервального ряду

$$n_i' = n \cdot p_i , \qquad (6.13)$$

де n – обсяг вибірки, p_i – теоретичні ймовірності потрапляння випадкової величини у відповідний інтервал і обчислюються за формулою:

$$p_i = P(x_i \le X < x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}\right) - \Phi\left(\frac{x_i - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}\right), \qquad (6.14)$$

де $\Phi(x)$ – інтегральна функція Лапласа, значення якої знаходимо за таблицею А.2 (див. Додаток А).

4) Знаходиться спостережене значення критерію Пірсона за формулою

$$\chi^2_{\text{спост}} = \sum_{i=1}^{q} \frac{(n_i - n_i')^2}{n_i'} . \qquad (6.15)$$

5) За таблицею критичних точок χ^2-розподілу, за заданим рівнем значущості α та числом степенів свободи $k=q-3$ (q – число груп для дискретного ряду або число інтервалів для інтервального ряду) знаходять критичну точку $\chi^2_{\text{кр}}(\alpha; k)$ правосторонньої критичної області (таблиця А.5, див. Додаток А).

6) Якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$, то немає підстав відхилити гіпотезу H_0 про нормальний розподіл генеральної сукупності. Інакше кажучи, емпіричні і теоретичні частоти різняться незначно. Якщо $\chi^2_{\text{спост}} \geq \chi^2_{\text{кр}}$, то гіпотезу H_0 про вид розподілу відхиляємо.

Зауваження 6.3. Для коректного застосування статистичних методів необхідно, щоб частота в кожному інтервалі становила не менше ніж 5 ($n_i \geq 5$). Якщо в певному інтервалі кількість спостережень $n_i < 5$, то доцільно об'єднати його з сусідніми до досягнення необхідного мінімуму. Тому при обчисленні числа степенів свободи величина q береться відповідно зменшеною на число об'єднаних інтервалів, тобто дорівнює числу інтервалів, що залишилися після об'єднання.

Приклад 40. У результаті перевірки на нестандартність 200 ящиків консервів отримано наступний емпіричний розподіл (у першому рядку вказано кількість x_i нестандартних коробок консервів в одному ящику; у другому рядку n_i – частота, тобто кількість ящиків, що містять коробки нестандартних консервів):

x_i	5	7	9	11	13	15	17	19	21
n_i	14	25	28	30	26	22	23	19	13

Потрібно за рівня значущості 0,05 перевірити за критерієм згоди Пірсона чи узгоджується гіпотеза H_0 про нормальний розподіл генеральної сукупності X – число нестандартних коробок з емпіричним розподілом вибірки.

Розв'язання. За умовою n=200, q=9. Оскільки маємо дискретний ряд, то для знаходження вибіркового середнього та вибіркової дисперсії, скористаємось формулами (6.6) та (6.8):

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{q} x_i \cdot n_i = \frac{1}{200}\cdot\left(5\cdot 14 + 7\cdot 25 + 9\cdot 28 + 11\cdot 30 + 13\cdot 26 + 15\cdot 22 + \right.$$

$$\left. + 17\cdot 23 + 19\cdot 19 + 21\cdot 13\right) = \frac{2520}{200} = 12{,}6\,.$$

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{q}(x_i - \bar{x}_{\text{в}})^2 \cdot n_i = \frac{1}{200}\left(\left(5-12{,}6\right)^2\cdot 14 + \left(7-12{,}6\right)^2\cdot 25 + \right.$$

$$+ \left(9-12{,}6\right)^2\cdot 28 + \left(11-12{,}6\right)^2\cdot 30 + \left(13-12{,}6\right)^2\cdot 26 + \left(15-12{,}6\right)^2\cdot 22 +$$

$$\left. + \left(17-12{,}6\right)^2\cdot 23 + \left(19-12{,}6\right)^2\cdot 19 + \left(21-12{,}6\right)^2\cdot 13\right) = \frac{4304}{200} = 21{,}52\,.$$

Тоді вибіркове середнє квадратичне відхилення за формулою (6.10) $\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{21{,}52} \approx 4{,}639$.

Обчислимо теоретичні частоти з огляду на те, що n=200, h=2, за формулою (6.12)

$$n_i' = \frac{n\cdot h}{\sigma_{\text{в}}}\cdot\varphi(u_i) = \frac{200\cdot 2}{4{,}639}\cdot\varphi(u_i) = 86{,}2255\cdot\varphi(u_i),\ \text{де}\ u_i = \frac{x_i - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}\,.$$

Для зручності отримані результати подамо у вигляді наступної таблиці 6.2:

Таблиця 6.2

x_i	$u_i = \frac{x_i - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}$	$\varphi(u_i)$	n_i'	n_i	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
5	−1,638	0,104	8,967	14	5,033	25,331	2,8249
7	−1,207	0,1919	16,547	25	8,453	71,453	4,3182
9	−0,776	0,2943	25,376	28	2,624	6,885	0,2713
11	−0,345	0,3758	32,404	30	−2,404	5,779	0,1783
13	0,086	0,3975	34,275	26	−8,275	68,476	1,9978
15	0,517	0,3485	30,05	22	−8,05	64,803	2,1565

Продовження Таблиці 6.2

x_i	$u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$	$\varphi(u_i)$	n_i'	n_i	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
17	0,948	0,2541	21,91	23	1,09	1,188	0,0542
19	1,38	0,1539	13,27	19	5,73	32,833	2,4742
21	1,811	0,0775	6,682	13	6,318	39,917	5,9738
Σ				200			$\chi^2_{спост}$ = =20,2492

З розрахункової таблиці знаходимо спостережене значення критерію Пірсона: $\chi^2_{спост} = 20{,}2492$.

За таблицею А.5 – Критичні точки розподілу Пірсона χ^2 (див. Додаток А) при рівні значущості α=0,05 та числу степенів свободи k=q−3=9−3=6 знаходимо критичну точку $\chi^2_{кр}(0{,}05; 6) = 12{,}592$. Тобто $\chi^2_{спост} > \chi^2_{кр}$ – отже, гіпотезу про нормальний розподіл генеральної сукупності відхиляємо. Інакше кажучи, емпіричні і теоретичні частоти різняться значимо.

Відповідь: Гіпотезу H_0 про нормальний розподіл генеральної сукупності X – число нестандартних коробок відхиляємо.

Приклад 41. Дано статистичне розподілення терміну служби інструменту до виходу за межі точності (у місяцях):

Інтервали – термін служби у місяцях	[15,20)	[20,25)	[25,30)	[30,35)	[35,40)	[40,45]
n_i – частота	6	9	24	35	16	10

Перевірити гіпотезу H_0 про нормальний розподіл терміну служби інструменту за допомогою критерію Пірсона (%) при рівні значущості α=0,05.

Розв'язання. За умовою n=100, q=6. Оскільки маємо інтервальний ряд, то знайдемо середини інтервалів за формулою $x_i' = \frac{x_i + x_{i+1}}{2}$. Складемо таблицю:

Інтервали – термін служби у місяцях	[15,20)	[20,25)	[25,30)	[30,35)	[35,40)	[40,45]
Середина інтервалу x_i'	17,5	22,5	27,5	32,5	37,5	42,5
n_i – частота	6	9	24	35	16	10

Для знаходження вибіркового середнього та вибіркової дисперсії, скористаємось формулами (6.7) та (6.9):

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{q} x_i' \cdot n_i = \frac{1}{100}\cdot(17{,}5\cdot 6 + 22{,}5\cdot 9 + 27{,}5\cdot 24 + 32{,}5\cdot 35 + 37{,}5\cdot 16 +$$

$$+ 42{,}5\cdot 10) = \frac{3130}{100} = 31{,}3\,.$$

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{q} (x_i)^2 \cdot n_i - \left(\bar{x}_{\text{в}}\right)^2 = \frac{1}{100}\cdot\Big((17{,}5)^2\cdot 6 + (22{,}5)^2\cdot 9 + (27{,}5)^2\cdot 24 +$$

$$+ (32{,}5)^2\cdot 35 + (37{,}5)^2\cdot 16 + (42{,}5)^2\cdot 10\Big) - (31{,}3)^2 = \frac{102075}{100} - 979{,}69 =$$

$$= 1020{,}75 - 979{,}69 = 41{,}06\,.$$

Тоді вибіркове середнє квадратичне відхилення за формулою (6.10)

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{36{,}4737} \approx 6{,}4078\,.$$

Обчислимо теоретичні частоти n_i' за формулою (6.13), а теоретичні ймовірності p_i за формулою (6.14). Для зручності отримані результати подамо у вигляді таблиці 6.3, де $z_{i+1} = \dfrac{x_{i+1} - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}$ і $z_i = \dfrac{x_i - \bar{x}_{\text{в}}}{\sigma_{\text{в}}}$:

Таблиця 6.3

x_i	x_{i+1}	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	p_i	$n\cdot p_i$
15	20	–2,54	–1,76	–0,4945	–0,4608	0,0337	3,37
20	25	–1,76	–0,98	–0,4608	–0,3365	0,1243	12,43
25	30	–0,98	–0,20	–0,3365	–0,0793	0,2572	2572,
30	35	–0,20	0,58	–0,0793	0,2190	0,2983	29,83
35	40	0,58	1,36	0,2190	0,4131	0,1941	19,41
40	45	1,36	2,14	0,4131	0,4838	0,0707	7,07

Знайдемо спостережене значення критерію Пірсона $\chi^2_{\text{спост}}$ за формулою (6.15).

Для цього складемо наступну розрахункову таблицю 6.4:

Таблиця 6.4

i	n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
1	6	3,37	2,63	6,9169	2,0525
2	9	12,43	–3,43	11,7649	0,9465
3	24	25,72	–1,72	2,9584	0,1150
4	35	29,83	5,17	26,7289	0,8960
5	16	19,41	–3,41	11,6281	0,5991
6	10	7,07	2,93	8,5849	1,2143
Σ	100	97,83			$\chi^2_{\text{спост}} = 5{,}8234$

З розрахункової таблиці знаходимо спостережене значення критерію Пірсона: $\chi^2_{\text{спост}} = 5{,}8234$.

За таблицею А.5 – Критичні точки χ^2-розподілу Пірсона (див. Додаток А) при рівні значущості α=0,05 та числу степенів свободи k=q−3=6−3=3 знаходимо критичну точку $\chi^2_{\text{кр}}(0{,}05;3) = 7{,}815$. Тобто $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$ – отже, немає підстав відхилити гіпотезу H_0 про нормальний розподіл терміну служби інструменту.

Відповідь: Гіпотезу H_0 про нормальний розподіл терміну служби інструменту не відхиляємо.

Приклад 42. Сформовано ранжовану вибірку по п'ятдесяти співробітникам виробничої ділянки, що відображає їх індивідуальну результативність у поточному періоді у відсотках щодо показників минулих років. Дані внесені в таблицю 6.5.

Таблиця 6.5

92	93	94	97	98	99	100	101	102	103
104	104	105	106	106	107	107	107	108	109
109	110	110	110	111	111	112	112	112	113
113	114	115	116	116	117	117	117	119	119
119	121	122	124	124	125	127	128	132	134

Перевірити гіпотезу H_0 про нормальний розподіл дослідних даних за допомогою критерію Пірсона (%) при рівні значущості α=0,05.

Розв'язання. Складемо інтервальний варіаційний ряд. Маємо x_{min}=92, x_{max}=134. Розмах вибірки: R=x_{max}–x_{min}=134–92=42. Визначаємо кількість інтервалів q за формулою Стерджесса:

$$q = 1 + 3{,}322 \cdot \lg(50) = 1 + 3{,}322 \cdot 1{,}699 \approx 1 + 5{,}644 \approx 6{,}644.$$

Округлюємо q: q=7. Обчислюємо довжину інтервалу h за формулою: h=R/q=42/7=6.

Визначаємо межі інтервалів: [92;98), [98;104), [104;110), [110;116), [116;122), [122;128), [128;134]. Для кожного інтервалу, визначаємо: середини інтервалів x_i'=(x_i+x_{i+1})/2, частоти n_i Складаємо таблицю 6.6.

Таблиця 6.6

	[92;98)	[98;104)	[104;110)	[110;116)	[116;122)	[122;128)	[128;134]
x_i'	95	101	107	113	119	125	131
n_i	4	6	11	12	9	5	3

Оскільки маємо інтервальний варіаційний ряд, то застосовуємо для знаходження вибіркового середнього $\bar{x}_в$ формулу (4.8):

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{q} x_i' \cdot n_i = \frac{1}{50} \cdot (95 \cdot 4 + 101 \cdot 6 + 107 \cdot 11 + 113 \cdot 12 + 119 \cdot 9 + 125 \cdot 5 +$$

$$+ 131 \cdot 3) = \frac{1}{50} \cdot (380 + 606 + 1177 + 1356 + 1071 + 625 + 393) = \frac{5608}{50} = 112{,}16$$

Для знаходження вибіркової дисперсії $D_в$ застосовуємо формулу (4.13):

$$D_в = \overline{x_в^2} - (\bar{x}_в)^2 = \frac{1}{n}\sum_{i=1}^{q} (x_i')^2 \cdot n_i - (\bar{x}_в)^2 = \frac{1}{50} \cdot (95^2 \cdot 4 + 101^2 \cdot 6 + 107^2 \cdot 11 +$$

$$+ 113^2 \cdot 12 + 119^2 \cdot 9 + 125^2 \cdot 5 + 131^2 \cdot 3) - (112{,}16)^2 = \frac{1}{50} \cdot (36100 + 61206 +$$

$$+ 125939 + 153228 + 127449 + 78125 + 51483) - 12579{,}8656 = \frac{633530}{50} -$$

$$- 12579{,}8656 = 12670{,}60 - 12579{,}8656 = 90{,}7344.$$

Вибіркове середнє квадратичне відхилення $\sigma_в$ знаходимо за формулою (4.12): $\sigma_в = \sqrt{D_в} = \sqrt{90{,}7344} \approx 9{,}5255$.

Обчислимо теоретичні частоти $n \cdot p_i$ за формулою (6.13), а теоретичні ймовірності p_i за формулою (6.14).

Для зручності отримані результати подамо у вигляді таблиці 6.7:

Таблиця 6.7

x_i	x_{i+1}	n_i	p_i	$n \cdot p_i$	$(n_i - n \cdot p_i)^2$	$\frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$
92	98	4	0,0511	2,555	1,221	0,1373
98	104	6 (10)	0,1268	6,34 (8,895)		
104	110	11	0,2141	10,705	0,087	0,0081
110	116	12	0,2464	12,32	0,102	0,0083
116	122	9	0,1931	9,655	0,429	0,0444
122	128	5	0,1030	5,15	0,951	0,1354
128	134	3 (8)	0,0375	1,875 (7,025)		
Σ		50	0,972	48,6		$\chi^2_{\text{спост}} = 0{,}3335$

З розрахункової таблиці знаходимо спостережене значення критерію Пірсона: $\chi^2_{\text{спост}}=0{,}3335$.

Оскільки в емпіричному розподілі кількість спостережень у першому і останньому інтервалах ($n_1=4$, $n_7=3$) менша за 5, для коректного застосування критерію χ^2-Пірсона ці інтервали було об'єднано із сусідніми (див. табл. 6.7). Після перегрупування кількість інтервалів становить $q=5$ і число степенів свободи $k=q-3=5-3=2$.

За таблицею А.5 – Критичні точки χ^2-розподілу Пірсона (див. Додаток А) при рівні значущості $\alpha=0{,}05$ знаходимо критичну точку $\chi^2_{\text{кр}}(0{,}05;2)=5{,}991$.

Оскільки $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$, то підстави для відхилення нульової гіпотези H_0 відсутні. Отже, припускається, що емпіричні дані розподілені за нормальним законом.

Відповідь: Гіпотеза H_0 про нормальний розподіл узгоджується з дослідними даними.

6.2.2 Критерій згоди Колмогорова

Критерій згоди Колмогорова (часто званий критерієм Колмогорова-Смирнова для однієї вибірки) дозволяє перевірити, чи відповідає емпіричний (реальний) розподіл даних якомусь теоретичному закону (наприклад, нормальному, рівномірному чи експоненційному).

Критерій Колмогорова досить часто застосовується практично завдяки своїй простоті. Найважливіша умова застосування критерію згоди Колмогорова – це неперервні випадкові величини (вага, зріст, час, дохід). Для дискретних даних (кількість дітей, кидки кубика) він працює некоректно. У класичному вигляді критерій вимагає, щоб заздалегідь були відомі параметри розподілу (наприклад, не просто "нормальний", а "нормальний із середнім 0 і відхиленням 1"). Тобто його застосування можливе лише тоді, коли теоретична функція розподілу $F_0(x)$ задана повністю. Якщо параметри (середнє, дисперсія) оцінюються за самою вибіркою, потрібно використовувати модифікацію – критерій Ліллієфорса. Він точніше, ніж критерій χ^2-Пірсона при невеликих обсягах даних, тому що не вимагає угруповання даних в інтервали (при якій втрачається інформація).

Незалежно від розподілу формулюється нульова гіпотеза H_0: $F_n(x)=F_0(x)$ і альтернативна гіпотеза H_1: $F_n(x)\neq F_0(x)$, де $F_0(x)$ – теоретична функція розподілу, $F_n(x)$ – емпірична функція розподілу.

Суть критерію Колмогорова полягає у порівнянні теоретичної функції розподілу $F_0(x)$, що відповідає припущеній гіпотезі, та емпіричної функції $F_n(x)$, побудованої на основі вибіркових даних. Мірою розбіжності між ними є статистика D_n, яка визначається як максимальне значення абсолютної різниці між теоретичною та емпіричною функціями розподілу в усій області визначення. Критерій шукає максимальну вертикальну відстань між цими двома кривими.

Ця відстань позначається як

$$D_n = \max_x \left| F_n(x) - F_0(x) \right| \tag{6.16}$$

і називається *статистикою критерію Колмогорова*.

Колмогоров довів, що при $n\to\infty$ закон розподілу випадкової величини

$$\lambda = D_n \cdot \sqrt{n} \tag{6.17}$$

незалежно від виду розподілу випадкової величини X прямує до закону розподілу Колмогорова.

Зауваження 6.4. Для малих вибірок, тобто при $n<40$ (у деяких джерелах $n<25$), розподіл статистики D_n залежить від конкретного числа спостережень. У цьому випадку використовується "Таблиця критичних значень $D_{\alpha,n}$ критерію Колмогорова-Смирнова" (таблиця А.8, див. Додаток А).

Зауваження 6.5. Для великих вибірок, тобто при $n\geq 40$, коли n зростає, величина $D_n \cdot \sqrt{n}$ перестає залежати від n і починає підпорядковуватися закону розподілу Колмогорова. У цьому випадку використовується статистика $\lambda = D_n \cdot \sqrt{n}$ і "Таблиця критичних значень λ_α розподілу Колмогорова" (таблиця А.9, див. Додаток А).

1. Перевірка гіпотези про нормальний закон розподілу за допомогою критерію згоди Колмогорова здійснюється за наступною схемою:

1) Для вибірки малого обсягу:

а) знаходимо емпіричну функцію розподілу $F_n(x)$ за формулою $F_n(x_i) = \frac{i}{n}$, де i – порядковий номер елемента x_i у відсортованій вибірці та значення теоретичної функції $F_0(x)$, яка використовується в критерії Колмогорова, – це накопичена ймовірність від $-\infty$ до x. Вона не може бути від'ємною та завжди знаходиться в межах від 0 до 1. Для кожного значення x_i вибірки знаходимо ймовірність ідеального нормального розподілу за формулою $F_0(x_i) = 0{,}5 + \Phi\left(\frac{x_i - a}{\sigma}\right)$, де $\Phi(x)$ – значення функції Лапласа, яке знаходимо за таблицею А.2 (див. Додаток А);

б) визначаємо міру розбіжності D_n між теоретичним та емпіричним розподілами за формулою (6.16);

в) за таблиця А.8 (див. Додаток А) обчислюють $D_{\alpha,n}$ за заданими n та рівнем значущості α.

Якщо $D_n \le D_{\alpha,n}$, то вважають, що гіпотеза H_0 не суперечить дослідним даним. Якщо $D_n > D_{\alpha,n}$, то гіпотеза H_0, що випадкова величина X має заданий закон розподілу, відхиляється.

2) Для вибірки великого обсягу:

а) складаємо інтервальний ряд;

б) знаходимо накопичені частоти n_i' для кожного інтервалу і визначаємо емпіричну функцію розподілу $F_n(x_i) = \dfrac{n_i'}{n}$ та для кожного значення правої межі інтервалу x_i знаходимо передбачувану теоретичну функцію розподілу $F_0(x_i) = 0{,}5 + \Phi\left(\dfrac{x_i - a}{\sigma}\right)$, де $\Phi(x)$ – значення функції Лапласа знаходимо за таблицею А.2 (див. Додаток А);

в) визначаємо міру розбіжності D_n між теоретичним та емпіричним розподілами за формулою (6.16);

г) за таблицею А.9 (див. Додаток А) обчислюють λ_α за заданим рівнем значущості α (це $\lambda_{кр}$). Знаходимо критичне значення $D_{кр}$ за формулою:

$$D_{кр} = \frac{\lambda_\alpha}{\sqrt{n}}. \qquad (6.18)$$

Якщо $D_n > D_{кр}$, то гіпотеза H_0, що випадкова величина X має заданий закон розподілу, відхиляється. Якщо $D_n \le D_{кр}$, то вважають, що гіпотеза H_0 не суперечить дослідним даним.

2. Перевірка гіпотез про закони розподілу: рівномірний, експоненціальний та Пуассона за допомогою критерію згоди Колмогорова здійснюється за наступною схемою:

1) Задану вибірку x_i $\left(i = \overline{1,n}\right)$ обсягу n відсортуємо за зростанням: $x_1 \le x_2 \le \ldots \le x_n$.

2) Знаходимо значення емпіричної функції $F_n(x)$ для кожного i-го елемента вибірки: $F_n(x_i) = i/n$.

3) Знаходимо значення теоретичної функції $F_0(x)$ для кожного x_i вибірки за формулою відповідного закону розподілу (попередньо обчисливши за вибіркою параметри типу λ):

Для рівномірного закону на $[a;b]$:

а) знаходимо межі: $a=x_{\min}$ (перше число x_1), $b=x_{\max}$ (останнє число x_n);

б) теоретична функція $F_0(x_i)=\dfrac{x_i-a}{b-a}$.

Для експоненціального закону:

а) знаходимо середнє $\bar{x}=\dfrac{1}{n}\sum_i x_i$;

б) знаходимо параметр λ: $\lambda=1/\bar{x}$;

в) теоретична функція $F_0(x_i)=1-\mathrm{e}^{-\lambda x_i}$.

Для закону Пуассона, оскільки він дискретний:

а) знаходимо середнє $\bar{x}=\dfrac{1}{n}\sum_i x_i$;

б) знаходимо параметр λ: $\lambda=\bar{x}$;

в) для кожного цілого k (0,1,2…) знаходимо ймовірності: $P(k)=\dfrac{\lambda^k\,\mathrm{e}^{-\lambda}}{k!}$;

г) теоретична функція $F_0(x_i)$ – це накопичена сума ймовірностей $P(k)$ для всіх $k<x_i$.

4) Шукаємо найбільшу різницю $|F_n(x)-F_0(x)|$.

Зауваження 6.6. Оскільки функція $F_n(x)$ – це сходинка, то розрив потрібно перевіряти у двох точках: "на початку сходинки" та "наприкінці". Тому дивимося різницю $F_0(x_i)$ з поточним значенням i/n та з попереднім $(i-1)/n$.

Для зручності складаємо розрахункову таблицю 6.8:

Таблиця 6.8

i	x_i	$F_n(x_i)=i/n$	$F_n(x_{i-1})=(i-1)/n$	$F_0(x_i)$ (за теорією)	$\lvert F_n(x_i)-F_0(x_i)\rvert$	$\lvert F_n(x_{i-1})-F_0(x_i)\rvert$
1	x_1	$1/n$	0	…	…	…
2	x_2	$2/n$	$1/n$	…	…	…
…	…	…	…	…	…	…
n	x_n	1	$(n-1)/n$	…	…	…

5) обчислити максимальне відхилення D_n між емпіричною функцією $F_n(x)$ та теоретичною $F_0(x)$. Для цього використовуємо формулу:

$$D_n = \max_{1\le i\le n}\left\{\max\left(\left|F_0(x_i)-\frac{i-1}{n}\right|,\left|F_0(x_i)-\frac{i}{n}\right|\right)\right\}. \qquad (6.19)$$

Тобто, з двох останніх колонок таблиці 6.8 вибираємо найбільше значення. Воно і є максимальним відхиленням – D_n.

6) За таблицею А.8 (див. Додаток А) обчислюють $D_{\alpha,n}$ за заданими n та рівнем значущості α.

Якщо $D_n \le D_{\alpha,n}$, то вважають, що гіпотеза H_0 не суперечить дослідним даним. Якщо $D_n > D_{\alpha,n}$, то гіпотеза H_0, що випадкова величина X має заданий закон розподілу, відхиляється.

Приклад 43. На заводі верстат фасує цукор у пакети по 1000 гр. З досвіду минулих років відомо, що похибка верстата розподілена нормально із середнім a=1000 і стандартним відхиленням σ=5. Беруть випадкову пробу з 10 пакетів і перевіряють: чи не збилися налаштування верстата? Отримали дані (вага в грамах):

992,998,1002,1005,1001,999,1003,1000,995,997.

Перевірити гіпотезу H_0, що дані розподілені нормально з заданими параметрами, використовуючи критерій згоди Колмогорова з рівнем значущості α=0,05.

Розв'язання. Маємо вибірку малого обсягу. Сортуємо значення вибірки за зростанням:

992, 995, 997, 998, 999, 1000, 1001, 1002, 1003, 1005.

Знаходимо значення емпіричної функції $F_n(x)$ і значення теоретичної функції $F_0(x)$: $F_n(x_i)=\frac{i}{n}$, де i – порядковий номер елемента x_i, $F_0(x_i)=0{,}5+\Phi\left(\frac{x_i-a}{\sigma}\right)$, де $\Phi(x)$ – значення функції Лапласа, яке знаходимо за таблицею А.2 (див. Додаток А).

Знайдемо статистику D_n за формулою (6.16). Знаходимо різниці $|F_n(x_i)-F_0(x_i)|$ та $|F_n(x_{i-1})-F_0(x_i)|$.

Складаємо таблицю 6.9.

Таблиця 6.9

x_i	992	995	997	998	999
$F_n(x_i)$	0,1	0,2	0,3	0,4	0,5
$F_0(x_i)$	0,0548	0,1587	0,2743	0,3446	0,4207
$\lvert F_n(x_i)-F_0(x_i)\rvert$	0,0452	0,0413	0,0257	0,0554	0,0793
$\lvert F_n(x_{i-1})-F_0(x_i)\rvert$	0,0548	0,0587	0,0743	0,0446	0,0207

Продовження Таблиці 6.9

x_i	1000	1001	1002	1003	1005
$F_n(x_i)$	0,6	0,7	0,8	0,9	1,0
$F_0(x_i)$	0,5	0,5793	0,6554	0,7257	0,8413
$\lvert F_n(x_i)-F_0(x_i)\rvert$	0,1	0,1207	0,1446	0,1743	0,1587
$\lvert F_n(x_{i-1})-F_0(x_i)\rvert$	0	0,0207	0,0446	0,0743	0,587

Вибираємо найбільше значення з останніх двох рядків. Це і буде статистика $D_n=D_{10}=0{,}1743$.

За Таблицею А.8 (див. Додаток А) знаходимо $D_{0,05;10}=0{,}40925$ та порівнюємо з отриманим значенням $D_{10}=0{,}1743$. Оскільки $D_{10}< D_{0,05;10}$, то гіпотезу H_0 що дані з заданими параметрами розподілені нормально не відхиляємо.

Відповідь: З ймовірністю 95% можна стверджувати, що поточні налаштування обладнання відповідають заданим стандартам точності.

Приклад 44. Верстат виготовляє деталі. Перевіряють калібрування верстата, взявши 100 деталей, згрупованих в інтервали:

Інтервали	$[988;992)$	$[992;996)$	$[996;1000)$	$[1000;1004)$	$[1004;1008]$
n_i	7	15	28	30	20

Припустимо, що похибка верстата розподілена нормально із середнім $a=1000$ і стандартним відхиленням $\sigma=5$. Перевірити гіпотезу H_0 про нормальний розподіл ваги деталей, використовуючи критерій згоди Колмогорова з рівнем значущості $\alpha=0{,}05$.

Розв'язання. Знаходимо значення емпіричної функції $F_n(x)$. Вона обчислюється як відношення накопиченої кількісті деталей n_i' для розглядуваного інтервалу до загальної їх кількісті n ($n=100$). Знаходимо значення теоретичної функції $F_0(x)$. Для кожного значення x_i (права межа інтервалу) потрібно знайти ймовірність ідеального нормального

розподілу за формулою: $F_0(x_i) = 0{,}5 + \Phi\left(\frac{x_i - a}{\sigma}\right)$, де $\Phi(x)$ – значення функції Лапласа знаходимо за таблицею А.2 (див. Додаток А).

Знайдемо значення різниці $|F_n(x)–F_0(x)|$ для кожного значення x_i (права межа інтервалу) та складемо таблицю 6.10.

Таблиця 6.10

Інтервали	Частота n_i	Накопичена частота n_i'	Функція $F_n(x)$	Права межа x_i	Функція $F_0(x)$	Різниця $\|F_n(x)–F_0(x)\|$
[988;992)	7	7	0,07	992	0,0548	0,0152
[992;996)	15	22	0,22	996	0,2119	0,0081
[996;1000)	28	50	0,50	1000	0,5	0
[1000;1004)	30	80	0,80	1004	0,7881	0,0119
[1004;1008]	20	100	1,0	1008	0,9452	0,0548

Вибираємо максимальну різницю: D_n=0,0548 (найбільше в останньому стовпці). За таблицею А.9 (див. Додаток А) знаходимо критичне значення λ_α за заданим рівнем значущості α=0,05: $\lambda_{кр}$=1,358. За формулою (6.18) знайдемо $D_{кр} = \frac{1{,}358}{\sqrt{100}} = \frac{1{,}358}{10} = 0{,}1358$.

Оскільки розрахункове D_n=0,0548 менше критичного $D_{кр}$=01358, то робимо висновок, що гіпотеза H_0 про нормальний розподіл ваги деталей не відхиляється. Верстат працює коректно відповідно до нормального закону.

Відповідь: Гіпотеза H_0 про нормальний розподіл ваги деталей не відхиляється.

Приклад 45. Пасажир, який приходить у випадкові моменти часу на зупинку автобуса, протягом п'яти поїздок фіксував час очікування автобуса: 5,1; 3,7; 1,2; 9,2; 4,8 (хвилин). Перевірити гіпотезу, що час очікування рівномірно розподілено на відрізку [0;10], використовуючи критерій згоди Колмогорова з рівнем значущості α=0,05.

Розв'язання. Перевіряємо гіпотезу H_0: "Час очікування рівномірно розподілено на відрізку [0;10]". Тоді альтернативна гіпотеза H_1: "Розподіл часу очікування відмінний від рівномірного на відрізку [0;10]".

Для рівномірного розподілу на відрізку [a;b] теоретична функція розподілу $F_0(x)$ має вигляд: $F_0(x)=\begin{cases} 0, & x<a, \\ \dfrac{x-a}{b-a}, & a\le x\le b, \\ 1, & x>b. \end{cases}$

За умовою a=0, b=10. Тому для $x\in[0;10]$ матимемо $F_0(x)=\dfrac{x}{10}$.

Спочатку упорядкуємо дані щодо зростання (варіаційний ряд): x_1=1,2; x_2=3,7; x_3=4,8; x_4=5,1; x_5=9,2.

Потрібно обчислити максимальне відхилення D_n між емпіричною функцією $F_n(x)$ та теоретичною $F_0(x)$. Для цього використовуємо формулу:

$$D_n=\max_{1\le i\le n}\left\{\max\left(\left|F_0(x_i)-\frac{i-1}{n}\right|,\left|F_0(x_i)-\frac{i}{n}\right|\right)\right\}.$$

Складемо наступну таблицю 6.11, враховуючи, що n=5, $F_0(x_i)=\dfrac{x_i}{10}$ і $D_i=\max\left(\left|F_0(x_i)-\dfrac{i-1}{5}\right|,\left|F_0(x_i)-\dfrac{i}{5}\right|\right)$.

Таблиця 6.11

i	x_i	$F_0(x_i)$	$(i-1)/5$	$i/5$	$\lvert F_0(x_i)-(i-1)/5\rvert$	$\lvert F_0(x_i)-i/5\rvert$	D_i
1	1,2	0,12	0	0,2	0,12	0,08	0,12
2	3,7	0,37	0,2	0,4	0,17	0,03	0,17
3	4,8	0,48	0,4	0,6	0,08	0,12	0,12
4	5,1	0,51	0,6	0,8	0,09	0,29	0,29
5	9,2	0,92	0,8	1,0	0,12	0,08	0,12

Значення статистики, що спостерігається: $D_5=\max\limits_{1\le i\le 5}(D_i)=0{,}29$. За таблицею А.8 (див. Додаток А) знаходимо $D_{0,05;5}$=0,563 та порівнюємо з отриманим значенням D_5=0,29. Оскільки $D_5<D_{0,05;5}$, то немає підстав відхилити гіпотезу H_0.

Відповідь: Дані не суперечать гіпотезі H_0 про те, що час очікування рівномірно розподілено на відрізку [0;10] при рівні значущості α=0,05.

6.2.3 Критерій Ліллієфорса

Критерій Ліллієфорса – це модифікація критерію Колмогорова-Смирнова для перевірки відповідності емпіричних даних нормальному закону. Особливість цього підходу полягає у використанні вибіркових оцінок для побудови теоретичної моделі. Оскільки параметри розподілу обчислюються безпосередньо з наявної вибірки, метод використовує специфічні таблиці критичних значень. Тест оцінює максимальну різницю між накопиченою частотою спостережень та теоретичною функцією розподілу, що дозволяє підтвердити або спростувати нульову гіпотезу.

Його застосовують для малих та середніх вибірках ($4 \le n \le 50$).

Алгоритм перевірки нормальності за критерієм Ліллієфорса.

1) Дані заданої вибірки обсягу n розтшовуємо у порядку зростання значень.

2) Формулюємо гіпотези: H_0 – дані розподілені нормально, H_1 – розподіл відрізняється від нормального.

3) Робимо оцінку параметрів та стандартизацію (необхідно привести дані до масштабу стандартного нормального розподілу $N(0;1)$:

а) знаходимо вибіркові характеристики: середнє $\bar{x}_{\text{в}}$ і виправлене стандартне відхилення s за формулами:

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \tag{6.20}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2} \,. \tag{6.21}$$

б) знаходимо Z-перетворення: для кожного значення x_i розраховуємо стандартизоване значення $z_i = \dfrac{x_i - \bar{x}_{\text{в}}}{s}$. Тепер дані наведені до масштабу стандартного нормального розподілу $N(0;1)$.

Зауваження 6.7. Для розрахунку Z-оцінок використовується виправлене стандартне відхилення s, оскільки воно є незміщеною оцінкою параметра генеральної сукупності. Це критично важливо для коректного порівняння отриманої статистики D з табличними критичними значеннями Ліллієфорсу.

4) Знаходимо значення функцій розподілу.

а) теоретична функція $F_0(z)$: для кожного z_i вибірки знаходимо ймовірність ідеального нормального розподілу за формулою $F_0(z_i)=0{,}5+\Phi(z_i)$, де $\Phi(z)$ – значення функції Лапласа, яке знаходимо за таблицею А.2 (див. Додаток А);

б) емпірична функція $S_n(x)$: розраховуємо накопичену частоту.

• якщо значення x_i унікальні, то $S_n(x_i)=\frac{i}{n}$, де i – порядковий номер елемента x_i у відсортованій вибірці;

• якщо є повтори, то для групи однакових значень S_n дорівнює відношенню кількості всіх елементів, що не перевищують дане значення, до загального числа n.

5) Розраховуємо статистику Ліллієфорса (D)

Для кожного спостереження необхідно знайти відхилення у двох точках (на краях «сходинки»):

різниці «зверху»: $D_i^+ = \left| F_0(z_i) - \frac{i}{n} \right|$ та «знизу»: $D_i^- = \left| F_0(z_i) - \frac{i-1}{n} \right|$.

Підсумкова статистика D – це максимальне значення з усіх обчислених відхилень D_i^+ і D_i^- по всій таблиці.

6) За таблицею А.10 (див. Додаток А) знаходимо критичне значення $D_{кр}$ критерію Ліллієфорсу за заданим рівнем значущості α (зазвичай α=0,05) та обсягом вибірки n.

Якщо $D \le D_{кр}$, то гіпотеза H_0, що дані розподілені нормально, не відхиляється. Якщо $D > D_{кр}$, то гіпотеза H_0, відхиляється (розподіл не є нормальним).

Приклад 46. Виміряли зріст 10 студентів. Отримали дані (в см): 170,172,175,178,180,182,185,188,190,195. Перевірити гіпотезу, що зріст

студентів у вибірці розподілено за нормальним законом, використовуючи критерій Ліллієфорса з рівнем значущості α=0,05.

Розв'язання. Будемо перевіряти гіпотезу H_0: "Дані зростання студентів відповідають нормальному розподілу".

Маємо вибірку обсягу n=10, в якій дані розташовані в порядку зростання значень без повторення.

Знайдемо середнє $\bar{x}_{\text{в}}$ за формулою (6.20) і виправлене стандартне відхилення s за формулою (6.21).

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\cdot(170+172+175+178+180+182+185+188+190+$$
$$+195) = 181{,}5;$$

$$\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2 = (170-181{,}5)^2 + (172-181{,}5)^2 + (175-181{,}5)^2 +$$
$$+(178-181{,}5)^2 + (180-181{,}5)^2 + (182-181{,}5)^2 + (185-181{,}5)^2 +$$
$$+(188-181{,}5)^2 + (190-181{,}5)^2 + (195-181{,}5)^2 = 588{,}5;$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2} = \sqrt{\frac{1}{9}\cdot 588{,}5} \approx \sqrt{65{,}3889} \approx 8{,}0863.$$

Знаходимо значення емпіричної функції $S_n(x)$:

$$S_n(x_i) = \frac{i}{n}, \text{ де } i \text{ – порядковий номер елемента } x_i.$$

Знаходимо значення теоретичної функції $F_0(z)$ для кожного $z_i = \dfrac{x_i - \bar{x}_{\text{в}}}{s}$ вибірки: $F_0(z_i)$=0,5+$\Phi(z_i)$, де $\Phi(z)$ – значення функції Лапласа, яке знаходимо за таблицею А.2 (див. Додаток А).

Враховуючи, що n=10, $\bar{x}_{\text{в}} = 181{,}5$ і s≈8,0863, складаємо таблицю 6.12, де $z_i = \dfrac{x_i - 181{,}5}{8{,}0863}$ і $S_n(x_i) = \dfrac{i}{10}$.

Таблиця 6.12

i	x_i	z_i	$F_0(z_i)$	$S_n(x_i)$	$\lvert F_0(z_i)-S_n(x_i)\rvert$	$\lvert F_0(z_i)-S_n(x_{i-1})\rvert$
1	170	–1,422	0,0778	0,1	0,0222	0,0778
2	172	–1,175	0,1190	0,2	0,0810	0,0190
3	175	–0,804	0,2119	0,3	0,0881	0,0119
4	178	–0,433	0,3336	0,4	0,0664	0,0336
5	180	–0,185	0,4247	0,5	0,0753	0,0247
6	182	0,062	0,5239	0,6	0,0761	0,0239
7	185	0,433	0,6664	0,7	0,0336	0,0664
8	188	0,804	0,7881	0,8	0,0119	0,0881
9	190	1,051	0,8531	0,9	0,0469	0,0531
10	195	1,669	0,9525	1,0	0,0475	0,0525

За визначенням статистика D критерію Ліллієфорса – це максимальне абсолютне відхилення між теорією та практикою. Переглянувши два останні стовпці, знаходимо, що максимальне значення досягається у кількох точках, але найбільші: D=0,0881.

За таблицею А.10 (див. Додаток А) за заданим рівнем значущості α=0,05 та обсягом вибірки n=10 знаходимо критичне значення критерію Ліллієфорсу $D_{кр}\approx$0,262.

Оскільки розрахункове значення D=0,0881 значно менше критичного $D_{кр}$=0,262, то немає підстав відхилити нульову гіпотезу H_0.

Відповідь: Дані зросту студентів відповідають нормальному розподілу.

6.3 Перевірка параметричних статистичних гіпотез

Параметрична гіпотеза – це статистична гіпотеза, яка робить припущення щодо параметрів відомого або передбачуваного розподілу генеральної сукупності (наприклад, про математичне сподівання a, дисперсію σ^2 або частку p).

Для перевірки параметричних гіпотез необхідно, щоб дані відповідали певному закону розподілу (найчастіше нормальному розподілу).

Параметричні гіпотези зазвичай формулюються як пара:

Нульова гіпотеза H_0: Основне припущення, яке перевіряється. Зазвичай формулюється як твердження про відсутність відмінностей чи рівність параметра деякому значенню.

Альтернативна (конкуруюча) гіпотеза H_1: Гіпотеза, що приймається у разі відхилення нульової гіпотези. Формулюється як твердження про наявність відмінностей чи нерівність.

Наприклад, (таблиця 6.13)

Таблиця 6.13

Тип параметра, що перевіряється	Нульова гіпотеза H_0	Альтернативна гіпотеза H_1
Про середнє a	$a=a_0$ (середнє дорівнює заданому значенню)	$a\neq a_0$ (двостороння) або $a>a_0$ (правостороння), або $a<a_0$ (лівостороння)
Про дисперсію σ^2	$\sigma^2=\sigma_0^2$ (дисперсія дорівнює заданому значенню)	$\sigma^2\neq\sigma_0^2$ або $\sigma^2>\sigma_0^2$, або $\sigma^2<\sigma_0^2$
Про рівність двох середніх	$a_1=a_2$ (середні у двох сукупностях рівні)	$a_1\neq a_2$ або $a_1>a_2$, або $a_1<a_2$

Перевірка параметричної гіпотези – це процес прийняття статистично обґрунтованого рішення про відхилення або ухвалення нульової гіпотези на основі даних вибірки.

Перевірка параметричних гіпотез здійснюється за наступною схемою:

1. Формулювання гіпотез. Спочатку формулюються гіпотези H_0 і H_1 щодо параметра, який перевіряється.

2. Вибір рівня значущості α. Рівень значущості – це максимально припустима ймовірність помилки I роду (відхилити гіпотезу H_0, коли вона вірна). Найчастіше використовувані значення: α=0,05 (5%), α=0,01 (1%).

3. Вибір статистичного критерію. Вибирається спеціальна статистика (наприклад, Z-критерій, t-критерій Стьюдента, F-критерій Фішера, χ^2-критерій), розподіл якої відомий і залежить від параметра, що перевіряється, і H_0. Вибір критерію залежить від: типу гіпотези (про середнє, про дисперсію тощо); виду розподілу (передбачається нормальний); обсяг вибірки.

4. Розрахунок спостережуваного значення критерію $T_{спост}$ За даними вибірки розраховується фактичне (спостережуване) значення обраного статистичного критерію.

5. Визначення критичної області та критичного значення $T_{кр}$. На основі обраного рівня значущості α та розподілу критерію визначається критична область (область, де гіпотеза H_0 відхиляється). Межами цієї області є критичні значення $T_{кр}$.

6. Прийняття рішення. Порівнюється значення критерію $T_{спост}$ з критичними значеннями $T_{кр}$:

якщо $T_{спост}$ потрапляє у критичну область (або $|T_{спост}|>T_{кр}$), то гіпотеза H_0 відхиляється на користь гіпотези H_1;

якщо $T_{спост}$ не потрапляє в критичну область, то немає підстав відхилити гіпотезу H_0.

6.3.1 Перевірка правильності нульової гіпотези H_0 про значення генеральної середньої

Перевірка гіпотези про генеральну середню дозволяє зрозуміти, чи є отриманий результат закономірністю або випадковим.

Загальний алгоритм перевірки гіпотези:

1. Формулювання гіпотез.

Нульова гіпотеза H_0 стверджує, що середня a дорівнює деякому еталонному значенню a_0, тобто H_0: $a=a_0$.

Альтернативна гіпотеза H_1 суперечить нульовій гіпотезі H_0 і може бути двосторонньою: $a\neq a_0$ або односторонньою: $a>a_0$ чи $a<a_0$.

2. Вибір рівня значущості α. Це ймовірність зробити помилку першого роду (відкинути правильну нульову гіпотезу). Найчастіше обирають α=0,05 (5%), α=0,01 (1%).

3. Вибір статистичного критерію. Вибір залежить від того, чи відома дисперсія σ^2 генеральної сукупності:

а) застосовуємо Z-критерій, якщо дисперсія σ^2 відома для всього процесу або обсяг вибірки великий $(n>30)$ і дисперсія σ^2 невідома (σ замінюємо на s):

$$Z_{\text{спост}} = \frac{\bar{x}_{\text{в}} - a_0}{\sigma / \sqrt{n}}. \qquad (6.22)$$

Зауваження 6.8. Якщо обсяг вибірки $n<30$, використовувати Z-критерій можна тільки в одному випадку, якщо впевнені, що генеральна сукупність розподілена нормально, і точно відома генеральна дисперсія σ^2. Однак на практиці σ^2 майже ніколи не відома. Тому, якщо $n<30$, стандартним рішенням є перехід до t-критерію Стьюдента.

б) застосовуємо t-критерій Стьюдента, якщо дисперсія σ^2 невідома (що найчастіше) і вибірка мала ($n\leq 30$):

$$t_{\text{спост}} = \frac{\bar{x}_{\text{в}} - a_0}{s / \sqrt{n}}. \qquad (6.23)$$

Якщо стандартне вибіркове відхилення s невідоме, то його необхідно розрахувати за наявними даними вибірки за формулою (6.21).

4. Визначення критичної області. За таблицями (Z чи t) знаходимо критичне значення для обраного рівня значущості α та степенів свободи $k=n-1$: $Z_{\text{кр}}$ чи $t_{\text{кр}}$.

1) Знаходження критичного значення $Z_{\text{кр}}$.

Найчастіші значення $Z_{\text{кр}}$ для двосторонніх гіпотез через функцію Лапласа беремо з таблиці 6.14:

Таблиця 6.14

Рівень значущості α	Критичне значення $Z_{\text{кр}}$
0,10 (10%)	1,645
0,05 (5%)	1,96
0,01 (1%)	2,576

Для односторонньої гіпотези правило пошуку $Z_{\text{кр}}$ таке:

а) знаходимо значення $\Phi(Z_{\text{кр}})=0,5-\alpha$;

б) шукаємо отримане значення всередині таблиці А.2 (див. Додаток А) та дивимося, якому аргументу x воно відповідає. Це і буде

критичне значення $Z_{кр}$. Якщо критична область знаходиться з лівої сторони, беремо значення зі знаком «мінус».

Найчастіші значення $Z_{кр}$ для односторонніх гіпотез через функцію Лапласа беремо з таблиці 6.15:

Таблиця 6.15

Рівень значущості α	$\Phi(Z_{кр})=0,5-\alpha$	Критичне значення $Z_{кр}$
0,10 (10%)	0,4	1,28
0,05 (5%)	0,45	1,645
0,01 (1%)	0,49	2,33

2) Знаходження критичного значення $t_{кр}$.

Для двосторонньої гіпотези знаходимо $t_{кр}$ за таблицею А.4 (див. Додаток А) з числом степенів свободи $k=n-1$ і заданим рівнем значущості α.

Для односторонньої гіпотези, якщо критична область знаходиться з лівої сторони, то знаходимо $t_{кр}$ за таблицею А.4 (див. Додаток А) з числом степенів свободи $k=n-1$ та заданим рівнем значущості α і беремо його зі знаком «мінус».

5. Якщо розрахункове значення критерію потрапляє у критичну область (зону відхилення), гіпотеза H_0 відхиляється; якщо розрахункове значення перебуває у довірчому інтервалі, гіпотеза H_0 не відхиляється.

Зауваження 6.9. Не знаючи вибіркове стандартне відхилення s (або без генерального σ) перевірити гіпотезу про середню неможливо.

Якщо потрібно не просто відповісти на запитання «так чи ні» (відкидаємо гіпотезу чи ні), а оцінити справжнє значення параметра з певною точністю, то застосовують підхід із побудовою довірчого інтервалу.

Цей підхід кращий за перевірки гіпотез, коли потрібна оцінка величини (наприклад, за гіпотезою відповідаємо «так чи ні», а довірчий інтервал вказує з заданою надійністю належність середньої «від і до», що дає набагато більше інформації для ухвалення рішень); коли потрібно оцінити точність виміру: чим вужчий інтервал, тим точніше дані, а якщо інтервал величезний, то вибірка дуже мала або дані занадто розкидані; при довгостроковому моніторингу (наприклад, в економіці

чи медицині часто дивляться, як довірчий інтервал зміщується зі часом).

Довірчий інтервал зручний для наукових досліджень та прогнозування, оскільки він показує діапазон можливих значень та степінь нашої невизначеності.

Загальний алгоритм побудови довірчого інтервалу.

1) *Вибір точкової оцінки*: Розраховуємо значення за вибіркою (наприклад, вибіркове середнє $\bar{x}_{\text{в}}$). Це центр інтервалу.

2) *Встановлення надійності* γ: Задаємо ймовірність (зазвичай 0,95 або 0,99).

3) *Визначення закону розподілу*: Вирішуємо, що використовувати Z (нормальне) або t (Стьюдента). Знаходимо критичну точку. За таблицею (Стьюдента чи Лапласа) знаходимо значення $t_{\text{кр}}$ або $Z_{\text{кр}}$ для обраної надійності γ та степенів свободи $k=n-1$.

4) *Розрахунок граничної помилки* Δ: Знаходимо «розмах» інтервалу:

якщо σ відоме (або $n>30$), то

$$\Delta = Z_{\text{кр}} \cdot \frac{\sigma}{\sqrt{n}}; \tag{6.24}$$

якщо σ невідоме (вибірка мала), то

$$\Delta = t_{\text{кр}} \cdot \frac{s}{\sqrt{n}}, \tag{6.25}$$

де s– вибіркове стандартне відхилення.

5) *Будуємо довірчий інтервал*:

якщо перевіряється двостороння гипотеза, то довірчий інтервал $\left[\bar{x}_{\text{в}} - \Delta; \bar{x}_{\text{в}} + \Delta\right]$;

якщо перевіряється гіпотеза «більше ніж» $a>a_0$, то довірчий інтервал $\left[\bar{x}_{\text{в}} - \Delta; +\infty\right)$;

якщо перевіряється гіпотеза «менше ніж» $a<a_0$, то довірчий інтервал $\left(-\infty; \bar{x}_{\text{в}} + \Delta\right]$.

6) *Приймаємо рішення*:

якщо $a_0 \in \left[\bar{x}_в - \Delta; \bar{x}_в + \Delta\right]$, то гіпотеза H_0 приймається. Відмінність між середнім та нормою випадкова (для односторонніх гіпотез аналогічно);

якщо $a_0 \notin \left[\bar{x}_в - \Delta; \bar{x}_в + \Delta\right]$, то гіпотеза H_0 відхиляється. Відмінність статистично значуща (для односторонніх гіпотез аналогічно).

Для побудови довірчих інтервалів при використанні закону розподілу Z (нормальне), можна користуватись таблицею 6.16 і таблицею 6.17:

Таблиця 6.16

Тип гіпотези	Інтервал	Формула
$a \neq a_0$ (двостороння)	двосторонній	$\bar{x}_в \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
$a > a_0$ (правостороння)	односторонній «більше ніж»	$\left[\bar{x}_в - Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}; +\infty\right)$
$a < a_0$ (лівостороння)	односторонній «менше ніж»	$\left(-\infty; \bar{x}_в + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right]$

Таблиця 6.17

Надійність	Рівень значущості α	$Z_{\alpha/2}$ (Двосторонній)	Z_α (Односторонній)
90%	0,10 (10%)	1,645	1,28
95%	0,05 (5%)	1,96	1,645
99%	0,01 (1%)	2,576	2,33

Зауваження 6.10. Коли використати той чи інший метод? Якщо поставлена мета прийняти рішення, то використовуємо перевірку гіпотез (H_0) і результат буде виглядати так: "Відмінність не значуща, H_0 не відхилена". Якщо поставлена мета оцінити реальність, то використовуємо довірчий інтервал і результат буде виглядати так: "Справжнє значення лежить між «нижня межа» та «верхня межа» довірчого інтервалу".

Приклад 47. Завод випускає деталі, середня довжина яких має бути a_0=100 мм. Відомо, що стандартне відхилення процесу σ=2 мм. Перевірили 100 деталей та отримали середню довжину $\bar{x}_в = 100{,}5$ мм. Чи свідчить це про збій у налаштуваннях верстата (прийняти рівень значущості α=0,05)?

Розв'язання. Сформулюємо гіпотези: H_0: a=100 ("Середня довжина деталі відповідає стандарту і дорівнює 100 мм. Відхилення, що спостерігається, в 0,5 мм випадково"), H_1: $a \neq 100$ ("Середня довжина деталі значно відрізняється від стандарту. Устаткування вийшло з ладу" – двостороння гіпотеза).

Маємо вибірку обсягу n=100>30 і дисперсія σ^2 відома (σ^2=4). За формулою (6.22) розраховуємо $Z_{спост}$:

$$Z_{спост} = \frac{\bar{x}_в - a_0}{\sigma / \sqrt{n}} = \frac{100{,}5 - 100}{2/\sqrt{100}} = \frac{0{,}5}{0{,}2} = 2{,}5 .$$

Для рівня значущості α=0,05 за таблицею 6.14 визначаємо $Z_{кр}$=1,96. Оскільки $Z_{спост} > Z_{кр}$ (2,5>1,96), то гіпотезу H_0 відхиляємо, тобто верстат потрібно переналаштувати.

Відповідь: Верстат потрібно переналаштувати.

Приклад 48. Фермер стверджує, що середній врожай деякого сорту пшениці становить a_0=40 ц/га. Дослідник посіяв цей сорт на 16 ділянках та отримав середню врожайність $\bar{x}_в = 38$ ц/га при вибірковому виправленому стандартному відхиленні s=4 ц/га. Чи справді середня врожайність нижча за заявлену (прийняти рівень значущості α=0,05)?

Розв'язання. Сформулюємо гіпотези: H_0: a=40 ("Врожайність відповідає заявленій"), H_1: a<40 ("Врожайність значно нижча за заявлену" – лівостороння гіпотеза).

Оскільки обсяг вибірки малий n=16<30 і дисперсія σ^2 невідома, то застосовуємо t-критерій Стьюдента. За формулою (6.23) розраховуємо $t_{спост}$:

$$t_{спост} = \frac{\bar{x}_в - a_0}{s / \sqrt{n}} = \frac{38 - 40}{4/\sqrt{16}} = \frac{-2}{1} = -2 .$$

Для односторонньої гіпотези (H_1: a<40) критична область знаходиться лише з лівої сторони. За таблицею А.4 (див. Додаток А) з числом степенів свободи $k=n-1=16-1=15$ та α=0,05 знаходимо

критичне значення $t_{кр}$=1,753. Оскільки перевіряємо лівосторонню гіпотезу, беремо його зі знаком «мінус»: $t_{кр}$=−1,753.

Отже $t_{спост}<t_{кр}$ (−2<−1,753), тому гіпотезу H_0 відхиляємо, значення потрапило до критичної області. Фермер перебільшив показники.

Відповідь: Середня врожайність значно нижча ніж заявлена фермером 40 ц/га.

Приклад 49. Виробник заявляє, що середній термін служби нової моделі лампочок становить a_0=1000 годин. Перевірили 25 ламп та отримали вибіркове середнє $\bar{x}_в = 980$ годин, вибіркове стандартне відхилення s=50 годин. Відомо, що термін служби лампочок має нормальний розподіл. Перевірити твердження виробника (прийняти рівень значущості α=0,05).

Розв'язання. Сформулюємо гіпотези: H_0: a=1000 ("Середній термін служби дорівнює 1000 годин"), H_1: $a\neq1000$ ("Середній термін служби відрізняється від 1000 годин" – двостороння гіпотеза).

Оскільки перевіряємо середнє при невідомій дисперсії σ^2 генеральної сукупності та малому обсязі вибірки (n=25<30), застосовуємо t-критерій Стьюдента.

За формулою (6.23) розраховуємо $t_{спост}$:

$$t_{спост} = \frac{\bar{x}_в - a_0}{s/\sqrt{n}} = \frac{980-1000}{50/\sqrt{25}} = \frac{-20}{50/5} = \frac{-20}{10} = -2.$$

За таблицею А.4 (див. Додаток А) для двосторонньої критичної області з числом степенів свободи $k=n-1=25-1=24$ та α=0,05знаходимо критичне значення $t_{кр}$=2,064.

Порівнюємо абсолютне значення $t_{спост}$ з $t_{кр}$: $|t_{спост}|=|-2|=2$ і $t_{кр}$=2,064. Оскільки 2<2,064, то $t_{спост}$ не потрапляє в критичну область, тому не можемо відхилити нульову гіпотезу H_0.

Відповідь: На рівні значущості 5% немає достатніх підстав стверджувати, що середній термін служби лампочок відрізняється від 1000 годин.

Приклад 50. За даними кондитерської фабрики середня вага плитки шоколаду становить a_0=100 грамів. Купили n=10 плиток шоколаду та зважили їх на точних терезах. Отримали дані (у грамах):

98, 102, 97, 99, 101, 96, 100, 98, 97, 99. Перевірити чи відповідає вага нормі (прийняти рівень значущості α=0,05).

Розв'язання. *1 метод: використовуємо перевірку гіпотези*. Сформулюємо гіпотези: H_0: a=100 ("Вага відповідає нормі"), H_1: $a\neq100$ ("вага відрізняється від норми, двостороння перевірка" – двостороння гіпотеза).

Знайдемо вибіркові характеристики: середнє $\bar{x}_{\text{в}}$ і виправлене стандартне відхилення s за формулами (6.20) і (6.21) відповідно:

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\left(98+102+97+99+101+96+100+98+97+99\right) = 98{,}7;$$

$$\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2 = (98-98{,}7)^2 + (102-98{,}7)^2 + (97-98{,}7)^2 + (99-98{,}7)^2 +$$
$$+ (101-98{,}7)^2 + (96-98{,}7)^2 + (100-98{,}7)^2 + (98-98{,}7)^2 + (97-98{,}7)^2 +$$
$$+ (99-98{,}7)^2 = 32{,}1;$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2} = \sqrt{\frac{1}{9}\cdot 32{,}1} \approx \sqrt{3{,}5667} \approx 1{,}8886.$$

Оскільки обсяг вибірки малий n=10<30 і точна дисперсія всієї фабрики невідома, використовуємо t-критерій Стьюдента. За формулою (6.23) розраховуємо $t_{\text{спост}}$:

$$t_{\text{спост}} = \frac{\bar{x}_{\text{в}} - a_0}{s/\sqrt{n}} = \frac{98{,}7-100}{1{,}8886/\sqrt{10}} = \frac{-1{,}3}{0{,}5972} \approx -2{,}18.$$

За таблицею А.4 (див. Додаток А) для двосторонньої критичної області з числом степенів свободи k=n−1=10−1=9 та α=0,05 знаходимо критичне значення $t_{\text{кр}}$=2,262.

Порівнюємо абсолютне значення $t_{\text{спост}}$ з $t_{\text{кр}}$: $|t_{\text{спост}}|=|-2{,}18|=2{,}18$ і $t_{\text{кр}}$=2,262. Отже $|t_{\text{спост}}|<t_{\text{кр}}$ (2,18<2,262). Розрахункове значення не потрапило до критичної області (хоча було дуже близько).

Відповідь: Недостатньо підстав, щоб відхилити нульову гіпотезу. Незважаючи на те, що середня вага в нашій вибірці $\bar{x}_{\text{в}} = 98{,}7$ менша за 100, статистично це відхилення при такій маленькій вибірці може бути випадковим. Фабрика «виправдана».

2 метод: використовуємо довірчий інтервал. Скористаємося даними, отриманими вище: $\bar{x}_{\text{в}} = 98{,}7$; s=1,8886; n=10; $t_{\text{кр}}$=2,262 (знайдене за таблицею А.3 (див. Додаток А) для $k=n-1=10-1=9$ і $\gamma=1-\alpha=1-0{,}05=0{,}95$).

Знаходимо граничну помилку Δ вибірки за формулою (6.25), оскільки σ невідоме і вибірка мала.

$$\Delta = t_{\text{кр}} \cdot \frac{s}{\sqrt{n}} = 2{,}262 \cdot \frac{1{,}8886}{\sqrt{10}} \approx 2{,}262 \cdot 0{,}5972 \approx 1{,}35 .$$

Будуємо довірчий інтервал: $\left[\bar{x}_{\text{в}} - \Delta; \bar{x}_{\text{в}} + \Delta\right]$. Маємо

$$(98{,}7 - 1{,}35; 98{,}7 + 1{,}35) \Rightarrow (97{,}35; 100{,}05).$$

Оскільки гіпотетичне значення a_0=100 потрапляє всередину довірчого інтервалу, то нульова гіпотеза H_0 приймається.

Відповідь: На рівні надійності 95% немає достатніх підстав звинувачувати фабрику в обмані покупців.

Приклад 51. Відомо, що на виробництві батарейок старе обладнання видає середній заряд a_0=100 одиниць із стандартним відхиленням σ=5. Купили нову деталь для верстата щоб збільшила середній заряд. Перевірили n=25 батарейок і отримали середнє $\bar{x}_{\text{в}} = 103$ одиниці. Перевірити ефективність придбаної деталі з рівнем значущості α=0,05 та визначити діапазон з надійністю 95%, в якому тепер знаходиться заряд насправді.

Розв'язання. Сформулюємо гіпотези: H_0: a=100 ("Середній заряд дорівнює 100 одиниць"), H_1: a>100 ("Середній заряд більше 100 одиниць"– правостороння гіпотеза).

Оскільки обсяг вибірки малий n=25<30 і дисперсія σ^2 відома (σ^2=25). За формулою (6.22) розраховуємо $Z_{\text{спост}}$:

$$Z_{\text{спост}} = \frac{\bar{x}_{\text{в}} - a_0}{\sigma / \sqrt{n}} = \frac{103 - 100}{5 / \sqrt{25}} = \frac{3}{1} = 3 .$$

Для рівня значущості α=0,05 за таблицею (табл. 6.15), оскільки гіпотеза одностороння, визначаємо $Z_{\text{кр}}$=1,645. Маємо $Z_{\text{спост}} > Z_{\text{кр}}$ (3>1,645), тому гіпотезу H_0 відхиляємо. Заряд виріс.

Знайдемо діапазон, в якому насправді тепер знаходиться заряд з надійністю γ=0,95 (95%). Довірчий інтервал односторонній (нижня

межа), оскільки нас цікавить, наскільки мінімум зріс заряд. Знаходимо нижню межу довірчого інтервалу.

Для цього за формулою (6.24) знаходимо граничну помилку Δ:

$$\Delta = Z_{\text{кр}} \cdot \frac{\sigma}{\sqrt{n}} = 1{,}645 \cdot \frac{5}{\sqrt{25}} = 1{,}645 .$$

Будуємо довірчий інтервал довкола свого вибіркового середнього $\bar{x}_{\text{в}} = 103 : \left[\bar{x}_{\text{в}} - \Delta; +\infty\right)$. Маємо

$$\left[103 - 1{,}645; +\infty\right) \Rightarrow \left[101{,}355; +\infty\right).$$

На 95% упевнені, що після модернізації верстатів справжній середній заряд батарей становить не менше 101,355 одиниць.

Оскільки старе значення a_0=100 не потрапляє до цього інтервалу (воно менше нижньої межі), то підтверджуємо висновок: зміни статистично значущі, і заряд реально зріс.

Відповідь: Нульову гіпотезу H_0 відхиляємо. Заряд батарейки виріс і знаходиться в діапазоні $\left[101{,}355; +\infty\right)$.

6.3.2 Порівняння двох середніх генеральних сукупностей

Часто зустрічаються ситуації, коли середнє значення даних одного експерименту відрізняється від середнього значення даних іншого експерименту, що проводиться за тих самих умов. Тоді виникає питання, чи можна вважати цю розбіжність незначною, тобто суто випадковою, або вона викликана суттєвою відмінністю двох генеральних сукупностей.

Вибір критерію для перевірки залежить від того, що відомо про генеральні сукупності та який обсяг даних. Якщо відомі дисперсії генеральних сукупностей, то застосовують *Z*-критерій (наприклад, у високоточному виробництві). Якщо генеральні дисперсії невідомі (найчастіший випадок), то застосовується *t*-критерій Стьюдента.

1. Застосування Z-тесту для перевірки гіпотези про рівність середніх значень

Застосування *Z*-тесту обґрунтовано у таких випадках:

1) Відомі дисперсії при будь-якому обсязі вибірки.

Якщо дисперсії σ_1^2 і σ_2^2 генеральних сукупностей відомі заздалегідь (з теоретичних моделей, стандартів або минулих

масштабних досліджень), то *Z*-тест є найбільш точним інструментом.

Зауваження 6.11. При малому обсязі вибірки ($n<30$) обов'язковою умовою є нормальний розподіл досліджуваної ознаки у генеральній сукупності. У цьому випадку *Z*-статистика суворо дотримується стандартного нормального розподілу.

2) Велика вибірка (незалежно від розподілу)

При обсягу вибірки $n\geq 30$ (у деяких джерелах $n\geq 50$ або $n\geq 100$) використання *Z*-тесту допустиме навіть у тому випадку, якщо дисперсія генеральної сукупності невідома, а вихідний розподіл даних відрізняється від нормального. Оскільки розподіл вибіркових середніх наближається до нормального зі збільшенням обсягу вибірки, то вибіркова дисперсія s^2 стає досить точною оцінкою генеральної дисперсії σ^2, що дозволяє використовувати критичні значення нормального розподілу.

Зазвичай при перевірці рівності середніх двох генеральних сукупностей гіпотези записуються так:

а) *Двостороння гіпотеза*.

Нульова гіпотеза H_0: $a_1=a_2$ ("Середні генеральних сукупностей рівні, значних відмінностей немає").

Альтернативна гіпотеза H_1: $a_1\neq a_2$ ("Середні не рівні" – двостороння).

б) *Правостороння гіпотеза (зростання показника)*.

Нульова гіпотеза H_0: $a_1=a_2$ або $a_1\leq a_2$.

Альтернативна гіпотеза H_1: $a_1>a_2$.

в) *Лівостороння гіпотеза (зниження показника)*.

Нульова гіпотеза H_0: $a_1=a_2$ або $a_1\geq a_2$.

Альтернативна гіпотеза H_1: $a_1<a_2$.

Формула *Z*-статистики при порівняння двох незалежних вибірок має вигляд:

$$Z_{\text{спост}}=\frac{\bar{x}_{\text{в}1}-\bar{x}_{\text{в}2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}, \qquad (6.26)$$

де n_1, n_2 – обсяги вибірок; $\bar{x}_{в1}$, $\bar{x}_{в2}$ – вибіркові середні двох вибірок (задані за умовою або знаходяться за формулою (6.20) для кожної вибірки); σ_1^2, σ_2^2 – дисперсії генеральних сукупностей відомі заздалегідь або при $n \geq 30$, якщо вони невідомі за умовою, то знаходимо вибіркова дисперсія s_1^2, s_2^2 за формулою $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2$ для кожної вибірки і знайдені значення беремо за σ_1^2, σ_2^2.

Після розрахунку значення $Z_{спост}$ його потрібно порівняти з критичним значення $Z_{кр}$, яке береться з таблиці нормального розподілу для обраного рівня значущості α. Можна для двосторонньої гіпотези користуватись таблицею 6.14, а для односторонньої гіпотези користуватись таблицю 6.15.

Якщо $|Z_{спост}| > Z_{кр}$, то нульова гіпотеза H_0 відхиляється на користь альтернативної гіпоьези H_1. Відмінності статистично значущі.

Якщо $|Z_{спост}| \leq Z_{кр}$, то нульова гіпотеза H_0 не відхиляється. Відмінності вважаються випадковими.

Таблиця порівняння критичних значень для α=0,05 (табл. 6.18).

Таблиця 6.18

Тип гіпотези H_1	Критичне значення $Z_{кр}$	Область відхилення H_0
$a_1 \neq a_2$ (двостороння)	1,96	$Z_{спост} > 1,96$ або $Z_{спост} < -1,96$
$a_1 > a_2$ (правостороння)	1,645	$Z_{спост} > 1,645$
$a_1 < a_2$ (лівостороння)	−1,645	$Z_{спост} < -1,645$

Приклад 52. Є дві групи студентів, які займаються за різними методиками. У першій групі n_1=20 студентів і середній бал $\bar{x}_{в1} = 82$, у другій групі n_2=18 студентів і середній бал $\bar{x}_{в2} = 78$. З багаторічної статистики відомо, що дисперсія балів за таких тестів однакова $\sigma_1^2 = \sigma_2^2 = 25$ для обох методик. Обидві генеральні сукупності

розподілені нормально. За рівнем значущості α=0,05 перевірити гіпотезу, що середні бали груп однакові.

Розв'язання. Для перевірки гіпотези про рівність середніх балів обираємо *Z*-критерій, попри малий обсяг вибірок (n_1=20, n_2=18). Цей вибір методологічно обґрунтований тим, що параметри генеральної дисперсії σ^2 є відомими величинами, встановленими з багаторічної статистики. Додатковою умовою легітимності тесту є прийняте припущення про нормальний розподіл досліджуваної ознаки в обох генеральних сукупностях.

Сформулюємо гіпотези: H_0: $a_1=a_2$ ("Середні бали груп однакові"), H_1: $a_1\neq a_2$ ("Середні бали розрізняються" – двостороння гіпотеза).

Знайдемо $Z_{\text{спост}}$ за формулою (6.26).

$$Z_{\text{спост}}=\frac{82-78}{\sqrt{\frac{25}{20}+\frac{25}{18}}}\approx\frac{4}{\sqrt{1{,}25+1{,}3889}}\approx\frac{4}{\sqrt{2{,}6389}}\approx\frac{4}{1{,}6245}\approx 2{,}4623.$$

За таблицею 6.18 знаходимо критичне значення $Z_{\text{кр}}$ для двосторонньої гіпотези з рівнем значущості α=0,05: $Z_{\text{кр}}$=1,96.

Оскільки виконується умова $Z_{\text{спост}}>Z_{\text{кр}}$ (2,4623>1,96), то нульова гіпотеза H_0 відхиляється. Різниця в 4 бали між методиками є статистично значущою.

Відповідь: Середні бали розрізняються. Різниця у 4 бали між методиками є статистично значущою.

Приклад 53. Необхідно дослідити міцність двох видів ниток. Отримали вибірку для першого виду ниток: 15,18,16,17,19 (n_1=5), а для другого виду ниток: 14,13,15,14 (n_2=4). Відповідно до нормативу, дисперсія міцності таких ниток завжди дорівнює σ^2=2. Прийняте припущення про нормальний розподіл досліджуваної ознаки обох генеральних сукупностей. Перевірити, чи є значущі відмінності між середніми при рівні значущості α=0,05.

Розв'язання. Заздалегідь припускаємо, що нитки першого виду міцніші, ніж другого. Сформулюємо гіпотези: H_0: $a_1\leq a_2$ ("Міцність ниток першого виду не вище міцності ниток другого виду"), H_1: $a_1>a_2$ ("Нитки першого виду значно міцніші").

Розрахуємо середні значення за вибірками:

$$\bar{x}_{\text{в1}}=\frac{15+18+16+17+19}{5}=\frac{85}{5}=17,\ \bar{x}_{\text{в2}}=\frac{14+13+15+14}{4}=\frac{56}{4}=14.$$

Знайдемо $Z_{\text{спост}}$ за формулою (6.26).

$$Z_{\text{спост}} = \frac{17-14}{\sqrt{\frac{2}{5}+\frac{2}{4}}} = \frac{3}{\sqrt{0{,}4+0{,}5}} = \frac{3}{\sqrt{0{,}9}} \approx \frac{3}{0{,}9487} \approx 3{,}1622.$$

За таблицею 6.18 знаходимо критичне значення $Z_{\text{кр}}$ для односторонньої (правосторонньої) гіпотези з рівнем значущості α=0,05 маємо $Z_{\text{кр}}$=1,645.

Оскільки виконується умова $Z_{\text{спост}} > Z_{\text{кр}}$ (3,1622>1,645), то нульова гіпотеза H_0 відхиляється. Підтверджуємо, що нитки першого виду значно міцніші за нитки другого виду.

Відповідь: Нитки першого виду значно міцніші за нитки другого виду.

Приклад 54. У ході клінічних випробувань досліджувався вплив двох препаратів (*A* та *B*) на рівень важливого біомаркера у крові. У групі *A* (n_1=100 пацієнтів) середній показник склав $\bar{x}_{\text{в1}} = 2500$ од. зі стандартним відхиленням s_1=500. У групі *B* (n_2=120 пацієнтів) середній показник склав $\bar{x}_{\text{в2}} = 2700$ од. зі стандартним відхиленням s_2=600. Для рівня значущості α=0,01 перевірити, чи препарат *B* не більш ефективний за препарат *A*.

Розв'язання. Сформулюємо гіпотези: H_0: $a_1 \geq a_2$ ("Препарат *B* не більш ефективний за препарат *A*"), H_1: $a_1 < a_2$ ("Препарат *B* ефективніший за препарат *A*").

Оскільки вибірки великі (n_1>30, n_2>30), використовуємо *Z*-критерій. Знайдемо $Z_{\text{спост}}$ замінивши в формулі (6.26) σ_1^2, σ_2^2 на s_1^2, s_2^2:

$$Z_{\text{спост}} = \frac{\bar{x}_{\text{в2}} - \bar{x}_{\text{в1}}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{2700-2500}{\sqrt{\frac{500^2}{100}+\frac{600^2}{120}}} = \frac{200}{\sqrt{5500}} \approx \frac{200}{74{,}162} \approx 2{,}6968.$$

За таблицею 6.17 знаходимо критичне значення $Z_{\text{кр}}$ для односторонньої (правосторонньої) гіпотези з рівнем значущості α=0,01: $Z_{\text{кр}}$=2,33.

Оскільки виконується умова $Z_{\text{спост}} > Z_{\text{кр}}$ (2,698>2,33), то нульова гіпотеза H_0 відхиляється. Препарат *B* значно кращий за препарат *A* з надійністю 99% .

Відповідь: Препарат *B* значно кращий за препарат *A* з надійністю 99%.

2. Застосування t-тесту для перевірки гіпотези про рівність середніх значень

t-критерій Стьюдента (найчастіший випадок) використовується, коли загальні дисперсії невідомі (оцінюємо їх за вибіркою як s_1^2 і s_2^2). Якщо вибірка мала (n<30) і умови застосування *Z*-статистики не виконані, то використання *t*-критерію Стьюдента обов'язкове.

При перевірці гіпотези про рівність середніх двох незалежних вибірок із невідомими теоретичними дисперсіями першочерговим завданням є оцінка однорідності їх вибіркових аналогів.

Оскільки дисперсії у генеральних сукупностях невідомі, тому використовуються виправлені вибіркові дисперсії s_1^2 і s_2^2. Від їх значень залежить вибір типу *t*-критерія. Розглянемо типи *t*-критеріїв (табл. 6.19).

Таблиця 6.19

Ситуація	Тип тесту	Особливості
Порівняння "До" і "Після" (одна група)	Парний *t*-тест	Аналізуємо різницю парних значень.
Дві різні групи (дисперсії рівні)	*t*-тест Стьюдента	Використовується об'єднана оцінка дисперсії.
Дві різні групи (дисперсії не рівні)	*t*-тест Уелча	Найнадійніший та універсальний метод.

1) *Парний t-тест*.

Коли використовується парний тест?

а) У медицині та фармакології – це основний інструмент клінічних випробувань (наприклад, перевірка мазі від алергії: на ліву руку пацієнта наносять препарат, на праву – нешкідливу речовину, приготовлену у вигляді ліків).

Чому парний? В однієї людини реакція шкіри може бути бурхливою, в іншої – слабкою. Якщо порівнювати "групу *A*" і

"групу *B*", ці відмінності все заплутають. Але порівнюючи дві руки однієї людини, ми бачимо чистий ефект мазі.

б) В IT та UX-дослідження (*A*/*B* тести всередині користувача). Наприклад, вимірювання швидкості роботи у двох версіях інтерфейсу.

Чому парний? Один користувач – "профі" і клікає швидко, інший – новачок. Якщо дамо «профі» версію *A*, а новачкові версію *B*, то не зрозуміємо, що краще. А якщо дамо кожному спробувати обидві (у різному порядку), тоді порівняємо їхній особистий прогрес.

в) В економіці та соціології (наприклад, вплив курсів підвищення кваліфікації на зарплату).

Чому парний? Ми порівнюємо зарплату одного й того ж чоловіка до курсів та після них. Це набагато точніше, ніж порівнювати різних людей.

Коли не можна застосувати парний тест?

Парний тест неможливий, якщо "вплив" не можна відіграти назад або застосувати паралельно (наприклад, не можна перевірити вплив статі на інтелект (людина не може бути одночасно і чоловіком, і жінкою немає пари); не можна перевірити ефективність двох методів навчання на тій самій людині, оскільки перший метод вже дав якісь знання – ефект перенесення).

Парний *t*-тест застосовується, коли досліджуємо один і той же об'єкт двічі ("до" і "після"). Наприклад, при дослідженні впливу нового препарату досліджують одну й ту саму групу осіб "до" і "після" його прийому.

Парний *t*-тест застосовується для пов'язаних вибірок обсягу n.

Формулювання гіпотез.

а) Двостороння гіпотеза.

Перевіряємо, чи є якийсь ефект, не уточнюючи заздалегідь, у який бік.

Нульова гіпотеза H_0: $\overline{d} = 0$ ("Середня різниця дорівнює нулю").

Альтернативна гіпотеза H_1: $\overline{d} \neq 0$ ("Середня різниця не дорівнює нулю").

б) Правостороння гіпотеза (зростання показника).

Використовується, коли успіх це зростання показника.

Нульова гіпотеза H_0: $\overline{d} = 0$ або $\overline{d} \le 0$.

Альтернативна гіпотеза H_1: $\overline{d} > 0$ ("Середнє "після" менше середнього "до" ").

в) Лівостороння гіпотеза (зниження показника).

Використовується, якщо очікуємо, що значення зменшилися (або друга змінна більша за першу).

Нульова гіпотеза H_0: $\overline{d} = 0$ або $\overline{d} \ge 0$.

Альтернативна гіпотеза H_1: $\overline{d} < 0$ ("Середнє "після" більше середнього "до" ").

Вибір сторони гіпотези впливає на те, де буде знаходитись критична область (область відхилення гіпотези H_0):

Нехай маємо об'єкт дослідження для якого відомі n показників x_i $\left(i = \overline{1,n}\right)$ "до початку" дослідження і n показників y_i $\left(i = \overline{1,n}\right)$ "після закінчення". У даному випадку важлива різниця d_i для кожного показника, яку обчислюємо за формулою:

$$d_i = x_i - y_i . \tag{6.27}$$

Знаходимо середню різницю

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i . \tag{6.28}$$

Знаходимо виправлене стандартне відхилення s_d різниць

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2} . \tag{6.29}$$

Розраховуємо $t_{\text{спост}}$ для парного t-тесту:

$$t_{\text{спост}} = \frac{\overline{d}}{s_d / \sqrt{n}} . \tag{6.30}$$

За таблицею А.4 (див. Додаток А) для степенів свободи k=n−1 та заданим рівнем значущості α знаходимо критичне значення $t_{\text{кр}}$.

При порівнюванні $t_{спост}$ з $t_{кр}$ важливо враховувати знак, оскільки різниця може бути відємною. Правило порівняння виглядає так:

Для двосторонньої гіпотези: порівнюємо абсолютне значення (модуль) $|t_{спост}|$ з $t_{кр}$: якщо $|t_{спост}|>t_{кр}$, то відхиляємо гіпотезу H_0 ("Результат значимий").

Для лівосторонньої гіпотези (зниження показника): якщо знаходили різницю як d="Після"−"До", то очікуємо, що $t_{спост}$ буде від'ємним. Якщо $t_{спост}\leq -t_{кр}$, то відхиляємо гіпотезу H_0.

Для правосторонньої гіпотези (зростання показника): якщо $t_{спост}\geq t_{кр}$, то відхиляємо гіпотезу H_0.

2) *t-тест Стьюдента*

Коли застосовується *t*-критерій Стьюдента?

а) Невідома дисперсія: у 99% реальних завдань невідома точна дисперсія σ^2 всієї генеральної сукупності, а маємо лише її оцінку s^2 із вибірки.

б) Малий обсяг вибірки: Якщо обсяг вибірки n малий (зазвичай $n<30$), то *t*-розподіл враховує ризик появи екстремальних значень через брак даних.

в) Консервативність: Навіть на великих вибірках *t*-тест дає майже той самий результат, що і *Z*-тест, але він біль строгий. Тому у сучасній науці *t*-тест вважається стандартом за умовчанням.

Алгоритм застосування *t*-тесту Стьюдента аналогічний застосуванню *Z*-тесту.

Розраховуємо *t*-статистику Стьюдента для незалежних вибірок із близькими дисперсіями за формулою:

$$t_{спост}=\frac{\bar{x}_{в1}-\bar{x}_{в2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}, \qquad (6.31)$$

де n_1, n_2 – обсяги вибірок; $\bar{x}_{в1}$, $\bar{x}_{в2}$ – вибіркові середні двох вибірок (задані за умовою або знаходяться за формулою (6.20) для кожної вибірки); s_1^2, s_2^2 – вибіркові дисперсії двох вибірок (задані за умовою або знаходяться за формулою $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_в)^2$ для кожної

вибірки).

Число степенів свободи знаходимо за формулою:

$$k = n_1 + n_2 - 2. \tag{6.32}$$

За таблицею А.4 (див. Додаток А) для степенів свободи k та заданого рівня значущості α знаходимо критичне значення $t_{кр}$.

Для двосторонньої гіпотези порівнюємо абсолютне значення $t_{спост}$ з $t_{кр}$. Якщо $|t_{спост}|<t_{кр}$, то гіпотеза H_0 не відхиляється. Для односторонніх гіпотез приймаємо рішення аналогічно розглянутому вище.

Зауваження 6.12. Якщо обсяги вибірок різні $n_1 \neq n_2$, вибіркові дисперсії двох вибірок s_1^2, s_2^2 близькі за значенням, застосовують об'єднану дисперсію s_p^2, яку знаходимо за формулою:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{6.33}$$

а t-статистику Стьюдента розраховуємо за формулою:

$$t_{спост} = \frac{\bar{x}_{в1} - \bar{x}_{в2}}{\sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{6.34}$$

Число степенів свободи знаходимо за формулою (6.32).

3) *t-тест Уелча*

t-тест Стьюдента (розглянутий вище) вимагає дотримання умови рівності дисперсій в обох групах. Але може бути, що одна група є дуже однорідною, а друга має величезний розкид, тобто дисперсії різні. t-тест Уелча дозволяє коригування цієї ситуації.

Розраховуємо t-статистику Уелча за формулою (6.31).

У t-тесті Уелча число степенів свободи розраховується за формулою Саттертуейта. Воно майже завжди виходить дробовим:

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}. \qquad (6.35)$$

Прийнято знайдене значення k округляти в меншу сторону до цілого значення.

Інші дії аналогічні розглянутим для t-тесту Стьюдента.

Приклад 55. Проводять тестування нового препарату для зниження артеріального тиску. Для групи з n=6 пацієнтів отримали дані (систолічний тиск, мм рт. ст.) "до" прийому препарату x_i: 150, 165, 140, 180, 155, 160 та "після" закінчення курсу прийому y_i: 140, 158, 135, 170, 148, 151. Необхідно з'ясувати ефективність препарату при рівні значущості α=0,05.

Розв'язання. Досліджують одну й ту саму групу осіб (n=6) на вплив нового препарату для зниження артеріального тиску "до" і "після" його прийому, тому застосовуємо парний t-тест.

Сформулюємо гіпотези. Маємо двосторонню гіпотезу, оскільки перевіряємо, чи є якийсь ефект від прийому препарату, не уточнюючи заздалегідь, у який бік. Нульова гіпотеза H_0: $\overline{d}$ =0 ("Препарат не впливає на артеріальний тиск"). Альтернативна гіпотеза H_1: $\overline{d} \neq 0$ ("Препарат має значний вплив на тиск (у будь-який бік)").

Знайдемо різницю d_i за формулою (6.27) та середню різницю $\overline{d}$ за формулою (6.28). Для зручності складемо таблицю 6.20:

Таблиця 6.20

i	x_i	y_i	d_i	$(d_i - \overline{d})^2$
1	150	140	10	4
2	165	158	7	1
3	140	135	5	9
4	180	170	10	4
5	155	148	7	1
6	160	151	9	1
n=6			$\overline{d}$ =8	Σ=20

Знаходимо виправлене стандартне відхилення s_d різниць за формулою (6.29):

$$s_d = \sqrt{20/5} = 2.$$

Розраховуємо $t_{\text{спост}}$ для парного t-тесту за формулою (6.30):

$$t_{\text{спост}} = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{8}{2/\sqrt{6}} \approx \frac{8}{2/2{,}4495} \approx 9{,}798.$$

За таблицею А.4 (див. Додаток А) для степенів свободи $k=n-1=6-1=5$ та заданим рівнем значущості $\alpha=0{,}05$ знаходимо критичне значення $t_{\text{кр}}=2{,}571$.

Порівнюємо $t_{\text{спост}}=9{,}798$ з $t_{\text{кр}}=2{,}571$. Оскільки $t_{\text{спост}}>t_{\text{кр}}$, то відхиляємо нульову гіпотезу H_0. Препарат працює. Ефект зниження тиску на 8 одиниць за такого низького розкиду даних ($s_d=2$) може бути випадковим.

Відповідь: Препарат ефективний.

Приклад 56. Проводять тестування на міцність двох видів бетону тип А і тип В. Оскільки немає можливості зробити сотні зразків, тож вибірки маленькі. Для типу А отримали наступні дані: $n_1=5$ зразків, середня міцність $\overline{x}_{\text{в}1}=42$ МПа, виправлена дисперсія $s_1^2=10$, а для типу В: $n_2=7$ зразків, середня міцність $\overline{x}_{\text{в}2}=46$ МПа, виправлена дисперсія $s_2^2=12$. Необхідно з'ясувати ідентичність по міцності цих типів бетону при рівні значущості $\alpha=0{,}05$.

Розв'язання. Сформулюємо гіпотези. Маємо двосторонню гіпотезу.

Нульова гіпотеза H_0: $a_1=a_2$ ("Міцність однакова"). Альтернативна гіпотеза H_1: $a_1\neq a_2$ ("Міцність різна" − двостороння гіпотеза).

Застосовуємо t-тест Стьюдента для відповіді про ідентичність по міцності досліджуваних типів бетону.

Розраховуємо t-статистику Стьюдента за формулою (6.31):

$$t_{\text{спост}} = \frac{42-46}{\sqrt{\frac{10}{5}+\frac{12}{7}}} \approx \frac{-4}{\sqrt{2+2{,}7143}} \approx -2{,}0755.$$

Число степенів свободи знаходимо за формулою (6.32):

$$k=n_1+n_2-2=5+7-2=10.$$

За таблицею А.4 (див. Додаток А) для степенів свободи k=10 та заданого рівня значущості α=0,05 знаходимо критичне значення $t_{кр}$=2,228.

Для двосторонньої гіпотези порівнюємо абсолютне значення $|t_{спост}|=|-2{,}0755|=2{,}0755$ з $t_{кр}$=2,228. Оскільки $|t_{спост}|<t_{кр}$, то нульова гіпотеза H_0 не відхиляється.

Відповідь: Нульова гіпотеза H_0 не відхиляється. Різниця в 4 МПа при таких малих вибірках та розкиданні даних може бути випадковою.

Приклад 57. Порівнюють дві методики навчання швидкості читання двох груп дітей. Отримали результати для дітей (n_1=5) групи А (методика №1): 40,42,38,41,39 (слів/хв), для дітей (n_2=3) групи Б (методика №2): 35,33,34 (слів/хв). Перевірити при рівні значущості α=0,05 чи впливає вибір методики на результат.

Розв'язання. Сформулюємо гіпотези. Нульова гіпотеза H_0: $a_1=a_2$ ("Середня швидкість читання в обох групах однакова. Вибір методики не впливає результат"). Альтернативна гіпотеза H_1: $a_1\neq a_2$ ("Середня швидкість читання у групах значно відрізняється" − двостороння гіпотеза).

Застосовуємо t-тест Стьюдента для вирішення питання про вплив методик на швидкість читання дітей.

За даними вибірками знайдемо для кожної групи дітей середні за формулою (4.6) та вибіркові дисперсії за формулою (4.14).

Група А: середнє $\bar{x}_{в1}$=(40+42+38+41+39)/5=200/5=40, вибіркова дисперсія $s_1^2=(0+2^2+(-2)^2+1+(-1)^2)/4=2{,}5$.

Група Б: середнє $\bar{x}_{в2}$=(35+33+34)/3=102/3=34, вибіркова дисперсія $s_2^2=(1+(-1)^2+0)/2=1{,}0$.

Дисперсії s_1^2=2,5 та s_2^2=1,0 досить близькі (розрізняються менш ніж у 3 рази), тому використовуємо t-тест Стьюдента з об'єднаною дисперсією s_p^2, яку знаходимо за формулою (6.33).

$$s_p^2=\frac{(5-1)\cdot 2{,}5+(3-1)\cdot 1}{5+3-2}=\frac{12}{6}=2.$$

t-статистику Стьюдента розраховуємо за формулою (6.34). Тоді

$$t_{\text{спост}} = \frac{40-34}{\sqrt{2\cdot\left(\frac{1}{5}+\frac{1}{3}\right)}} \approx \frac{6}{\sqrt{2\cdot 0{,}533}} = \frac{6}{\sqrt{1{,}066}} \approx \frac{6}{1{,}032} \approx 5{,}814.$$

Знайдемо число степенів свободи $k=n_1+n_2-2=5+3-2=6$.

За таблицею А.4 (див. Додаток А) для степенів свободи $k=6$ та заданого рівня значущості $\alpha=0{,}05$ знаходимо критичне значення $t_{\text{кр}}=2{,}447$.

Оскільки $t_{\text{спост}}>t_{\text{кр}}$ (5,814>2,447), то відхиляємо нульову гіпотезу H_0.

Відповідь: Нульова гіпотеза H_0 відхиляється. Різниця між методиками статистично значуща. Діти з групи А читають достовірно швидше, і це не пояснюється простою випадковістю.

Приклад 58. Порівнюють час роботи батарейок двох виробників. У першого якість стабільна, у другого – «як пощастить» (високий розкид). Для першого виробника отримали дані: обсяг вибірки $n_1=10$, середнє $\bar{x}_{\text{в}1}=100$ год., дисперсія $s_1^2=5$. Для другого виробника: обсяг вибірки $n_2=8$, середнє $\bar{x}_{\text{в}2}=90$ год., дисперсія $s_2^2=40$. Необхідно з'ясувати ідентичність середнього часу роботи батарейок двох виробників при рівні значущості $\alpha=0{,}05$.

Розв'язання. Різниця середніх $\bar{x}_{\text{в}1}-\bar{x}_{\text{в}2}=100-90=10$, а стандартні відхилення суттєво різняться: $s_1\approx 2{,}236$ і $s_2\approx 6{,}325$ (майже в три рази). У цьому випадку краще застосовувати t-тест Уелча. Сформулюємо гіпотези.

Нульова гіпотеза H_0: $a_1=a_2$ ("Середній час роботи батарейок двох виробників однаковий. Спостережувана різниця в 10 годин є випадковою.").

Альтернативна гіпотеза H_1: $a_1\neq a_2$ ("Середній час роботи батарейок першого виробника статистично значно перевищує середній час роботи батарейок другого виробника. Різниця обумовлена якістю виробництва, а не випадковістю вибірки." − двостороння гіпотеза).

Розраховуємо t-статистику Стьюдента за формулою (6.31):

$$t_{\text{спост}} = \frac{100-90}{\sqrt{\frac{5}{10}+\frac{40}{8}}} \approx \frac{10}{\sqrt{0{,}5+5}} \approx 4{,}264.$$

Число степенів свободи розраховується за формулою Саттертуейта (6.35):

$$k = \frac{(5/10 + 40/8)^2}{\frac{(5/10)^2}{9} + \frac{(40/8)^2}{7}} \approx \frac{5{,}5^2}{0{,}0278 + 3{,}5714} \approx \frac{30{,}25}{3{,}5992} \approx 8{,}405\,.$$

Знайдене значення k округлюємо в меншу сторону до цілого значення: k=8.

За таблицею А.4 (див. Додаток А) для заданого рівня значущості α=0,05 та степенів свободи k=8 знаходимо критичне значення $t_{кр}$=2,306.

Оскільки $t_{спост} > t_{кр}$ (4,264>2,306), то відхиляємо нульову гіпотезу H_0. Перший виробник демонструє не тільки стабільнішу, а й вищу середню якість роботи.

Відповідь: Нульова гіпотеза H_0 відхиляється. У першого виробника якість продукції стабільна.

6.3.3 Перевірка правильності нульової гіпотези H_0 про значення дисперсії

Для перевірки гіпотези про дисперсію однієї вибірки (при порівнянні її з якимось еталоном чи нормативом) χ^2-критерій є основним і загальноприйнятим. Проте важливо пам'ятати: цей тест дуже чутливий до нормальності розподілу. Необхідно мати вибірку обсягу n, дані якої мають бути розподілені нормально, та значення дисперсії σ_0^2 (з техзавдання або минулого досвіду), з яким порівнюється отриманий результат при перевірки гіпотези.

Загальний алгоритм перевірки гіпотези про значення дисперсії:

1. Формулювання гіпотез. Вибір тесту в залежності від питання, що перевіряємо.

а) Двосторонній тест.

Нульова гіпотеза H_0: $\sigma^2=\sigma_0^2$ ("Дисперсія дорівнює заданому значенню").

Альтернативна гіпотеза H_1: $\sigma^2\neq\sigma_0^2$ ("Дисперсія змінилася у будь-яку сторону" − двостороння гіпотеза).

б) Правосторонній тест.

Нульова гіпотеза H_0: $\sigma^2=\sigma_0^2$.

Альтернативна гіпотеза H_1: $\sigma^2>\sigma_0^2$ ("Розкид став більшим, стабільність знизилась" − правостороння гіпотеза).

в) Лівосторонній тест.

Нульова гіпотеза H_0: $\sigma^2=\sigma_0^2$.

Альтернативна гіпотеза H_1: $\sigma^2<\sigma_0^2$ ("Розкид зменшився, процес став точнішим" − лівостороння гіпотеза).

2. Вибір рівня значущості α. Найчастіше обирають α=0,05 (5%), α=0,01 (1%).

3. Розрахунок критерію, що спостерігається ($\chi^2_{\text{спост}}$) за формулою:

$$\chi^2_{\text{спост}} = \frac{(n-1)\cdot s^2}{\sigma_0^2}, \tag{6.36}$$

де n – обсяг вибірки, s^2– виправлена вибіркова дисперсія, σ_0^2 – задане значення дисперсії (з техзавдання або минулого досвіду), з яким порівнюється отриманий результат.

4. Визначення критичної області.

За таблицею А.5 – Критичні точки розподілу Пірсона $\chi^2_{\alpha;k}$ (див. Додаток А) для степенів свободи k= n–1 та заданого рівня значущості α знаходимо критичне значення $\chi^2_{\text{кр}} = \chi^2_{\alpha;k}$.

Якщо $\chi^2_{\text{спост}} > \chi^2_{\text{кр}}$, то відхиляємо нульову гіпотезу H_0; якщо $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$, то нульову гіпотезу H_0 приймаємо.

Приклад 59. Отримана вибірка n=10 замовлень. Час відхилення від графіка доставки становить: 2, 5, 4, 8, 6, 3, 7, 5, 4, 6 (хв.). Сервіс доставки стверджує, що розкид (дисперсія) часу доставки в спальний район не перевищує 25 (стандартне відхилення σ=5 хвилин). Необхідно перевірити при рівні значущості α=0,05, чи не став розкид більше.

Розв'язання. Передбачувана дисперсія σ_0^2=25. Сформулюємо гіпотези.

Нульова гіпотеза H_0: $\sigma^2=\sigma_0^2$ ("Дисперсія дорівнює заявленій.").

Альтернативна гіпотеза H_1: $\sigma^2>\sigma_0^2$ ("Дисперсія більше заявленої." − правостороння гіпотеза).

Знайдемо вибіркові характеристики: середнє $\bar{x}_{\text{в}}$

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\left(2+5+4+8+6+3+7+5+4+6\right) = \frac{50}{10} = 5$$

і вибіркову дисперсію s^2

$$\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2 = (2-5)^2 + (5-5)^2 + (4-5)^2 + (8-5)^2 + (6-5)^2 + (3-5)^2 +$$

$$+ (7-5)^2 + (5-5)^2 + (4-5)^2 + (6-5)^2 = 30.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_{\text{в}})^2 = \frac{1}{9}\cdot 30 \approx 3{,}333.$$

Розраховуємо $\chi^2_{\text{спост}}$ за формулою (6.36):

$$\chi^2_{\text{спост}} = \frac{9\cdot 3{,}333}{25} = \frac{30}{5} = 6.$$

За таблицею А.5 (див. Додаток А) для степенів свободи k=10–1=9 та заданого рівня значущості α=0,05 знаходимо критичне значення $\chi^2_{\text{кр}}$ =16,919.

Оскільки $\chi^2_{\text{спост}} < \chi^2_{\text{кр}}$ (6<16,919), то нульову гіпотезу H_0 приймаємо. Дані не дають підстав стверджувати, що розкид часу доставки перевищує норму.

Відповідь: Нульова гіпотеза H_0 приймається. Дані не дають підстав стверджувати, що розкид часу доставки перевищує норму.

6.3.4 Перевірка правильності нульової гіпотези H_0 про рівність двох дисперсій

Для визначення статистичної значущості відмінностей між дисперсіями двох незалежних вибірок застосовується *F*-тест (критерій) Фішера. При цьому для коректності статистичного висновку потрібне виконання наступних вимог:

Основні умови (*припущення*): Дані у кожній із порівнюваних вибірок повинні підпорядковуватись закону нормального розподілу.

Незалежність вибірок: Об'єкти в групах мають бути обрані випадково та незалежно один від одного. Спостереження в першій вибірці не повинні бути пов'язані зі спостереженнями в другій

(наприклад, це не можуть бути результати тих самих випробувань до і після експерименту).

Кількісний тип даних: Дані мають бути подані в кількісній шкалі (інтервальній або шкалі відносин).

Загальний алгоритм перевірки гіпотези про рівність двох дисперсій:

1. Формулювання гіпотез.

Нульова гіпотеза H_0: $\sigma_1^2=\sigma_2^2$ ("Дисперсії рівні").

Альтернативна гіпотеза H_1: $\sigma_1^2\neq\sigma_2^2$ ("Дисперсії значно відрізняються").

2. Розрахунок вибіркових дисперсій s_1^2 і s_2^2двох вибірок обсягів n_1 і n_2 відповідно, якщо вони не задані за умовою.

3. Обчислюється F-статистика як відношення двох дисперсій:

$$F_{\text{спост}} = \frac{s_1^2}{s_2^2}\,. \tag{6.37}$$

Зауваження 6.13. При розрахунку $F_{\text{спост}}$ у чисельнику завжди записують більшу дисперсію ($s_1^2 > s_2^2$), щоб значення $F_{\text{спост}}$ було більше або дорівнювало 1.

4. Визначаємо степіні свободи: $k_1=n_1-1$ і $k_2=n_2-1$.

5 За таблицею А.11 (див. Додаток А) для степенів свободи k_1 і k_2 та заданим рівнем значущості α знаходимо критичне значення $F_{\text{кр}}$.

Зауваження 6.14. При користуванні таблицею враховуємо, що по горизонталі (верхній рядок) – це k_1 (степінь свободи для вибірки з дисперсією s_1^2), а по вертикалі (перший стовпець) – це k_2 (степінь свободи для вибірки з дисперсією s_2^2). Якщо реальне число степенів свободи відсутнє у таблиці, зазвичай використовують один із двох шляхів: а) беруть найближче менше значення з таблиці до заданого, оскільки це дає трохи суворіший критичний поріг (запас надійності); б) використовують лінійну інтерполяцію між двома значеннями між якими знаходиться реальне число степенів свободи.

Якщо $F_{\text{спост}}\leq F_{\text{кр}}$, то нульова гіпотеза H_0 приймається (відмінності статистично не значущі).

Якщо $F_{спост}>F_{кр}$, то нульову гіпотезу H_0 відхиляємо (дисперсії різняться значно).

Приклад 60. Інвестор збирається купувати акції двох компаній. Порівнює ризик їх придбання. Дисперсія прибутковості – це міра ризику. Для першої компанії отримали наступні дані: n_1=21 день спостережень, дисперсія s_1^2=12, а для другої компанії: n_2=16 день спостережень, дисперсія s_2^2=4. Визначити при рівні значущості α=0,05, наскільки ризики придбання акцій компаній різні.

Розв'язання. Сформулюємо гіпотези.

Нульова гіпотеза H_0: $\sigma_1^2=\sigma_2^2$ ("Ризики однакові").

Альтернативна гіпотеза H_1: $\sigma_1^2\neq\sigma_2^2$ ("Ризики значно відрізняються").

Знайдемо $F_{спост}$ за формулою (6.37) (Більшу дисперсію ділимо на меншу):

$$F_{спост}=\frac{s_1^2}{s_2^2}=\frac{12}{4}=3.$$

Визначимо степіні свободи: $k_1=n_1-1=21-1=20$ і $k_2=n_2-1=16-1=15$.

За таблицею А.11 (див. Додаток А) для степенів свободи k_1=20 і k_2=15 та заданим рівнем значущості α=0,05 знаходимо критичне значення $F_{кр}\approx$2,33.

Оскільки $F_{спост}>F_{кр}$ (3>2,33), то нульову гіпотезу H_0 відхиляємо.

Відповідь: Придбання акцій першої компанії статистично значно більш ризиковано, ніж акцій другої компанії.

Приклад 61. Проводять експеримент, щоби порівняти точність двох приладів. Для першого приладу провели n_1=11 вимірів та визначили дисперсію s_1^2=4,5. Для другого приладу провели n_2=25 вимірювань та визначили дисперсію s_2^2=2,1. Встановити чи відрізняються ці дисперсії при рівні значущості α=0,05.

Розв'язання. Сформулюємо гіпотези.

Нульова гіпотеза H_0: $\sigma_1^2=\sigma_2^2$ ("Дисперсії рівні").

Альтернативна гіпотеза H_1: $\sigma_1^2\neq\sigma_2^2$ ("Дисперсії значно відрізняються").

Знайдемо $F_{спост}$ за формулою (6.37) (Більшу дисперсію ділимо на меншу):

$$F_{\text{спост}} = \frac{s_1^2}{s_2^2} = \frac{4{,}5}{2{,}1} \approx 2{,}14.$$

Визначимо степіні свободи: $k_1=n_1-1=11-1=10$ і $k_2=n_2-1=25-1=24$.

За таблицею А.11 (див. Додаток А) для степенів свободи $k_1=10$ і $k_2=24$ та заданим рівнем значущості $\alpha=0{,}05$ знаходимо $F_{\text{кр}}\approx 2{,}25$.

Оскільки $F_{\text{спост}}<F_{\text{кр}}$ (2,14<2,25), то нульову гіпотезу H_0 приймаємо. Відмінність між дисперсіями статистично незначна (вона обумовлена випадковими факторами).

Відповідь: Відмінність між дисперсіями статистично незначна (вона обумовлена випадковими факторами).

6.4 Завдання для самостійної роботи

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>1. Для інтервального статистичного ряду перевірити гіпотезу про нормальний розподіл за допомогою критерію Пірсона при рівні значущості $\alpha=0{,}05$:
<table>
<tr><td>Межі інтервалів</td><td>5−7</td><td>7−9</td><td>9−11</td><td>11−13</td><td>13−15</td><td>15−17</td></tr>
<tr><td>n_i</td><td>8</td><td>14</td><td>40</td><td>26</td><td>6</td><td>4</td></tr>
</table></td><td>$\chi^2_{\text{спост}} = 3{,}9955$, $\chi^2_{\text{кр}}(0{,}05;\,2) = 5{,}991$ Гіпотезу H_0 про нормальний розподіл приймаємо.</td></tr>
<tr><td>2. Для заданого варіаційного ряду
<table>
<tr><td>x_i</td><td>2</td><td>5</td><td>8</td><td>11</td><td>14</td><td>17</td><td>20</td></tr>
<tr><td>n_i</td><td>14</td><td>25</td><td>28</td><td>30</td><td>26</td><td>22</td><td>23</td></tr>
</table>
використовуючи критерій Пірсона, перевірити гіпотезу про відповідність статистичного розподілу нормальному закону при рівні значущості $\alpha=0{,}01$.</td><td>$\chi^2_{\text{спост}} = 7{,}0{,}17$, $\chi^2_{\text{кр}}(0{,}01;\,3) = 13{,}3$ Гіпотезу H_0 про нормальний розподіл приймаємо.</td></tr>
<tr><td>3. Перевіряють екологічність очисних споруд. За нормою середній вміст шкідливої речовини не повинен перевищувати 50 мг/л. Дисперсія відома з паспорта устаткування: $\sigma^2=16$. Зробили 16 вимірів і отримали середнє значення $\bar{x}_{\text{в}} = 52$ мг/л. Потрібно зрозуміти: чи це випадкове коливання, чи ми реально порушуємо норми? Побудувати верхню межу довірчого інтервалу для рівня надійності 99%.</td><td>$(-\infty;54{,}33)$ Справжній середній вміст домішок не перевищує 54,33 мг/л.</td></tr>
</table>

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>4. Виміряли вхідний опір 130 електронних ламп в Ом. Дані згрупували і отримали:
<table>
<tr><td>x_i</td><td>3-4,2</td><td>4,2-4,8</td><td>4,8-5,4</td><td>5,4-6,0</td><td>6,0-6,6</td><td>6,6-7,2</td></tr>
<tr><td>n_i</td><td>11</td><td>11</td><td>20</td><td>27</td><td>36</td><td>25</td></tr>
</table>
У першому інтервалі об’єднали два, оскільки n_1=3, n_2=8. Прийнявши рівень значущості 10%, перевірити гіпотезу про те, що дані отримані з нормально розподіленої генеральної сукупності.</td><td>$\chi^2_{\text{спост}} \approx 8{,}39$;
$\chi^2_{\text{кр}}(0{,}1;3) = 6{,}25$.
Гіпотеза про нормальний розподіл вхідного опору ламп відкидається.</td></tr>
<tr><td>5. Дві автоматичні лінії фасують чай. З технічних паспортів відомо, що стандартне відхилення маси пачки чаю на першій лінії становить σ_1=0,8 г, на другій – σ_2=1,2 г. Для перевірки роботи ліній були відібрані проби: на першій лінії n_1=40 пачок, середня вага $\bar{x}_{в1}$=102г; на другій лінії n_2=50 пачок, середня вага $\bar{x}_{в2}$=101г. Перевірити на рівні значущості α=0,05, чи суттєво відрізняється середня вага пачок на цих лініях.</td><td>$Z_{\text{спост}}$=4,72,
$Z_{\text{кр}}$=1,96.
Різниця в середній вазі пачок на двох лініях є статистично значущою (перша лінія в середньому пакує більше чаю ніж друга)</td></tr>
<tr><td>6. За 6 робочих днів попит на деякий товар становив: 104, 80, 96, 120, 113, 82. На рівні значущості α=0,1 за допомогою критерію Колмогорова перевірити гіпотезу про те, що попит рівномірно розподілено на відрізку [75;125].</td><td>D=0,253;
$D_{\text{кр}}$=0,473.
Гіпотеза про рівномірний розподіл приймається</td></tr>
<tr><td>7. Виміряли час очікування відповіді оператора у 5 клієнтів: 10, 12, 11, 9, 13 хвилин. Норматив – 10 хвилин. Перевірити, чи відрізняється від нормативу в 10 хвилин за рівня значущості 5% середній час очікування. Побудувати довірчій інтервал для нормативу з надійністю 95%.</td><td>$t_{\text{спост}}$=1,41;
$t_{\text{кр}}$=2,776;
(9,04;12,96).</td></tr>
<tr><td>8. Компанія займається доставкою готової їжі містом. Порівняли швидкість доставки їжі в центрі та в спальному районі. Отримали дані: центр n_1=10, середній час $\bar{x}_{в1}$=100хв., s_1^2=9, а для спального району: n_2=12, середній час $\bar{x}_{в2}$=35хв., дисперсія s_2^2=64. З’ясувати ідентичність середнього часу доставки готової їжі в центр та в спальний район при рівні значущості α=0,05.</td><td>$t_{\text{спост}}$=–2,01;
$t_{\text{кр}}$=2,12.
Статистично значимих відмінностей у швидкості доставки не виявлено.</td></tr>
</table>

Завдання	Відповідь
9. Завод випускає деталі. Відомо, що стандартне відхилення ваги деталі становить σ=2 г. Впровадили нову технологію і треба перевірити, чи стала середня вага деталі меншою, ніж стандартні 100 г. Для цього вибрали 9 деталей та отримали значення $\bar{x}_в$ =98,5. Прийняти рівень значущості α=0,05.	$Z_{спост}$=−2,25; $Z_{кр}$=−1,645. Достатньо підстав стверджувати, що середня вага деталі зменшилась.
10. Необхідно порівняти стабільність доставки замовлення до центру та спального району міста при рівні значущості α=0,05. Отримано дані доставки замовлення для центру: n_1=6 замовлень, виправлена дисперсія s_1^2=100; для району: n_2=11 замовлень, виправлена дисперсія s_2^2=25.	$F_{спост}$=4; $F_{кр}$≈3,33. Розкид часу доставки у центрі статистично значно вище, ніж у спальному районі.
11. За двома незалежними вибірками обсягами n_1=9 і n_2=16, отриманими з нормально розподілених генеральних сукупностей X і Y, розраховано виправлені вибіркові дисперсії s_1^2=34,02 та s_2^2=12,5. Необхідно при рівні значущості α=0,01 перевірити нульову гіпотезу H_0: про рівність генеральних дисперсій.	$F_{спост}$=12,8; $F_{кр}$≈4. Немає підстав відкинути гіпотезу про рівність генеральних дисперсій.
12. За вибіркою обсягу n=16, отриманою з нормально розподіленої генеральної сукупності, знайдено вибіркове середнє $\bar{x}_в$ =12,4 та виправлене середнє квадратичне відхилення s=1,2. Необхідно при рівні значущості α=0,05 перевірити нульову гіпотезу H_0: a=11,8 про рівність генерального середнього заданому числу.	$t_{спост}$=2; $t_{кр}$=2,13. Немає підстав відкинути нульову гіпотезу.
13. Точність роботи верстата-автомата перевіряється за дисперсією розмірів виробів, яка має не перевищувати σ_0^2 =0,01(мм^2). За вибіркою з n=25 виробів отримано виправлену вибіркову дисперсію s^2=0,02(мм^2). На рівні значущості α=0,05 перевірити, чи верстат забезпечує необхідну точність.	$\chi^2_{спост} = 48$; $\chi^2_{кр}(0{,}05; 24) = 36{,}4$ Верстат не забезпечує необхідної точності.

7 Елементи кореляційно-регресійного аналізу

Існують різні типи залежностей між випадковими величинами X та Y, наприклад, функціональна (жорстко детермінована) та статистична (стохастично детермінована).

Функціональна (детермінована) залежність – це залежність, за якої кожному значенню однієї змінної (незалежної, *X*) суворо відповідає одне (або кілька) певне значення іншої змінної (залежної, *Y*). Такі залежності описуються точними математичними формулами, у яких немає місця випадковості. Наприклад, закон Ома, площа круга та інші.

Статистична (стохастична, кореляційна) залежність – це залежність, за якої зміна однієї змінної (незалежної, *X*) тягне за собою зміну середнього значення інший змінної (залежної, *Y*), але не кожного її окремого значення. Ця залежність не є жорсткою і точною, оскільки на залежну змінну, крім досліджуваного фактору, впливають випадкові (невраховані) фактори. Наприклад, залежність між зростом людини та її вагою, залежність між витратами на рекламу та обсягом продажів, залежність між рівнем освіти та доходом та інші.

Ключова відмінність між ними полягає в тому, що в функціональній залежності, знаючи *X*, можна точно прогнозувати *Y*, а в статистичній залежності, знаючи *X*, можна передбачити середнє значення *Y* або діапазон її значень, але завжди присутній елемент випадковості.

Кореляційно-регресійний аналіз – це потужний інструментарій математичної статистики, який використовується для вивчення взаємозв'язків між змінними.

Він дозволяє кількісно виміряти тісноту, напрямок зв'язку (кореляційний аналіз), а також встановити аналітичне вираження залежності результату від конкретних факторів за сталості решти діючих на результативну ознаку факторних ознак (регресійний аналіз).

Тобто він дозволяє не тільки визначити наявність і силу зв'язку, а й побудувати модель, яка дозволяє прогнозувати значення однієї змінної з урахуванням інших.

7.1 Кореляційний аналіз

Мета кореляційного аналізу – це оцінити наявність, напрям і тісноту (силу) статистичного зв'язку між двома чи більше випадковими величинами.

Зауваження 7.1 Кореляція не означає причинно-наслідкового зв'язку. Дві змінні можуть бути сильно корельовані просто тому, що на

них одночасно впливає третя, прихована змінна, або це суто випадковий збіг.

Кореляційна залежність — це окремий випадок статистичної залежності.

Означення 7.1. Кореляційна залежність — це статистичний взаємозв'язок між двома чи більше випадковими величинами, коли зміна значень однієї чи кількох із цих величин супроводжується систематичною зміною середнього значення іншої чи інших величин.

Ключові моменти:

Статистична природа: Це не жорсткий, а імовірнісний зв'язок.

Зміна середнього: При зміні фактору X закономірно зміщується не кожне індивідуальне значення Y, а середнє значення (умовне математичне сподівання).

Неоднозначність: При одному і тому ж значенні факторної ознаки можуть спостерігатися різні значення результативної ознаки, що пояснюється впливом безлічі інших, не врахованих факторів.

Нерівнозначність регресії: На відміну від регресії, де ми чітко виділяємо залежну (*Y*) та незалежну (*X*) змінні, в кореляції часто обидві змінні можуть розглядатися як взаємопов'язані без строгого причинно-наслідкового зв'язку (хоча іноді причинність очевидна).

Кореляція відповідає на питання: "Наскільки сильно змінні пов'язані?", тоді як регресія відповідає на питання: "Як одна змінна пророкує іншу?".

Теорія кореляції вирішує два основні завдання:

1) *Визначення форми зв'язку* (*вид функції регресії*):

- це завдання полягає у встановленні математичної функції (рівняння регресії), яка найкраще описує взаємозв'язок між змінними;
- це може бути лінійна (може бути описана прямою лінією) та нелінійна (може бути описана квадратичною, показовою або іншою нелінійною функцією);
- візуально форму зв'язку можна попередньо оцінити за діаграмою розсіювання (кореляційному полю), де кожна точка відповідає парі значень (x_i,y_i).

2) *Оцінка тісноти* (*сили*) *зв'язку*:

- це завдання спрямовано на кількісний вимір степеня близькості

між фактичними значеннями залежної змінної та значеннями, передбаченими рівнянням регресії, тобто, наскільки щільно точки на кореляційному полі згруповані навколо лінії регресії.

- для цього використовуються коефіцієнти кореляції. Чим ближче значення коефіцієнта до абсолютної одиниці (+1 або –1), тим тісніше зв'язок. Чим ближче до нуля, тим слабший зв'язок або він відсутній.

- напрямок зв'язку: прямий (додатний) зв'язок (збільшення однієї змінної супроводжується збільшенням іншої (і навпаки), коефіцієнт кореляції буде додатний), наприклад, зріст та вага; зворотний (від'ємний) зв'язок (збільшення однієї змінної супроводжується зменшенням іншої (і навпаки), коефіцієнт кореляції буде від'ємним), наприклад, ціна товару та попит на нього.

Види кореляційного зв'язку:

Класифікація кореляційних зв'язків ґрунтується на кількості досліджуваних змінних та формі взаємозв'язку.

1) За кількістю ознак, що корелюються:

а) *Парна кореляція* (*проста*):

Вивчає взаємозв'язок між двома змінними (одним факторним та одним результативним або двома рівнозначними).

Наприклад, зв'язок між доходом та витратами, між рівнем освіти та зарплатою.

Для вимірювання тісноти парного лінійного зв'язку найчастіше використовується коефіцієнт лінійної кореляції Пірсона (r_{xy}).

б) *Множинна кореляція*:

Вивчає взаємозв'язок між однією залежною змінною (Y) та двома або більше незалежними змінними (X_1, X_2, …, X_k) одночасно.

Дозволяє оцінити сукупний вплив кількох факторів на результативну ознаку.

Наприклад, вплив досвіду роботи, освіти та віку на зарплату; вплив температури, вологості та освітленості на зростання рослин.

Для вимірювання тісноти множинного зв'язку використовується коефіцієнт множинної кореляції (R). Він змінюється від 0 до 1 і показує, як добре кілька незалежних змінних пояснюють варіацію залежної змінної.

в) *Часткова кореляція*:

Вивчає тісноту зв'язку між двома змінними, виключаючи (або "фіксуючи") вплив інших змінних.

Наприклад, зв'язок між ціною та попитом, виключивши вплив реклами та доходу населення.

Використовуються часткові коефіцієнти кореляції.

2) За формою (характером) зв'язку:

а) *Лінійна кореляція*:

Передбачається, що залежність між змінними може бути описана адекватно прямою лінією.

Зміна однієї змінної веде до пропорційної (або майже пропорційної) зміни середнього значення іншої змінної.

Графік розсіювання буде нагадувати хмару точок, витягнуту вздовж прямої лінії.

Наприклад, зв'язок між кількістю споживаної їжі та набраною вагу.

Для вимірювання тісноти нелінійного зв'язку використовують коефіцієнт Пірсона (

б) *Нелінійна* (*криволінійна*) *кореляція*:

Передбачається, що залежність між змінними описується кривою лінією (параболою, гіперболою, експонентою тощо).

Зміна однієї змінної веде до непропорційної зміни середнього значення іншою.

Графік розсіювання буде нагадувати вигнуту хмару точок.

Наприклад, залежність продуктивності праці від стажу роботи (спочатку зростає, потім стабілізується і навіть падає); залежність урожайності від кількості добрив, що вносяться (до певної межі, потім може знижуватися).

Для вимірювання тісноти нелінійного зв'язку використовують кореляційні відносини (η). Коефіцієнт Пірсона для нелінійного зв'язку може бути близьким до нуля, навіть якщо зв'язок сильний (але нелінійний). Кореляційне відношення η не чутливе до форми зв'язку і завжди від'ємне.

Маємо вибірку (x_i,y_i), $i=1,2,\ldots,n$. Для того, щоб результати кореляційного аналізу були достовірними, вибірка повинна відповідати ряду критеріїв:

Репрезентативність: Вибірка має адекватно відображати властивості всієї генеральної сукупності.

Випадковість: Кожна одиниця сукупності повинна мати рівний шанс потрапити у вибірку.

Нормальність розподілу: Класичний коефіцієнт Пірсона вимагає, щоб змінні *X* та *Y* були розподілені за нормальним законом (або близьким до нього).

Обсяг вибірки (*n*): Для лінійної кореляції бажано мати щонайменше 20–30 спостережень. На малих вибірках (наприклад, n=5) висновки мають швидше ілюстративний характер.

7.1.1 Кореляційне поле та кореляційна таблиця

Кореляційне поле та кореляційна таблиця є допоміжними засобами під час аналізу вибіркових даних.

Перш ніж переходити до формул, теоретично завжди згадують графічний метод – побудова поля кореляції.

При нанесенні на координатну площину вибіркових точок (x_i,y_i) одержують *кореляційне поле* або *поле кореляції*.

Візуалізація кореляційного поля (діаграма розсіювання) дає змогу висунути припущення щодо форми зв'язку між досліджуваними показниками *X* та *Y*. За характером аналітичного виразу кореляційні залежності поділяють на лінійні (рис. 7.1) та нелінійні (рис. 7.2).

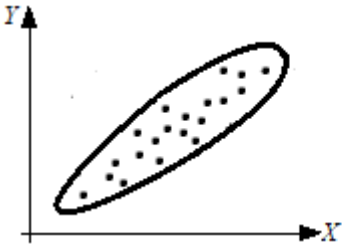


Рисунок 7.1

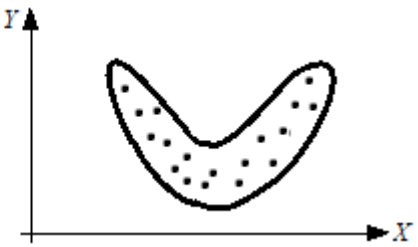


Рисунок 7.2

При лінійній залежності обвідна кореляційного поля має вигляд більш менш правильного еліпса зі згущенням точок у центрі та порівняно рідкісним їх розташування на периферії (кількість пар спостережень велика: n>100). Відхилення осей еліпса від координатних напрямів вказує на наявність кореляції. Витягнутість же еліпса не завжди є її об'єктивним показником, оскільки залежить від прийнятих масштабів по осях координат. По кореляційному полю можна будувати висновки про форму і тісноту зв'язку.

Лінійний взаємозв'язок між двома випадковими величинами

характеризується тим, що при зміні однієї з них інша виявляє тенденцію до зростання (або спадання) згідно з лінійною функцією. Виявлення форми статистичної залежності є необхідним для вибору методу оцінки тісноти (сили) взаємозв'язку.

Напрямленість є додатною, якщо збільшення значення однієї ознаки призводить до збільшення значення іншої (рис. 7.3). Напрямленість є від'ємною, якщо збільшення значення однієї ознаки призводить до зменшення значення іншої (рис. 7.4). Залежність може не мати напрямленості.

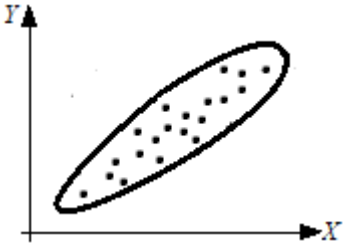


Рисунок 7.3

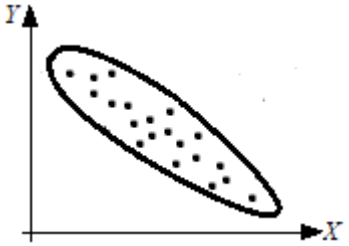


Рисунок 7.4

Для визначення лінійності зв'язку необхідно насамперед побудувати графік цього зв'язку. Для цього на графік у прямокутній системі координат наносять у вигляді точки дані кожної пари значень x і y. Графік крім форми зв'язку дозволяє побачити і тісноту зв'язку.

Для чисельної обробки результати вибірки зазвичай групують і репрезентують у формі *кореляційної таблиці*.

Нехай в результаті незалежних спостережень над величинами X і Y з генеральної сукупності отримали вибірку обсягу $n<30$, причому обсяг вибіркової n сукупності визначається як кількість наявних у вибірці пар (x_i,y_i), $i=1,2,\ldots,n$. Тоді для заданих значень вхідної змінної X та значень вихідної змінної Y складають таблицю 7.1:

Таблиця 7.1

X	x_1	...	x_n
Y	y_1	...	y_n

Тут x_i позначає незалежну змінну (аргумент), а y_i – залежну змінну (функцію), i – будь-який порядковий номер x або y, від 1 до n, n – загальна кількість пар спостережень x і y.

При великій кількості спостережень (наприклад, $n>30$) проводять

систематизацію даних шляхом їх угруповання та побудови *кореляційної таблиці* (табл. 7.2), не забуваючи при цьому, що групування вносить похибки в результати обчислень. Кореляційна таблиця будується за інтервалами значень x і y, обраними самостійно.

Нехай величина X у вибірці приймає значення $x_1,x_2,\ldots,x_k$, де k – кількість значень цієї величини, що різняться між собою, причому в загальному випадку кожне з них у вибірці може повторюватися, а величина Y у вибірці приймає значення $y_1,y_2,\ldots,y_m$, де m – кількість значень цієї величини, що різняться між собою, причому в загальному випадку кожне з них у вибірці також може повторюватися.

Припускаючи, що довжини інтервалів угруповання (по кожній із змінних x та y) рівні між собою, проводиться вибірка даних по цих інтервалах: Δx_i та Δy_j (i=1,2,…,k; j=1,2,…,m). В результаті одержують частоту n_{ij} (i=1,2,…,k; j=1,2,…,m) поєднань значень x та y для певних інтервалів, тобто на перетині кожного вертикального стовпця і горизонтального рядка записують частоту n_{ij}, що показує скільки разів при даному значенні $x\in\Delta x_i$ зустрічалися вказані значення $y\in\Delta y_j$ або навпаки.

Як основу для розрахунків таблицю спрощують, вибравши середини x_i та y_j (i=1,2,…,k; j=1,2,…,m) цих інтервалів та числа n_{ij}. Таблицю доповнюють, вписуючи в передостанній рядок та передостанній стовпець суми частот n_x по стовпцям та рядкам n_y:

$$n_x = \sum_{i=1}^{k} n_{x_i}\ , \quad n_y = \sum_{j=1}^{m} n_{y_j}\ , \tag{7.1}$$

де

$$n_{x_i} = \sum_{j=1}^{m} n_{ij} \text{ – частота ознаки } x_i, \tag{7.2}$$

$$n_{y_j} = \sum_{i=1}^{k} n_{ij} \text{ – частота ознаки } y_j, \tag{7.3}$$

а в останній рядок і останній стовпець вписують зважені за частотами умовні середні ($\overline{y}_{x_i}$ і $\overline{x}_{y_j}$) значень y і x по стовпцям і рядкам:

$$\bar{y}_{x_i} = \frac{1}{n_{x_i}} \sum_{j=1}^{m} y_j n_{ij} \ (i=1,2,\ldots,k), \quad \bar{x}_{y_j} = \frac{1}{n_{y_j}} \sum_{i=1}^{k} x_i n_{ij} \ (j=1,2,\ldots,m). \quad (7.4)$$

Суми величин, що стоять у передостанньому рядку та передостанньому стовпці, повинні дорівнювати загальній кількості спостережень n:

$$n = \sum_{i=1}^{k} n_{x_i} = \sum_{j=1}^{m} n_{y_j} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} \ . \quad (7.5)$$

У передостанні клітини останнього рядка та останнього стовпця вписують підраховані загальні середні виважені значення всіх y і x, тобто $\bar{y}$ та $\bar{x}$. Вони можуть бути обчислені як середні всіх y або x, зважені за частотами n_{ij}, або як середні з $\bar{y}_x$ і $\bar{x}_y$, зважені за частотами n_x або n_y:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} y_j \quad \text{або} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{k} n_{x_i} \cdot \bar{y}_{x_i} \ ; \quad (7.6)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} x_i \quad \text{або} \quad \bar{x} = \frac{1}{n} \sum_{j=1}^{m} n_{y_j} \cdot \bar{x}_{y_j} \ . \quad (7.7)$$

Таблиця 7.2

Інтервал Δx	Середина інтервалу Δy	Інтервал Δx				n_y	$\bar{x}_{y_j}$
		Δx_1	Δx_2	...	Δx_k		
		Середина інтервалу Δx					
		x_1	x_2	...	x_k		
Δy_1	y_1	n_{11}	n_{21}	...	n_{k1}	n_{y_1}	$\bar{x}_{y_1}$
Δy_2	y_2	n_{12}	n_{22}	...	n_{k2}	n_{y_2}	$\bar{x}_{y_2}$
...	...	...	...	...	...	...	
Δy_m	y_m	n_{1m}	n_{2m}	...	n_{km}	n_{y_m}	$\bar{x}_{y_m}$
n_x		n_{x_1}	n_{x_2}	...	n_{x_k}	n	$\bar{x}$
$\bar{y}_{x_i}$		$\bar{y}_{x_1}$	$\bar{y}_{x_2}$		$\bar{y}_{x_k}$	$\bar{y}$	

За розташуванням заповнених клітинок у таблиці можна побачити напрямок зв'язку: якщо значення концентруються вздовж діагоналі з верхнього лівого кута в правий нижній (при збільшенні x збільшується y), то зв'язок прямий; якщо з лівого нижнього до правого верхнього (при збільшенні x зменшується y), то зв'язок зворотний.

7.1.2 Коефіцієнт лінійної кореляції Пірсона та його властивості

Оскільки важлива не сама залежність однієї змінної від іншої, а саме характеристика тісноти зв'язку між ними, то саме вибірковий коефіцієнт лінійної кореляції Пірсоном оцінює тісноту лінійної кореляційної залежності і вказує її напрям (прямий чи зворотний).

1. Випадок несгрупованих даних (табл. 7.1).

1) *Коефіцієнт вибіркової коваріації* (*кореляційний момент*) (K_{XY} або cov(X,Y)) визначається формулою:

$$K_{XY} = \operatorname{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \tag{7.8}$$

де $\bar{x}$, $\bar{y}$ – вибіркові середні випадкових величин X та Y:

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n}x_i, \quad \bar{y} = \frac{1}{n}\cdot\sum_{i=1}^{n}y_i \tag{7.9}$$

або

$$K_{XY} = \overline{xy} - \bar{x}\cdot\bar{y}, \tag{7.10}$$

де

$$\overline{xy} = \frac{1}{n}\cdot\sum_{i=1}^{n}x_i\cdot y_i\ . \tag{7.11}$$

Кореляційний момент відображає характер взаємозв'язку між X та Y. Якщо $K_{XY}>0$, то взаємозв'язок прямий. Якщо $K_{XY}<0$, то взаємозв'язок зворотний. Показує напрямок зв'язку (додатний або від'ємний). Його недолік – залежність від одиниць виміру змінних, що ускладнює порівняння сили зв'язку між різними парами змінних.

2) *Вибірковий коефіцієнт лінійної кореляції Пірсона* (r_{xy} або $r_в$) знаходиться за формулою:

$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{K_{XY}}{\sigma_x \cdot \sigma_y}, \tag{7.12}$$

де σ_x та σ_y – вибіркові середньо квадратичні відхилення випадкових величин X та Y:

$$\sigma_{\text{x}} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\overline{x}\right)^2} = \sqrt{\overline{x^2} - \left(\overline{x}\right)^2}, \tag{7.13}$$

$$\sigma_y = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - \left(\overline{y}\right)^2} = \sqrt{\overline{y^2} - \left(\overline{y}\right)^2}. \tag{7.14}$$

Формулу (7.12) можна записати у вигляді

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{7.15}$$

або

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{n} \Delta x \cdot \Delta y}{\sqrt{\sum_{i=1}^{n} (\Delta x)^2} \cdot \sqrt{\sum_{i=1}^{n} (\Delta y)^2}}, \tag{7.16}$$

де $\Delta x = x_i - \overline{x}$, $\Delta y = y_i - \overline{y}$.

2. Випадок згрупованих даних (табл. 7.2).

1) *Коефіцієнт вибіркової коваріації* (*кореляційний момент*) (K_{XY} або cov(X,Y)) визначається формулою:

$$K_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij} - \overline{x} \cdot \overline{y}, \tag{7.17}$$

де $\overline{x}$, $\overline{y}$ – середні арифметичні:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_{x_i}, \quad \bar{y} = \frac{1}{n} \cdot \sum_{j=1}^{m} y_j \cdot n_{y_j}, \tag{7.18}$$

а x_i, y_j – середини відповідних інтервалів; n_{x_i}, n_{y_j} – частоти потрапляння випадкових величин X та Y (відповідно) у зазначені інтервали.

2) *Вибірковий коефіцієнт лінійної кореляції Пірсона* (r_{xy} або $r_в$) знаходиться за формулою:

$$r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \tag{7.19}$$

де $n = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$, σ_x та σ_y – вибіркові середньо квадратичні відхилення випадкових величин X та Y:

$$\sigma_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 \cdot n_{x_i} - \left(\bar{x}\right)^2}, \quad \sigma_y = \sqrt{\frac{1}{n} \cdot \sum_{j=1}^{m} y_j^2 \cdot n_{y_j} - \left(\bar{y}\right)^2}. \tag{7.20}$$

Застосування умовних варіант

Для спрощення розрахунків часто використовуються умовні варіанти, які визначаються за формулами:

$$u_i = (x_i - C_1)/h_1, \qquad v_j = (y_j - C_2)/h_2, \tag{7.21}$$

де C_1, C_2 – "хибний початок", нулі (зазвичай вибираються); h_1, h_2 – різниці між сусідніми значеннями X та Y.

Для спрощення розрахунків коефіцієнта кореляції при використанні умовних варіантів, значення C (хибний початок) рекомендується вибирати, виходячи з наступних пріоритетів:

1) *Пріоритет максимальної частоти*: Як C вибирається середина того інтервалу, який має найбільшу частоту (n_x або n_y), оскільки при

множенні частоти на умовний варіант (n_i u_i) у цьому рядку/стовпці вийде 0, що виключає найбільші числа з подальших розрахунків.

2) *Пріоритет центрального положення*: Якщо частоти в кількох центральних інтервалах близькі за значенням (наприклад, 29 і 27), перевага віддається геометричній середині варіаційного ряду, оскільки це робить ряд умовних варіантів симетричним (наприклад, −2,−1,0,1,2), що спрощує підсумкове підсумовування та перевірку розрахунків.

Наприклад, 5 інтервалів (непарно), то вибирають 3-й інтервал, якщо його частота одна з найвищих.

Наприклад, інтервалів 6 (парно), то порівнюють частоти 3-го та 4- го інтервалів і вибирають той, де частота більша.

Якщо є пік (аномально висока частота), то вибирають цей інтервал за *C*, навіть якщо він зміщений щодо центру.

Для розрахунку вибіркового коефіцієнта кореляції у цьому випадку використовуються формула:

$$r_{xy} = \frac{\sum n_{uv} uv - n \cdot \bar{u} \cdot \bar{v}}{n \cdot \sigma_u \cdot \sigma_v}, \qquad (7.22)$$

де $\bar{u} = \frac{1}{n} \cdot \sum n_u \cdot u$, $\bar{v} = \frac{1}{n} \cdot \sum n_v \cdot v$ – середні значення умовних варіант; σ_u, σ_v – середні квадратичні відхилення умовних варіант: $\sigma_u^2 = \frac{1}{n} \cdot \sum n_u \cdot u^2 - \left(\bar{u}\right)^2$, $\sigma_v^2 = \frac{1}{n} \cdot \sum n_v \cdot v^2 - \left(\bar{v}\right)^2$

Відповідно, для зворотного переходу застосовуються вирази:

$$x_i = h_1 \cdot u_i + C_1, \qquad y_j = h_2 \cdot v_j + C_2,$$

$$\bar{x} = h_1 \cdot \bar{u} + C_1, \qquad \bar{y} = h_2 \cdot \bar{v} + C_2,$$

$$\sigma_x = h_1 \cdot \sigma_u, \qquad \sigma_y = h_2 \cdot \sigma_v.$$

Основні властивості вибіркового коефіцієнта лінійної кореляції

Якщо коефіцієнт кореляції двох величин r_{xy}=+1, то маємо ідеальний прямий лінійний зв'язок у разі зростаючої залежності.

Якщо коефіцієнт кореляції двох величин r_{xy} =−1, то маємо ідеальний зворотний лінійний зв'язок у разі спадної залежності.

Якщо r_{xy} =0, то лінійний зв'язок відсутній (це не означає відсутність будь-якого зв'язку, він може бути нелінійним).

Якщо дві величини пов'язані лінійною кореляційною залежністю, то абсолютна величина коефіцієнта кореляції задовольняє нерівності 0<|r|<1. При цьому коефіцієнт кореляції додатний при зростаючій кореляційній залежності і від'ємний, якщо кореляційна залежність спадна.

Чим ближче значення r_{xy} до +1 або −1, тим сильніший лінійний зв'язок і тим більше прямолінійна кореляція між величинами X і Y.

Для оцінки сили зв'язку зазвичай використовують шкалу Чеддока (табл. 7.3):

Таблиця 7.3

Сила зв’язку	Характер зв’язку	
	Прямий (+)	Зворотній (–)
Повна	1	–1
Дуже сильна	від 0,9 до 0,99	від –0,9 до –0,99
Сильна	від 0,7 до 0,9	від –0,7 до –0,9
Середня	від 0,5 до 0,7	від –0,5 до –0,7
Слабка	від 0,3 до 0,5	від –0,3 до –0,5
Дуже слабка	від 0,01 до 0,3	від –0,01 до –0,3
Зв’язок відсутній	0	0

Перевірка значущості коефіцієнта кореляції

Для перевірки значущості коефіцієнта кореляції використовують *t*- критерій Стьюдента.

Загальний алгоритм перевірки гіпотези про значущість коефіцієнта кореляції:

1. Формулювання гіпотез.

Нульова гіпотеза H_0: ρ=0 ("Зв'язку в генеральній сукупності немає, результат випадковий").

Альтернативна гіпотеза H_1: $\rho\neq0$ ("Зв'язок статистично значущий").

ρ –позначення справжнього коефіцієнта кореляції в генеральній сукупності, який зазвичай невідомий.

2. Розрахунок вибіркового коефіцієнту кореляції Пірсона (r_{xy}) за формулою (7.12).

3. Обчислюється фактичне значення t-критерію за формулою:

$$t_{\text{факт}} = \frac{\left|r_{xy}\right| \cdot \sqrt{n-2}}{\sqrt{1-\left(r_{xy}\right)^2}}, \tag{7.23}$$

де r_{xy} – розрахований вибірковий коефіцієнт кореляції, n – обсяг вибірки (кількість пар даних), $n-2=k$ – число степенів свободи.

За таблицею А.4 (див. Додаток А) для степенів свободи k та заданим рівнем значущості α (зазвичай 0,05 або 0,01) знаходимо критичне значення $t_{\text{кр}}$.

Якщо $t_{\text{факт}}>t_{\text{кр}}$, то відхиляємо гіпотезу H_0 ("Зв'язок статистично значущий"). Якщо $t_{\text{факт}}\le t_{\text{кр}}$, то приймаємо гіпотезу H_0 ("Зв'язок випадковий, довіряти коефіцієнту r_{xy} не можна").

Зауваження 7.2. Перевірити значущість коефіцієнта кореляції можна так званим "Інженерним методом". Якщо не застосовувати t- критерій Стьюдента, то є спрощена формула для критичної помилки коефіцієнта кореляції (m_r):

$$m_r = \frac{1-\left(r_{xy}\right)^2}{\sqrt{n}}. \tag{7.24}$$

Правило надійності: коефіцієнт кореляції вважається значущим, якщо він перевищує свою помилку щонайменше в 3 рази.

Коефіцієнт детермінації R^2

Для оцінки частки варіації (зміни) результативної ознаки Y, обумовленої впливом фактору X, розраховується коефіцієнт детермінації R^2.

Для випадку лінійної залежності коефіцієнт детермінації дорівнює квадрату коефіцієнта кореляції Пірсона:

$$R^2=(r_{xy})^2. \tag{7.25}$$

Економічний та статистичний зміст: якщо, наприклад, r_{xy}=0,9 (високий зв'язок), то R^2=0,81. Це означає, що 81% змін залежної змінної (Y) обумовлено впливом фактору X, а 19%, що залишилися, припадають на частку інших факторів, не врахованих в моделі (помилка моделі, випадкові коливання).

Коефіцієнт детермінації більш наглядний, оскільки відсотки сприймаються легше, ніж абстрактні коефіцієнти від −1 до +1. Також дозволяє зробити оцінку якості моделі (у регресійному аналізі R^2 є основним критерієм того, наскільки вдало підібрано рівняння регресії: чим ближче R^2 до 1, тим вище прогностична здатність моделі) (табл. 7.4).

Таблиця 7.4

Значення R^2	Якість моделі
>0,8	Дуже хороша
0,5–0,8	Хороша / Прийнятна
<0,5	Низька (модель вимагає доопрацювання)

7.1.3 Коефіцієнт рангової кореляції Спірмена ρ

Коефіцієнт рангової кореляції Спірмена − статистичний критерій, який найчастіше використовується при обробці емпіричних даних. Цей критерій належить до типу непараметричних і не вимагає, щоб дані були розподілені за нормальним законом. Достатньо, якщо показники представлені в порядковій шкалі, тобто враховується лише той факт, що один показник більший або менший, ніж інший.

Коефіцієнт Пірсона шукає строгий лінійний зв'язок (пряму лінію), а коефіцієнт Спірмена набагато гнучкіший. Він оцінює монотонність: чи росте одна змінна при зростанні іншої, неважливо, по прямій або по кривій.

Коли використовують коефіцієнт Спірмена:

- дані не розподілені нормально;
- є очевидні викиди (Спірмен їх майже не боїться).
- дані представлені у вигляді рангів (місця у рейтингу) або якісних оцінок (низький, середній, високий).

Зауваження 7.3. Коефіцієнт рангової кореляції Спірмена застосовують у випадку, коли обсяг вибірки не менший 5, оскільки при

n=5 будь-яка висунута гіпотеза, швидше за все, буде визнана «незначною», навіть якщо зв'язок насправді є.

Для підрахунку рангової кореляції необхідно мати два ряди значень: величин X, яка приймає значення $x_1,x_2,\ldots,x_n$, де n – кількість значень цієї величини, та Y, яка приймає значення $y_1,y_2,\ldots,y_n$, де n – кількість значень цієї величини, які можуть бути проранжовані. Треба перейти від реальних значень до порядкових номерів (рангів).

Для ранжування необхідно значення для ознаки X розташувати від меншого до більшого. Найменше значення отримує ранг1, наступне – 2 і т.д. Те ж саме виконують з ознакою Y.

Краще застосовувати коефіцієнт рангової кореляції Спірмена коли обидва ранжовані ряди (випадкові величини X та Y) мають послідовність значень, що не збігаються. При великій кількості однакових рангів по одній або обох порівнюваних величин необхідно вносити поправку на однакові ранги.

Зауваження 7.4. Якщо в ранжуванні ознак X або Y значення однакові, їм надається середній ранг (наприклад, якщо два числа ділять 3 і 4 місця, обидва отримують ранг 3.5).

При застосуванні коефіцієнта Спірмена розглядаються гіпотези.

Нульова гіпотеза H_0: "Статистично значущий взаємозв'язок між ознаками X та Y відсутній (коефіцієнт кореляції дорівнює 0)".

Альтернативна гіпотеза H_1: "Існує статистично значущий взаємозв'язок між ознаками X і Y (коефіцієнт кореляції значно відрізняється від 0)".

Для підрахунку рангової кореляції необхідно визначити різниці (d) між рангами, тобто для кожної пари треба відняти з рангу X ранг Y:

$$d=R_x-R_y. \qquad (7.26)$$

Перевірочне правило: сума всіх d (без квадрата) завжди повинна дорівнювати 0. Якщо не нуль, то десь помилка в відніманні.

Щоб позбутися мінусів, отримане d треба піднести до квадрату (d^2) та підрахувати суму значень d^2.

За наявності однакових рангів необхідно розрахувати поправки T для кожного ряду (X і Y) за формулою:

$$T = \sum_i \frac{t_i^3 - t_i}{12}, \tag{7.27}$$

де t_i – скільки разів повторюється однакове значення рангу ознаки.

Коефіцієнт рангової кореляції $\rho_{\text{факт}}$ розраховується за формулою:

а) за відсутності однакових рангів

$$\rho_{\text{факт}} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}, \tag{7.28}$$

де n – кількість пар даних, що брали участь у ранжируванні.

б) при наявності однакових рангів

$$\rho_{\text{факт}} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n \cdot \left(n^2 - 1\right)}, \tag{7.29}$$

де T_X, T_Y – це поправка для кожного ряду (X та Y)

Якщо збігів багато (більше 25% вибірки), то використовують розширену формулу:

$$\rho_{\text{факт}} = 1 - \frac{\frac{n^3 - n}{6} - \sum_{i=1}^{n} d_i^2 - T_X - T_Y}{\sqrt{\frac{n^3 - n}{6} - 2 \cdot T_X} \cdot \sqrt{\frac{n^3 - n}{6} - 2 \cdot T_Y}}. \tag{7.30}$$

Якщо $\rho_{\text{факт}} > 0$, то показники зростають разом. Якщо $\rho_{\text{факт}} < 0$, то один зростає, а інший спадає.

За таблицею А.12 (див. Додаток А) для кількості значень n та заданим рівнем значущості α (зазвичай 0,05 або 0,01) знаходимо критичне значення $\rho_{\text{кр}}$.

Якщо $\rho_{факт} \geq \rho_{кр}$, то відхиляємо гіпотезу H_0. Приймається альтернативна гіпотеза H_1 про наявність статистично значущого взаємозв'язку.

Якщо $\rho_{факт} < \rho_{кр}$, то приймаємо гіпотезу H_0. Гіпотеза про наявність взаємозв'язку не підтверджується.

Якщо вибірка велика (n>40), то звичайні таблиці критичних значень Спірмена незручні. У цьому випадку переходять до t-критерію Стьюдента.

Перевіряємо, наскільки розрахований коефіцієнт $\rho_{факт}$ відрізняється від чистого нуля (тобто відсутність зв'язку).

Розраховуємо значення $t_{факт}$ за формулою:

$$t_{факт} = \rho \cdot \frac{\sqrt{n-2}}{\sqrt{1-\rho^2}}, \tag{7.31}$$

де n – кількість значень, ρ – розрахований коефіцієнт $\rho_{факт}$.

За таблицею А.4 (див. Додаток А) для степенів свободи $k = n–2$ та заданим рівнем значущості α (зазвичай 0,05 або 0,01) знаходимо критичне значення $t_{кр}$.

Якщо $|t_{факт}| > t_{кр}$, то зв'язок статистично значущий (це не випадковість). Якщо $|t_{факт}| \leq t_{кр}$, то зв'язок випадковий, навіть якщо коефіцієнт здається великим.

7.1.4 Кореляційне відношення η

Для оцінки тісноти та характеру взаємозв'язку між показниками використовується коефіцієнт кореляції (Пірсона/Спірмена). На додаток до нього розраховується кореляційне відношення (η).

Означення 7.2. *Кореляційне відношення* (η) – це показник, який вимірює тісноту зв'язку між змінними за будь-якої форми залежності: як лінійної, так і нелінійної.

Статистичний аналіз взаємозв'язків проводиться таким чином:

1) Здійснюється перевірка розподілу ознак на нормальність.

- відповідно до закону нормального розподілу, для оцінки лінійного зв'язку застосовується коефіцієнт кореляції Пірсона (r_{xy}).

- у разі відхилення розподілу від нормального або під час роботи з порядковими даними використовується коефіцієнт рангової кореляції Спірмена (ρ).

2) Для додаткової перевірки форми зв'язку (лінійна/криволінійна) та оцінки загальної детермінації (визначити яку частку варіації (мінливості) залежної змінної Y можна пояснити впливом незалежної змінної X) розраховується кореляційне відношення (η). Порівняння коефіцієнта кореляції (r_{xy} або ρ) з кореляційним відношенням (η) дозволяє підтвердити адекватність обраної лінійної моделі або зробити висновок про наявність складної нелінійної залежності.

Розрахунок кореляційного відношення η базується на правилі складання дисперсій. Сенс кореляційного відношення η у тому, щоб порівняти два види розкиду (дисперсії). Кореляційне відношення η обчислюється за формулою:

$$\eta = \sqrt{\frac{\sigma^2_{\text{міжгр}}}{\sigma^2_{\text{загал}}}}\,, \tag{7.32}$$

де $\sigma^2_{\text{міжгр}}$ –міжгрупова дисперсія, $\sigma^2_{\text{загал}}$ – загальна дисперсія. $\bar{x}$

Загальна дисперсія $\sigma^2_{\text{загал}}$ показує загальний розкид всіх значень Y щодо їх загального середнього $\bar{y}$. Розраховуємо її за формулою:

$$\sigma^2_{\text{загал}} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\,, \tag{7.33}$$

де y_i – кожне окреме значення Y, $\bar{y}$ – загальне середнє арифметичне всіх значень Y, n – загальне число значень Y (обсяг вибірки).

Міжгрупова дисперсія $\sigma^2_{\text{міжгр}}$ показує, наскільки сильно різняться між собою середні значення у групах, сформованих фактором X. Саме ця частина дисперсії вважається «поясненою» фактором X. Для її знаходження необхідно розбити всі значення Y на групи за значеннями X. Знайти середнє значення Y всередині кожної групи ($\overline{y_j}$) та порахувати розкид цих групових середніх щодо загального середнього. Розраховуємо її за формулою:

$$\sigma^2_{\text{міжгр}} = \frac{1}{n} \cdot \sum_{j=1}^{k} n_j \cdot \left(\overline{y_j} - \overline{y}\right)^2 , \tag{7.34}$$

де k – кількість унікальних значень (груп) фактору X, n_j – кількість спостережень в кожній j-групі (частота), $\overline{y_j}$ – середнє значення Y безпосередньо для j-групи, $\overline{y}$ – загальне середнє арифметичне всіх значень Y.

Зауваження 7.5. Кореляційне відношення η адекватно працює тільки тоді, коли дані за фактором X можуть бути згруповані (або за однаковими значеннями, або за інтервалами). Якщо кожне значення X унікальне, то кореляційне відношення η завжди дорівнюватиме 1, що не має сенсу.

Кореляційне відношення η розраховується, якщо є підозра на нелінійність або коефіцієнт кореляції r виявився підозріло низьким.

Означення 7.3. Різниця (η^2–r^2) називається *показником нелінійності*. Якщо вона велика, використання лінійних моделей (регресії) призведе до грубих помилок у прогнозі.

Порівнюємо η^2 та r^2.

Якщо $\eta^2 \approx r^2$, то зв'язок лінійний ($\eta^2 - r^2 \leq 0{,}1$). Достатньо знаходження коефіцієнта лінійної кореляції Пірсона. Для перевірки значущості коефіцієнта кореляції застосовують t-критерій Стьюдента.

Якщо $\eta^2 > r^2$ (суттєва різниця), то зв'язок нелінійний. Застосовуємо F-критерій Фішера. Це спосіб перевірити, чи справді різниця між середніми значеннями в групах є суттєвою, чи це просто випадковий розкид.

Знаходимо значення $F_{\text{факт}}$ за формулою:

$$F_{\text{факт}} = \frac{\eta^2 / (k-1)}{(1-\eta^2)/(n-k)}, \tag{7.35}$$

де n – загальний обсяг вибірки (загальна кількість випробувань чи спостережень), η^2 – квадрат кореляційного відношення (показує частку

дисперсії, пояснену впливом фактору), k – кількість груп за фактором X (кількість рівнів незалежної змінної).

По суті формули: у чисельнику маємо «пояснену» фактором частину мінливості (містить η^2 і ділиться на свої степені свободи (k–1), що відображає кількість міжгрупових порівнянь), а знаменнику – «помилку», яка пояснюється фактором (містить залишок ($1–\eta^2$) і ділиться на свої степені свободи ($n– k$), що відображає кількість степенів свободи всередині груп).

Чим більше F, тим сильніший зв'язок.

Для знаходження критичного значення критерію Фішера $F_{кр}$ необхідно знати три параметри: рівень значущості α (зазвичай 0,05 або 0,01), число степенів свободи $k_1= k–1$ і $k_2=n–k$.

За таблицею А.11 (див. Додаток А) для степенів свободи k_1 (горизонталь) та k_2 (вертикаль) і заданим рівнем значущості α знаходимо критичне значення $F_{кр}$. *Пояснення*: стовпці (k_2) – це степені свободи, пов'язані з фактором (кількістю груп): k −1, а рядки (k_1) – це степені свободи, пов'язані з помилкою (кількістю спостережень): $n− k$.

Якщо $F_{факт}>F_{кр}$, то зв'язок статистично значущий (це не випадковість). Якщо $F_{факт}\leq F_{кр}$, то зв'язок недоведений (відмінності між групами випадкові, потрібно більше даних).

Зауваження 7.6. Критерій Фішера дуже чутливий до обсягу вибірки. На величезних вибірках навіть нікчемний зв'язок може виявитися «значущим», тому завжди треба дивитись на сам коефіцієнт η^2 (силу зв'язку).

При дослідженні можна користуватись таблицею 7.5:

Таблиця 7.5

Метод	Зв’язок	Критерій	Степені свободи
Пірсона (r_{xy})	Лінійний	t-критерій Стьюдента	n–2
Спірмена (ρ)	Ранговий (монотонний)	t-критерій Стьюдента	n–2
Кореляційне відношення (η)	Будь-який закономірний	F-критерій Фішера	$k_1= k–1$, $k_2=n–k$

Приклад 62. Досліджують, як кількість годин сну (*X*) впливає на концентрацію уваги (*Y*) за 10-бальною шкалою. Існує 6 піддослідних, які розбиті на 3 групи за кількістю годин сну.Отримали наступні дані для групи 1 (4 години сну): 2, 4 (бали); для групи 2 (8 годин сну): 9, 10 (балів); для групи 3 (12 годин сну): 5, 6 (балів).

Розв'язання. За отриманими даними складено таблицю 7.6, де $\overline{y_j}$ – середнє значення *Y* безпосередньо для *j*- групи.

Таблиця 7.6

Група (*X*, год.)	Значення уваги (*Y*)	Групове середнє ($\overline{y_j}$)	Кількість осіб (n_j)
4	2, 4	3	2
8	9, 10	9,5	2
12	5, 6	5,5	2

Визначити тісноту зв'язку між кількістю годин сну та концентрацію уваги і який саме зв'язок.

Розв'язання. Знаходимо загальне середнє $\overline{y}$ за формулою (7.9):

$$\overline{y}=(2+4+9+10+5+6)/6=36/6=6.$$

Загальну дисперсію знаходимо за формулою (7.33)

$$\sigma^2_{\text{загал}}=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=((2-6)^2+(4-6)^2+(9-6)^2+(10-6)^2+(5-6)^2+(6-6)^2)/6=$$

$$=(16+4+9+16+1+0)/6=46/6\approx 7{,}6667.$$

Міжгрупову дисперсію розраховуємо за формулою (7.34), де $k=3$, $n_j=2$ ($j=1,2,3$), $\overline{y_j}$ беремо з таблиці (3, 9.5, 5.5). Тоді

$$\sigma^2_{\text{міжгр}}=(2\cdot(3-6)^2+2\cdot(9.5-6)^2+2\cdot(5.5-6)^2)/6=(18+24.5+0.5)/6=$$

$$=43/6\approx 7{,}1667.$$

Підставляємо знайдені значення у формулу (7.32) для розрахунку кореляційного відношення:

$$\eta=\sqrt{\frac{7{,}1667}{7{,}6667}}\approx\sqrt{0{,}9348}\approx 0{,}9669.$$

Значення $\eta\approx 0{,}9669$ свідчить про дуже сильний зв'язок. Сон майже повністю визначає концентрацію. Висновок: У разі фактор *X* пояснює

93,5% ($\eta^2 \cdot 100\%$) всієї варіації уваги, але зв'язок носить нелінійний характер.

Для розрахунку коефіцієнта Пірсона потрібні пари (x,y). Маємо 3 групи по 2 особи: група 1 (x=4): y_1=2, y_2=4; група 2 (x=8): y_3=9, y_4=10; група 3 (x=12): y_5=5, y_6=6. За формулою (7.9) знайдемо вибіркові середні $\bar{x}$, $\bar{y}$ випадкових величин X та Y:

$\bar{x}$=(4+4+8+8+12+12)/6=48/6=8, $\bar{y}$=(2+4+9+10+5+6)/6=36/6=6.

Розрахуємо $\Delta x = x_i - \bar{x}$, $\Delta y = y_i - \bar{y}$ та складаємо таблицю 7.7:

Таблиця 7.7

X	Y	Δx	Δy	$\Delta x \cdot \Delta y$	$(\Delta x)^2$	$(\Delta y)^2$
4	2	–4	–4	16	16	16
4	4	–4	–2	8	16	4
8	9	0	3	0	0	9
8	10	0	4	0	0	16
12	5	4	–1	–4	16	1
12	6	4	0	0	16	0
$\bar{x}$=8	$\bar{y}$=6			Σ=20	Σ=64	Σ=46

За формулою (7.16) знаходимо коефіцієнт лінійної кореляції Пірсона

$$r_{xy} = \frac{20}{\sqrt{64} \cdot \sqrt{46}} = \frac{20}{\sqrt{2944}} \approx \frac{20}{54{,}2586} \approx 0{,}3686.$$

Оскільки коефіцієнт лінійної кореляції Пірсона r_{xy}≈0,3685 суттєво нижчий за кореляційне відношення η≈0,9669, то лінійна модель неадекватно описує залежність. Високе значення η при низькому r_{xy} вказує на наявність сильного, але криволінійного зв'язку між кількістю сну та концентрацією уваги.

Відповідь: η≈0,9669; r_{xy}≈0,3685. Між кількістю сну та концентрацією уваги існує криволінійний зв'язок.

Приклад 63. Є дані про стаж роботи (X, років) та продуктивність праці (Y, шт./год) для n=8 співробітників, подані у вигляді таблиці:

i	1	2	3	4	5	6	7	8
X	3	4	6	2	1	8	7	9
Y	7	10	11	6	4	14	13	15

Знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості α=0,05.

Розв'язання. За умовою дані несгруповані. Вибірковий коефіцієнт лінійної кореляції Пірсона знайдемо за формулою (7.16). Розрахуємо $\Delta x = x_i - \overline{x}$, $\Delta y = y_i - \overline{y}$ та складемо таблицю 7.8:

Таблиця 7.8

i	X	Y	Δx	Δy	$\Delta x \cdot \Delta y$	$(\Delta x)^2$	$(\Delta y)^2$
1	3	7	–2	–3	6	4	9
2	4	10	–1	0	0	1	0
3	6	11	1	1	1	1	1
4	2	6	–3	–4	12	9	16
5	1	4	–4	–6	24	16	36
6	8	14	3	4	12	9	16
7	7	13	2	3	6	4	9
8	9	15	4	5	20	16	25
Σ	40	80			81	60	112
	$\overline{x}$=5	$\overline{y}$=10					

Тоді

$$r_{xy} = \frac{\sum_{i=1}^{n} \Delta x \cdot \Delta y}{\sqrt{\sum_{i=1}^{n} (\Delta x)^2} \cdot \sqrt{\sum_{i=1}^{n} (\Delta y)^2}} = \frac{81}{\sqrt{60} \cdot \sqrt{112}} \approx \frac{81}{81{,}9756} \approx 0{,}988.$$

Отримане значення $r_{xy}\approx$0,988 говорить про наявність прямого, практично функціонального (дуже високого за шкалою Чеддока) лінійного зв'язку між стажем і продуктивністю.

Оскільки маємо лінійний зв'язок між стажем і продуктивністю, то знайдемо коефіцієнт детермінації R^2 за формулою (7.25)

$$R^2=(r_{xy})^2=(0{,}988)^2\approx 0{,}976.$$

Це означає, що продуктивність праці на 97,6% залежить від стажу роботи працівника. Лише 2,4% змін продуктивності пояснюються іншими причинами (здоров'я, настрій, складність конкретної деталі тощо). Можемо зробити висновок, що модель має виняткову точність.

Для перевірки гіпотези про значущість коефіцієнта кореляції розглянемо дві гіпотези:

Нульова гіпотеза H_0: "Зв'язку між стажем і продуктивністю праці немає, результат випадковий".

Альтернативна гіпотеза H_1: "Зв'язок статистично значущий".

Обчислюємо фактичне значення t-критерію за формулою (7.23):

$$t_{\text{факт}} = \frac{0{,}988 \cdot \sqrt{8-2}}{\sqrt{1-(0{,}988)^2}} \approx \frac{0{,}988 \cdot \sqrt{6}}{0{,}1545} \approx 15{,}664.$$

За таблицею А.4 (див. Додаток А) для степенів свободи $k = n-2=6$ та заданим рівнем значущості $\alpha=0{,}05$ знаходимо критичне значення $t_{\text{кр}}=2{,}447$.

Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (15,664>2,447), то відхиляємо гіпотезу H_0.

З ймовірністю 95% стверджуємо, що зв'язок між стажем і продуктивністю праці статистично значущий.

Відповідь: $r_{xy} \approx 0{,}988$. Зв'язок між стажем і продуктивністю праці статистично значущий.

Приклад 64. Досліджують зв'язок між віком співробітників (X) та їх продуктивністю у балах (Y) на великому підприємстві. Загальна вибірка $n=100$ осіб. Дані сгрупували по 5 інтервалів для кожної ознаки та склали таблицю 7.9:

Таблиця 7.9

X(вік) (x_i) / Y(бали) (y_i)	20−25 (22,5)	25−30 (27,5)	30−35 (32,5)	35−40 (37,5)	40−45 (42,5)	n_y
80−90 (85)	−	−	2	5	3	10
70−80 (75)	−	4	12	10	4	30
60−70 (65)	5	15	8	2	−	30
50−60 (55)	8	7	6	−	−	20
40−50 (45)	7	3	−	−	−	10
n_x	20	29	27	17	7	n=100

Знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості $\alpha=0{,}05$.

Розв'язання. В таблицю 7.9 вносимо значення середин інтервалів для X і Y, значення n_{x_i} та n_{y_j}, розраховані за формулами (7.2) і (7.3), та значення n_x та n_y, розраховані за формулою (7.1).

Для спрощення розрахунків введемо умовні варіанти:
для X: виберемо центр C_1=32,5 (середина), крок h_1=5; для Y: виберемо центр C_2=65 (середина), крок h_2=10.

Тоді $u_i=(x_i-C_1)/h_1=(x_i-32,5)/5$ і $v_j=(y_j-C_2)/h_2=(y_j-65)/10$.

Для розрахунку вибіркового коефіцієнта кореляції в цьому випадку використовуємо формулу (7.22).

Знайдемо суми: $\Sigma n_v\, v$, $\Sigma n_u\, u$, $\Sigma n_v\, v^2$, $\Sigma n_u\, u^2$. Складемо перетворену кореляційну таблицю з умовними варіантами, в яку внесемо значення n_u, n_v та відповідні добутки (табл. 7.10):

Таблиця 7.10

V \ U	–2	–1	0	1	2	n_v	$n_v \cdot v$	$n_v \cdot v^2$
2	–	–	2	5	3	**10**	20	40
1	–	4	12	10	4	**30**	30	30
0	5	15	8	2	–	**30**	0	0
–1	8	7	5	–	–	**20**	–20	20
–2	7	3	–	–	–	**10**	–20	40
n_u	**20**	**29**	**27**	**17**	**7**	$\boldsymbol{n}$**=100**	**Σ=10**	**Σ=130**
$n_u\, u$	–40	–29	0	17	14	**Σ=–38**		
$n_u\, u^2$	80	29	0	17	28	**Σ=154**		

Тоді $\bar{u} = \frac{1}{n} \cdot \sum n_u \cdot u$ =–38/100=–0,38; $\bar{v} = \frac{1}{n} \cdot \sum n_v \cdot v$ =10/100=0,1.

$\sigma_u^2 = \frac{1}{n} \cdot \sum n_u \cdot u^2 - \left(\bar{u}\right)^2$ =154/100–(–0,38)^2=1,54–0,1444=1,3956; $\sigma_u \approx 1,18$.

$\sigma_v^2 = \frac{1}{n} \cdot \sum n_v \cdot v^2 - \left(\bar{v}\right)^2$ =130/100–(0,1)^2=1,3–0,01=1,29; $\sigma_v \approx 1,1358$.

$\sum n_{uv} uv$=5·1·2+3·2·2+4·(–1) 1+10·1·1+4·2·1+8·(–2) (–1)+7·(–1) (–1)+7 (–2) (–2)+3 (–1) (–2)=93.

$$r_{xy}=\frac{\sum n_{uv}uv-n\cdot\bar{u}\cdot\bar{v}}{n\cdot\sigma_u\cdot\sigma_v}=\frac{93-100\cdot(-0{,}38)\cdot 0{,}1}{100\cdot 1{,}18\cdot 1{,}1358}=\frac{96{,}8}{134{,}0244}\approx 0{,}72.$$

Отримане значення $r_{xy}\approx 0{,}72$ згідно шкали Чеддока свідчить про високий (сильний) додатний лінійний зв'язок.

Оскільки маємо лінійний зв'язок між віком і продуктивністю, то знайдемо коефіцієнт детермінації R^2 за формулою (7.25)

$$R^2=(r_{xy})^2=(0{,}72)^2\approx 0{,}5184.$$

Це означає, що продуктивність праці на 51,84% залежить від віку працівника. Інші 48,16% припадають на фактори, не враховані в даній моделі (стаж на конкретній посаді, рівень освіти, індивідуальні психофізіологічні особливості та інше).

Перевіримо значущість коефіцієнта кореляції. Сформулюємо гіпотези.

Нульова гіпотеза H_0: "Лінійний взаємозв'язок між віком співробітників та їх продуктивністю у генеральній сукупності відсутній".

Альтернативна гіпотеза H_1: "Існує статистично значущий лінійний взаємозв'язок між віком співробітників та їх продуктивністю".

Обчислюємо фактичне значення t-критерію за формулою (7.23):

$$t_{\text{факт}}=\frac{0{,}72\cdot\sqrt{100-2}}{\sqrt{1-(0{,}72)^2}}=\frac{0{,}72\cdot\sqrt{98}}{\sqrt{0{,}4816}}\approx\frac{7{,}1276}{0{,}694}\approx 10{,}27.$$

де $r_{xy}=0{,}72$ – розрахований вибірковий коефіцієнт кореляції, $n=100$ – обсяг вибірки (кількість пар даних).

За таблицею А.4 (див. Додаток А) для степенів свободи $k=n-2=98$ та заданим рівнем значущості $\alpha=0{,}05$ критичне значення $t_{\text{кр}}$ знаходиться в інтервалі [1,98;2,00].

Оскільки $t_{\text{факт}}$ значно перевищує верхню межу даного інтервалу, то відхиляємо гіпотезу H_0. Зв'язок між віком та продуктивністю статистично значущий.

Відповідь: $r_{xy}\approx 0{,}72$. Зв'язок між стажем і продуктивністю праці статистично значущий.

Приклад 65. Для 100 автомобілів Hissan Juke з двигуном 1,6 л (передньоприводна версія) та варіатором CVT дослідили витрати пального (Y, л/100 км, бензин А-95) від швидкості (X, км/год.) при

ідеальних умовах (без вітру, маневрів та підйомів). Треба знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості α=0,05.

Розв'язання. За даними складаємо таблицю 7.11, в яку внесемо значення n_{x_i}, n_{y_j}, розраховані за формулами (7.2) і (7.3), та значення n_x і n_y, розраховані за формулою (7.1).

Таблиця 7.11

Y \ X	**40**	**60**	**80**	**100**	n_y
5	6	4	–	–	**10**
5,5	5	10	–	–	**15**
6,5	–	10	20	5	**35**
7,5	–	5	15	20	**40**
n_x	**11**	**29**	**35**	**25**	**100**

Вибірковий коефіцієнт лінійної кореляції Пірсона знайдемо за формулою (7.19). Вибіркові середні $\bar{x}$, $\bar{y}$ випадкових величин X та Y знаходимо за формулами (7.18), де k=m=4.

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} n_{x_i} \cdot x_i = (11 \cdot 40 + 29 \cdot 60 + 35 \cdot 80 + 25 \cdot 100)/100 = 7480/100 = 74{,}8.$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{j=1}^{m} n_{y_j} \cdot y_j = (10 \cdot 5 + 15 \cdot 5{,}5 + 35 \cdot 6{,}5 + 40 \cdot 7{,}5)/100 = 660/100 = 6{,}6.$$

Вибіркові середньо квадратичні відхилення σ_x, σ_y випадкових величин X та Y знаходимо за формулами (7.20), де k=m=3.

$$\frac{1}{n} \cdot \sum_{i=1}^{k} n_{x_i} \cdot x_i^2 = (11 \cdot 1600 + 29 \cdot 3600 + 35 \cdot 6400 + 25 \cdot 10000)/100 =$$

$$= 586000/100 = 5860.$$

$$\frac{1}{n} \cdot \sum_{j=1}^{m} n_{y_j} \cdot y_j^2 = (10 \cdot 25 + 15 \cdot 30{,}25 + 35 \cdot 42{,}25 + 40 \cdot 56{,}25)/100 =$$

$$= 4432{,}50/100 = 44{,}325.$$

$$\sigma_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 \cdot n_{x_i} - \left(\bar{x}\right)^2} = \sqrt{5860 - 5595{,}04} = \sqrt{264{,}96} \approx 16{,}2776.$$

$$\sigma_y = \sqrt{\frac{1}{n}\cdot\sum_{j=1}^{m} y_j^2 \cdot n_{y_j} - \left(\bar{y}\right)^2} = \sqrt{44{,}325 - 43{,}56} = \sqrt{0{,}765} \approx 0{,}8746.$$

Знайдемо суми добутків ($\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij}$). Потрібно перемножити $x \cdot y \cdot n$ для кожної заповненої клітини таблиці:

$$\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} = \sum_{i=1}^{k} x_i \cdot \left(\sum_{j=1}^{m} y_j n_{ij}\right) = 40\cdot(5\cdot 6+5{,}5\cdot 5)+60\cdot(5\cdot 4+5{,}5\cdot 10+$$
$$+6{,}5\cdot 10+7{,}5\cdot 5)+80\cdot(6{,}5\cdot 20+7{,}5\cdot 15)+100\cdot(6{,}5\cdot 5+7{,}5\cdot 20)=40\cdot 57{,}5+$$
$$+60\cdot 177{,}5+80\cdot 242{,}5+100\cdot 182{,}5=2300+10650+19400+18250=$$
$$=50600.$$

Тоді $\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} = 50600/100 = 506.$

Вибірковий коефіцієнт лінійної кореляції Пірсона

$$r_{xy} = \frac{506 - 74{,}8\cdot 6{,}6}{16{,}2776\cdot 0{,}8746} \approx \frac{506 - 493{,}68}{14{,}2364} \approx \frac{12{,}32}{14{,}2364} \approx 0{,}865.$$

Отримане значення $r_{xy} \approx 0{,}865$ згідно шкали Чеддока свідчить про високий (сильний) додатний лінійний зв'язок.

Оскільки маємо лінійний зв'язок між витратами пального і швидкістю, то знайдемо коефіцієнт детермінації R^2 за формулою (7.25)

$$R^2 = (r_{xy})^2 = (0{,}865)^2 \approx 0{,}7482.$$

Це означає, що витрати пального на 74,82% залежать від швидкості. Інші 25,18% припадають на фактори, не враховані в даній моделі (рік випуску, тиск у шинах, завантаження та інше).

Перевіримо значущість коефіцієнта кореляції. Сформулюємо гіпотези.

Нульова гіпотеза H_0: "Зв'язку між швидкістю та витратою пального немає".

Альтернативна гіпотеза H_1: "Існує статистично значущий лінійний взаємозв'язок між швидкістю та витратою пального".

Обчислюємо фактичне значення t-критерію за формулою (7.22):

$$t_{\text{факт}} = \frac{0{,}865\cdot\sqrt{100-2}}{\sqrt{1-(0{,}865)^2}} \approx \frac{0{,}865\cdot\sqrt{98}}{\sqrt{0{,}2518}} \approx \frac{8{,}563}{0{,}5018} \approx 17{,}065.$$

де r_{xy}=0,865 – розрахований вибірковий коефіцієнт кореляції, n=100 – обсяг вибірки (кількість пар даних), $n-2=k$ – число степенів свободи.

За таблицею А.4 (див. Додаток А) для степенів свободи $k= n-2=98$ та заданим рівнем значущості α=0,05 критичне значення $t_{кр}$ знаходиться в інтервалі [1,98;2,00].

Оскільки $t_{факт}$ значно перевищує верхню межу даного інтервалу, то відхиляємо гіпотезу H_0. Зв'язок між швидкістю та витратою пального статистично значущий.

Відповідь: $r_{xy}\approx$0,865. Зв'язок між швидкістю та витратою пального статистично значущий.

Приклад 66. Проводять тестування n=6 студентів: один тест (величина X) на логіку, інший (величина Y) – на уважність. Для тесту на логіку отримали дані для студентів (у балах) x_i: 20, 55, 30, 60, 40, 65. Для тесту на уважність отримали дані для студентів (у балах) y_i: 15, 40, 10, 50, 35, 55. Потрібно з'ясувати, чи є зв'язок між цими навичками при рівні значущості α=0,05.

Розв'язання. Оскільки даних n<40 (n=6), то для з'ясування наявності зв'язку між навичками застосуємо коефіцієнт рангової кореляції Спірмена. Ранжуємо значення величина X та Y. Для зручності складемо таблицю 7.12.

Таблиця 7.12

Студент	Тест на логіку x_i	Тест на уважність y_i	Ранг R_x	Ранг R_y	$d=R_x-R_y$	d^2
1	20	15	1	2	–1	1
2	55	40	4	4	0	0
3	30	10	2	1	1	1
4	60	50	5	5	0	0
5	40	35	3	3	0	0
6	65	55	6	6	0	0
Σ					**0**	**2**

Розраховуємо коефіцієнт рангової кореляції $\rho_{факт}$ за формулою (7.28), оскільки однакові ранги відсутні:

$$\rho_{факт}=1-6\cdot\frac{2}{6\cdot\left(6^2-1\right)}=1-\frac{2}{35}\approx 0{,}94\,.$$

Коефіцієнт 0,94 говорить про дуже високий додатний зв'язок. Це означає, що ті, хто добре розв'язує завдання на логіку, майже завжди показують відмінні результати і в тестах на уважність ($\rho_{факт}>0$).

Сформулюємо гіпотези.

Нульова гіпотеза H_0: "Статистично значущий взаємозв'язок між ознаками X та Y відсутній".

Альтернативна гіпотеза H_1: "Існує статистично значущий взаємозв'язок між ознаками X і Y".

За таблицею А.12 (див. Додаток А) для кількості значень n=6 та заданим рівнем значущості α=0,05 критичне значення $\rho_{кр}$=0,85.

Оскільки $\rho_{факт}>\rho_{кр}$, то відхиляємо гіпотезу H_0. Приймається альтернативна гіпотеза H_1 про наявність статистично значущого взаємозв'язку.

Відповідь: Взаємозв'язок статистично значущий.

Приклад 67. Досліджують вплив отриманих балів за тест із розділу «Матстатистика» (макс. 40 балів) та часу, витраченого на його виконання (у хвилинах) для 6 студентів. Отримали дані для *X:* 10,20,20,30,30,40 і для *Y*: 5,15,10,20,15,20. Необхідно встановити наявність та вид зв'язку між отриманими балами та витраченим часом при рівні значущості α=0,05.

Розв'язання. Оскільки даних $n<40$ (n=6), то для з'ясування наявності зв'язку між отриманими балами та витраченим часом застосуємо коефіцієнт рангової кореляції Спірмена. Ранжуємо значення величина X та Y. Для зручності складемо таблицю 7.13, з урахуванням пов'язаних рангів.

Таблиця 7.13

№	Тест (X, бали)	Ранг R_x	Час (Y, хвилини)	Ранг R_y	$d=R_x-R_y$	d^2
1	10	1	5	1	0	0
2	20	2,5	15	3,5	–1	1
3	20	2,5	10	2	0,5	0,25
4	30	4,5	20	5,5	–1	1
5	30	4,5	15	3,5	1	1
6	40	6	20	5,5	0,5	0,25
Σ					**0**	**3,5**

Оскільки маємо для R_x і R_y пов'язані ранги, то вводимо поправки поправку T_X і T_Y, розраховані за формулою (7.27):

Для X: дві групи по 2 повтори (20, 20 та 30, 30).

$$T_X=(2^3-2)/12+(2^3-2)/12=0,5+0,5=1.$$

Для Y: дві групи по 2 повтори (15, 15 та 20, 20).

$$T_Y=(2^3-2)/12+(2^3-2)/12=0,5+0,5=1.$$

За наявності однакових рангів коефіцієнт рангової кореляції $\rho_{\text{факт}}$ розраховуємо за формулою (7.28):

$$\rho_{\text{факт}} = 1-6\cdot\frac{3,5+1+1}{6\cdot(36-1)} = 1-\frac{5,5}{35} \approx 1-0,1571 = 0,8429 .$$

Значення $\rho_{\text{факт}}$ =0,8429 свідчить про високий додатний зв'язок.

Для перевірки існування нелінійного зв'язку, знайдемо кореляційне відношення за формулою (7.32). Воно показує, яку частку варіації залежної змінної Y пояснює угруповання фактором X. Для розрахунку потрібно згрупувати Y за значеннями X:

X=10: Y={5},
X=20: Y={15, 10} (середнє 12.5),
X=30: Y={20, 15} (середнє 17.5),
X=40: Y={20}.

Знаходимо загальне середнє $\bar{y}$ за формулою (7.9):

$$\bar{y}=(5+15+10+20+15+20)/6=85/6\approx 14,17.$$

Знайдемо суми для чисельника і знаменника формули (7.32) (n=6, k=4):

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=(5-14,17)^2+(15-14,17)^2+(10-14,17)^2+(20-14,17)^2+$$
$$+(15-14,17)^2+(20-14,17)^2=84,0889+0,5889+17,3889+33,9889+0,6889+$$
$$+33,9889=170,8334.$$

$$\sum_{j=1}^{k} n_j\cdot\left(\overline{y_j}-\bar{y}\right)^2=1\cdot(5-14,17)^2+2\cdot(12,5-14,17)^2+2\cdot(17,5-14,17)^2+$$
$$1\cdot(20-14,17)^2=84,0889+5.5778+22.1778+33,9889=145,8334.$$

Підставляємо знайдені значення сум у формулу (7.32) для розрахунку кореляційного відношення:

$$\eta=\sqrt{\frac{145,8334}{170,8334}}\approx\sqrt{0,8537}\approx 0,924 .$$

Значення $\eta \approx 0{,}924$ свідчить про практично функціональну залежність.

Перевіримо значущість кореляційного відношення за F-критерієм Фішера. Сформулюємо гіпотези.

Нульова гіпотеза H_0: "Кореляційне відношення в генеральній сукупності дорівнює нулю ($\eta=0$). Вплив фактору X (Бали) на результативну ознаку Y (Час) у генеральній сукупності відсутній (зв'язок випадковий)".

Альтернативна гіпотеза H_1: "Кореляційне відношення суттєво відмінно від нуля ($\eta \neq 0$). вплив фактору X на Y статистично значущий".

Знайдемо значення $F_{\text{факт}}$ за формулою (7.35):

$$F_{\text{факт}} = \frac{0{,}924^2 \cdot (6-4)}{(1-0{,}924^2)\cdot(4-1)} \approx \frac{0{,}8538 \cdot 2}{0{,}1462 \cdot 3} \approx 3{,}89.$$

Для рівня значущості $\alpha=0{,}05$ і степенів свободи $k_1=k-1=3$ (між групами) та $k_2=n-k=2$ (всередині груп) за таблицею А.11 (див. Додаток А) знаходимо критичне значення $F_{\text{кр}}=19{,}16$.

Оскільки $F_{\text{факт}} \leq F_{\text{кр}}$ (3,89<19,16), дані, що спостерігаються, не дають підстав для відхилення нульової гіпотези H_0.

Щоб з'ясувати наскільки зв'язок відхиляється від прямої лінії, розрахуємо коефіцієнт кореляції Пірсона r_{xy}, та порівняємо його з кореляційним відношенням $\eta=0{,}924$.

Знаходимо загальне середнє $\bar{x}$ за формулою (7.9):

$$\bar{x} = (10+20+20+30+30+40)/6 = 150/6 = 25.$$

Знаходимо суму квадратів відхилень X:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = (10-25)^2+(20-25)^2+(20-25)^2+(30-25)^2+(30-25)^2+$$

$$+(40-25)^2 = 225+25+25+25+25+225 = 550.$$

Сума квадратів відхилень Y знайдена вище $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 = 170{,}8334$.

Знаходимо суму добутків відхилень:

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (10-25)\,(5-14{,}17)+(20-25)\,(15-14{,}17)+$$

$$+(20-25)\,(10-14{,}17)+(30-25)\,(20-14{,}17)+(30-25)\,(15-14{,}17)+$$

$$+(40-25)\,(20-14{,}17) = 137{,}55-4{,}15+20{,}85+29{,}15+4{,}15+87{,}45 = 275.$$

Коефіцієнт кореляції Пірсона знаходимо за формулою (7.16):

$$r_{xy} = \frac{275}{\sqrt{550} \cdot \sqrt{170{,}8334}} \approx \frac{275}{306{,}5263} \approx 0{,}897 .$$

Порівняємо η=0,924 (загальна тіснота зв'язку, будь-якої форми) та r_{xy}=0,897 (тіснота тільки лінійного зв'язку). η^2=0,854, r^2_{xy}=0,805, а різниця 0,854–0,805=0,049 невелика (загалом близько 5%). Це означає, що залежність у даних близька до лінійної. Хоча групові середні і описують дані трохи точніше ($\eta > r_{xy}$), проста пряма лінія також дає дуже високий коефіцієнт детермінації (80%). У статистиці таку різницю на маленьких вибірках зазвичай вважають несуттєвою. Тобто, якщо говорити, що зв'язок лінійний, то великої помилки не буде, але формально значення η дійсно свідчать про невеликі нелінійні коливання.

Незважаючи на високе вибіркове значення кореляційного відношення (η=0,924), на рівні значущості 5% зв'язок між балами та часом визнається статистично незначним. Отриманий результат пояснюється малим обсягом вибірки (n=6), при якому навіть сильна зовнішня схожість залежності не дозволяє виключити фактор випадковості.

Відповідь: На досліджуваній вибірці спостерігається сильна, переважно лінійна залежність із легким нелінійним характером. Однак, через критично малий обсяг вибірки (n=6), виявлений зв'язок не є статистично значущим на рівні 5%.

7.2 Регресійний аналіз

Кореляційний аналіз з’ясовує наявність зв’язку між однією змінною (залежною Y) і іншою змінною (незалежною X) та тісноту цього зв’язку, а регресійний аналіз визначає який вплив одна змінна має на іншу і дозволяє передбачати значення залежної змінної Y за будь-якого заданого значення фактору X. Якщо кореляція статистично незначима (близька до 0), то будувати регресійну модель немає сенсу, оскільки вона не матиме передбачуваної сили.

Означення 7.4. *Регресія* – це математична модель (рівняння регресії), яка описує залежність однієї змінної (залежної, результативної, Y) від однієї чи кількох інших змінних (незалежних, факторних, X).

Основні завдання:

1) Встановлення форми зв'язку: вибір виду рівняння (математичної функції), що найкраще описує взаємозв'язок між змінними (лінійна, параболічна тощо).

2) Оцінка параметрів: знаходження коефіцієнтів рівняння (зазвичай шляхом найменших квадратів — МНК).

3) Прогноз: передбачення значення Y при заданому X.

4) Оцінити вплив: визначити, наскільки сильно та у якому напрямку (додатному чи від'ємному) зміна незалежної змінної впливає на залежну.

Види рівнянь регресії (за кількістю незалежних змінних та формою зв'язку):

За кількістю незалежних змінних:

1) Парна (проста) регресія: Одна залежна змінна Y та одна незалежна змінна X.

2) Множинна регресія: Одна залежна змінна Y і дві або більше незалежних змінних $X_1, X_2, \ldots, X_k$.

За формою зв'язку (лінії регресії):

1) Лінійна регресія: Залежність описується прямою лінією. Це найпростіший і найчастіше використовуваний вид.

2) Нелінійна регресія: Залежність описується кривою.

7.2.1 Лінійна регресія

Означення 7.4. *Лінійна регресія* – це статистичний метод, який використовується для моделювання взаємозв'язку між однією залежною змінною (яку ми хочемо передбачити або пояснити) та однією чи декількома незалежними змінними (які використовуються для передбачення/пояснення).

Галузі застосування:

Інженерія: Прогнозування зношування матеріалів, оптимізація виробничих процесів.

Сільське господарство: Прогнозування врожайності з урахуванням кількості добрив, температури, опадів.

Медицина та охорона здоров'я: Прогнозування ризику захворювань на основі факторів життя, оцінка ефективності лікування.

Бізнес та економіка: Прогнозування продажів, оцінка впливу цінової політики на попит, прогноз ВВП.

Фінанси: Оцінка ризику інвестицій, прогнозування цін акцій.

Соціологія та психологія: Вивчення впливу освіти на дохід, зв'язки між рисами особистості та поведінкою.

У найпростішому випадку проста лінійна регресія досліджує зв'язок між двома кількісними змінними. Передбачається, що взаємозв'язок між залежною змінною (Y) та незалежною змінною (X) може бути апроксимований прямою лінією. Мета – знайти рівняння цієї прямої, яке найкраще описує дані.

Лінійна парна регресійна модель має вигляд:

$$Y_{\text{теор}}=b_0+b_1\cdot X+\varepsilon, \tag{7.36}$$

де: $Y_{\text{теор}}$ – передбачене (очікуване) значення залежної змінної; X – незалежна змінна (змінна, що пояснює); b_0 (вільний член) – перетин: очікуване значення $Y_{\text{теор}}$, коли X дорівнює 0. Це точка, де лінія регресії перетинає вісь Y; b_1 (коефіцієнт регресії) – коефіцієнт нахилу: показує, наскільки в середньому зміниться $Y_{\text{теор}}$ при зміні X на одну одиницю. Це "нахил" лінії регресії. Якщо $b_1>0$, то зв'язок прямий. Якщо $b_1<0$, то зв'язок зворотний; ε – помилка: є випадковою, незрозумілою частиною змінної $Y_{\text{теор}}$, яка не може бути пояснена змінною X. Включає вплив всіх інших факторів, які не враховані в моделі, а також випадковий шум.

Оцінка якості моделі (*надійності*) регресійної моделі (для простої лінійної регресії) визначається коефіцієнтом детермінації (R^2):

$$R^2=r_{xy}^2, \tag{7.37}$$

який набуває значення від 0 до 1 (r_{xy} – коефіцієнт кореляції Пірсона). Він показує, яка частка варіації залежної змінної (Y) пояснюється варіацією незалежної змінної (X) за допомогою побудованої моделі. Чим ближче значення R^2 до 1, тим точніше модель підходить для прогнозування. Наприклад, якщо R^2=0,88, то це означає, що 88% варіації Y пояснюється моделлю, а 12% – випадковими факторами або іншими змінними.

Для того щоб лінійна регресія була коректною, повинні виконуватися наступні умови:

Лінійність: залежність між змінними має бути візуально близька до прямої лінії (перевіряється по полю кореляції).

Незалежність спостережень: дані щодо однієї змінної не повинні залежати від даних щодо іншої.

Розкид (дисперсія) помилок має бути приблизно однаковим на всьому діапазоні значень X.

Нормальність: помилки моделі мають бути розподілені за нормальним законом.

Метод найменших квадратів (МНК)

Найбільш поширений метод знаходження коефіцієнтів регресії (b_0, b_1 тощо) – це *метод найменших квадратів* (МНК). Його суть полягає у знаходженні такої лінії (або кривої), яка мінімізує суму квадратів відхилень фактичних значень Y_i від теоретичних (лежачих на лінії) $Y_{\text{теор}}$:

$$Q = \sum_{i=1}^{n} \left(Y_i - Y_{i\,\text{теор}} \right)^2 \rightarrow \min \ , \tag{7.38}$$

де Q – цільова функція, вигляд якої залежить від способу подання даних.

Оцінкою моделі (7.36) є рівняння лінійної регресії Y на X

$$Y = b_0 + b_1 \cdot x, \tag{7.39}$$

де b_1 – вибірковий коефіцієнт регресії. Коефіцієнт регресії b_1 (його називають кутовим коефіцієнтом) показує середнє змінювання результативної ознаки Y зі збільшенням факторної ознаки X на одиницю її виміру. Якщо $b_1>0$, то зв'язок прямий, якщо $b_1<0$ – зв'язок зворотний.

Для знаходження коефіцієнтів лінійної регресії b_0 і b_1 застосуємо метод найменших квадратів (МНК). Суть методу найменших квадратів (МНК) залишається незмінною: мінімізація суми квадратів відхилень. Однак форма запису даних змінює те, як підсумовуються ці відхилення.

Цільова функція Q визначається формулою:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \to \min , \qquad (7.40)$$

де n – обсяг вибірки, Y_i – розрахункова ордината, y_i – задана (експериментальна) ордината.

Для знаходження мінімуму функції $Q(b_0,b_1)$ необхідно, перш за все, прирівняти нулю її частинні похідні по b_0 і b_1: $\frac{\partial Q}{\partial b_0} = 0$ і $\frac{\partial Q}{\partial b_1} = 0$. Отримаємо систему рівнянь для визначення b_0 і b_1:

$$\begin{cases} b_1 \cdot \sum X^2 + b_0 \cdot \sum X = \sum XY, \\ b_1 \sum X + n \cdot b_0 = \sum Y. \end{cases} \qquad (7.41)$$

Систему (7.41) називають системою нормальних рівнянь за методом найменших квадратів для визначення параметрів b_0 та b_1.

Розглянемо випадки для даних поданих набором пар (x,y) і кореляційною таблицею.

Нехай в результаті проведення n спостережень маємо вибірку з n пар чисел (x_i,y_i), i=1,2,…,n і кожна з них унікальна:

x_i	x_1	x_2	…	x_n
y_i	y_1	y_2	…	y_n

Тоді система рівнянь (7.41) для визначення b_0 і b_1 має вигляд:

$$\begin{cases} b_1 \cdot \sum_{i=1}^{n} x_i^2 + b_0 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \cdot y_i , \\ b_1 \cdot \sum_{i=1}^{n} x_i + b_0 \cdot n = \sum_{i=1}^{n} y_i . \end{cases} . \qquad (7.42)$$

Розв’язуючи її, знаходимо коефіцієнти лінійної регресії b_0 і b_1 за формулами:

$$b_1 = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{7.43}$$

$$b_0 = \frac{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \bar{y} - b_1 \cdot \bar{x}, \tag{7.44}$$

де $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ і $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$ – вибіркові середні величин X та Y відповідно.

Підставляючи значення $b_0 = \bar{y} - b_1 \cdot \bar{x}$ в рівняння регресії (7.39) отримуємо:

$$Y - \bar{y} = b_1 \cdot \left(x - \bar{x} \right). \tag{7.45}$$

Коефіцієнт b_1 в рівнянні регресії (7.45) є вибірковим коефіцієнтом регресії Y на X і позначається ρ_{yx}. Тоді регресію Y на X можна записати у вигляді:

$$Y - \bar{y} = \rho_{yx} \cdot \left(x - \bar{x} \right). \tag{7.46}$$

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

$$X = c_0 + c_1 \cdot y, \tag{7.47}$$

де c_1 – вибірковий коефіцієнт регресії.

Для знаходження коефіцієнтів лінійної регресії c_0 і c_1 можна провести перетворення, аналогічні тим, що були використані при знаходженні b_0 і b_1. Отримаємо формули:

$$c_1 = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2}, \tag{7.48}$$

$$c_0 = \frac{\sum_{i=1}^{n} y_i^2 \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i y_i}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2}. \tag{7.49}$$

Рівняння регресії (7.47) можна записати у вигляді:

$$X - \bar{x} = c_1 \cdot \left(y - \bar{y} \right). \tag{7.50}$$

Коефіцієнт c_1 в рівнянні регресії (7.50) є вибірковим коефіцієнтом регресії X на Y і позначається ρ_{xy}. Тоді регресію X на Y можна записати у вигляді:

$$X - \bar{x} = \rho_{xy} \cdot \left(y - \bar{y} \right), \tag{7.51}$$

Отже, $Y=b_0+b_1 \cdot x$ і $X=c_0+c_1 \cdot y$ – різні прямі. Перша пряма виходить у результаті розв'язання задачі про мінімізацію суми квадратів відхилень по вертикалі, а друга – при розв'язанні задачі про мінімізацію суми квадратів відхилень по горизонталі. Ці прямі перетинаються у точці з координатами $M\left(\bar{x}; \bar{y}\right)$.

Добуток коефіцієнтів регресії b_1 і c_1 дорівнює квадрату коефіцієнта кореляції.

Приклад 68. Знайти рівняння лінійної регресії $Y=b_0+b_1 \cdot x$ методом найменших квадратів та побудувати графік для наступних даних:

x_i	1	3	4	5	7
y_i	1,3	3,2	6,5	9,6	12,1

Розв'язання. Для знаходження коефіцієнтів лінійної регресії b_0 і b_1 застосуємо формули (7.43) і (7.44). Маємо $n=5$ значень змінної X і змінної Y. Знайдемо

$$\sum_{i=1}^{5} x_i = 1+3+4+5+7=20; \qquad \sum_{i=1}^{5} y_i = 1{,}3+3{,}2+6{,}5+9{,}6+12{,}1=32{,}7;$$

$$\sum_{i=1}^{5} x_i y_i = 1\cdot 1{,}3+3\cdot 3{,}2+4\cdot 6{,}5+5\cdot 9{,}6+7\cdot 12{,}1=169{,}6;$$

$$\sum_{i=1}^{5} x_i^2 = 1^2+3^2+4^2+5^2+7^2=100.$$

Знайдені значення підставимо у формули:

$$b_1 = \frac{5\cdot 169{,}6 - 20\cdot 32{,}7}{5\cdot 100 - (20)^2} = \frac{194}{100} = 1{,}94;$$

$$b_0 = \frac{100\cdot 32{,}7 - 20\cdot 169{,}6}{5\cdot 100 - (20)^2} = \frac{3270 - 3392}{100} = \frac{-122}{100} = -1{,}22.$$

Найкраща пряма, побудована за методом найменших квадратів, має вигляд: $Y = -1{,}22+1{,}94\cdot x$.

Побудуємо графік прямої лінії регресії $Y = -1{,}22+1{,}94\cdot x$ (рис. 7.1)

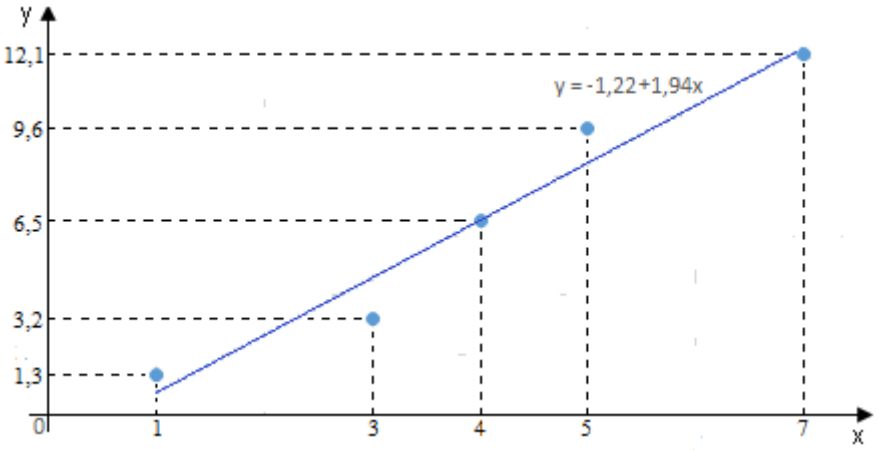


Рисунок 7.1

Відповідь: Рівняння лінійної регресії $Y = -1{,}22+1{,}94\cdot x$.

У класичній лінійній моделі МНК y_i беремо безпосередньо з даних. Однак, якщо є дані у вигляді вибірки з частотами, то в цьому випадку застосовується зважений метод найменших квадратів (ЗМНК), де

кожна точка даних зважується залежно від її частоти, тобто ЗМНК використовується, коли маємо групи спостережень, і кожна група має свою "вагу" або частоту. Основна ідея полягає в тому, щоб спостереження з більш високою частотою більш впливали на підсумковий розв'язок. Якщо вибірка для величини X має частоти, це означає, що кожне значення X спостерігалося певну кількість разів. Ці частоти й будуть у ролі "ваг".

Наприклад, маємо значення:

Значення X	x_1	x_2	…	x_k
Частота n_i	n_1	n_2	…	n_k
Значення Y	y_1	y_2	…	y_k

Тут y_i – це середнє чи типове значення Y для кожного x_i, а n_i – це частота, з якою зустрічається пара (x_i, y_i).

ЗМНК мінімізує зважену суму квадратів залишків:

$$Q(b_0, b_1) = \sum_{i=1}^{k} n_i \cdot \left(y_i - \left(b_0 + b_1 x_i\right)\right)^2 \to \min \ . \tag{7.52}$$

Формули для знаходження коефіцієнтів лінійної регресії b_0 і b_1 мають вигляд:

$$b_1 = \frac{\sum\limits_{i=1}^{k} n_i \cdot \sum\limits_{i=1}^{k} n_i x_i y_i - \sum\limits_{i=1}^{k} n_i x_i \cdot \sum\limits_{i=1}^{k} n_i y_i}{\sum\limits_{i=1}^{k} n_i \cdot \sum\limits_{i=1}^{k} n_i x_i^2 - \left(\sum\limits_{i=1}^{k} n_i x_i\right)^2}, \tag{7.53}$$

$$b_0 = \frac{\sum\limits_{i=1}^{k} n_i y_i - b_1 \cdot \sum\limits_{i=1}^{k} n_i x_i}{\sum\limits_{i=1}^{k} n_i} = \overline{y_n} - b_1 \overline{x_n}, \tag{7.54}$$

де $\overline{x_n} = \dfrac{\sum\limits_{i=1}^{k} n_i x_i}{\sum\limits_{i=1}^{k} n_i}$ і $\overline{y_n} = \dfrac{\sum\limits_{i=1}^{k} n_i y_i}{\sum\limits_{i=1}^{k} n_i}$ – зважене середнє x і y відповідно.

Використання зваженого МНК з частотами дозволяє врахувати, що деякі пари значень (x_i, y_i) зустрічаються частіше, і тому повинні мати більший вплив на визначення лінії регресії. Це робить модель адекватнішою для наявних даних.

Приклад 69. Знайти рівняння лінійної регресії $Y=b_0+b_1\cdot x$ методом зважених найменших квадратів для наступних даних:

x_i	3	2	4	3	5	2	3	4	3	3	5	2	4	3	3	2	4	5	3	4
y_i	5	5	2	4	4	3	5	3	4	6	3	4	4	2	4	4	3	8	2	3

Розв'язання. Впорядкуємо дані x_i:

2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,5.

Для застосування МНК із частотами необхідно мати таблицю з унікальними x_i, їх частотами n_i та середніми y_i. Підрахуємо їх:

$x_1=2$, $n_1=4$, $y_1=\frac{1}{4}(5+3+4+4)=\frac{16}{4}=4$;

$x_2=3$, $n_2=8$, $y_2=\frac{1}{8}(5+4+5+4+6+2+4+2)=\frac{32}{8}=4$;

$x_3=4$, $n_3=5$, $y_3=\frac{1}{5}(2+3+4+3+3)=\frac{15}{5}=3$;

$x_4=5$, $n_4=3$, $y_4=\frac{1}{3}(4+3+8)=\frac{15}{3}=5$.

Для зручності знаходження сум складемо таблицю 7.14:

Таблиця 7.14

x_i	y_i	n_i	$n_i\cdot x_i$	$n_i\cdot y_i$	$n_i\cdot x_i^2$	$n_i\cdot x_i\cdot y_i$
2	4	4	8	16	16	32
3	4	8	24	32	72	96
4	3	5	20	15	80	60
5	5	3	15	15	75	75
		$\sum_{i=1}^{4} n_i = 20$	$\sum_{i=1}^{4} n_i x_i = 67$	$\sum_{i=1}^{4} n_i y_i = 78$	$\sum_{i=1}^{4} n_i x_i^2 = 243$	$\sum_{i=1}^{4} n_i x_i y_i = 263$

Маємо $\sum_{i=1}^{4} n_i = 20$, $\sum_{i=1}^{4} n_i x_i = 67$, $\sum_{i=1}^{4} n_i y_i = 78$, $\sum_{i=1}^{4} n_i x_i^2 = 243$, $\sum_{i=1}^{4} n_i x_i y_i = 263$.

Коефіцієнти лінійної регресії b_0 і b_1 знаходимо за формулами (7.53) і (7.54).

$$b_1 = \frac{20 \cdot 263 - 67 \cdot 78}{20 \cdot 243 - (67)^2} = \frac{5260 - 5226}{4860 - 4489} = \frac{34}{371} \approx 0{,}0916;$$

$$b_0 = \frac{78 - 0{,}0916 \cdot 67}{20} \approx \frac{71{,}8628}{20} \approx 3{,}5931.$$

Рівняння зваженої лінійної регресії, отримане методом найменших квадратів, має вигляд: Y=3,5931+0,0916·x.

Відповідь: Рівняння лінійної регресії Y=3,5931+0,0916·x.

Розглянемо випадок, коли кількість спостережень n велике. Для спрощення розрахунків експериментальні дані групують, тобто будують кореляційну таблицю 7.15, де вказують частоти n_{ij} (i=1,2,…,k; j=1,2,…,m) – скільки разів зустрілася комбінація конкретного значення x_i (i=1,2,…,k) та y_j (j=1,2,…,m).

Таблиця 7.15

Y \ X	x_1	x_2	…	x_k	n_y
y_1	n_{11}	n_{21}	…	n_{k1}	n_{y_1}
y_2	n_{12}	n_{22}	…	n_{k2}	n_{y_2}
…	…	…	…	…	…
y_m	n_{1m}	n_{2m}	…	n_{km}	n_{y_m}
n_x	n_{x_1}	n_{x_2}	…	n_{x_k}	n

Якщо в таблиці дані сгруповані за інтервалами, то таблицю спрощують, вибравши середини x_i та y_j (i=1,2,…,k; j=1,2,…,m) цих інтервалів та числа n_{ij}.

У випадку кореляційної таблиці формула мінімізації цільової функції Q має вигляд:

$$Q(b_0, b_1) = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} \left(y_j - (b_0 + b_1 x_i) \right)^2 \to \min . \qquad (7.55)$$

Для знаходження коефіцієнтів регресії b_0 і b_1 за згрупованими даними формули аналогічні формулам (7.42)–(7.43) для незгрупованих даних, але суми $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$ замінюють відповідно на $\sum_{i=1}^{k} x_i \cdot n_{x_i}$, $\sum_{j=1}^{m} y_j \cdot n_{y_j}$, $\sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij}$, $\sum_{i=1}^{k} x_i^2 \cdot n_{x_i}$, $\sum_{j=1}^{m} y_j^2 \cdot n_{y_j}$.

Тоді формули для знаходження коефіцієнтів b_0 і b_1 лінійної регресії мають вигляд:

$$b_1 = \frac{n \cdot \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij} - \sum_{i=1}^{k} x_i \cdot n_{x_i} \cdot \sum_{j=1}^{m} y_j \cdot n_{y_j}}{n \cdot \sum_{i=1}^{k} x_i^2 \cdot n_{x_i} - \left(\sum_{i=1}^{k} x_i \cdot n_{x_i} \right)^2}, \qquad (7.56)$$

$$b_0 = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_{x_i} \cdot \sum_{j=1}^{m} y_j \cdot n_{y_j} - \sum_{i=1}^{k} x_i \cdot n_{x_i} \cdot \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j n_{ij}}{n \cdot \sum_{i=1}^{k} x_i^2 \cdot n_{x_i} - \left(\sum_{i=1}^{k} x_i \cdot n_{x_i} \right)^2}, \qquad (7.57)$$

За отриманим рівнянням регресії можна зробити прогноз щодо значення залежної змінної при конкретному значенні незалежної змінної, підставивши її значення в рівняння регресії. Важливим і практично значущим результатом лінійної регресії є те, що вона дозволяє «передбачати» значення залежної змінної навіть для таких незалежних значень, які реально не спостерігалися.

Оцінка якості моделі

Нехай в результаті проведення n спостережень маємо вибірку з n пар чисел (x_i, y_i), $i=1,2,\ldots,n$ і кожна з них унікальна.

Коефіцієнт детермінації R^2 – це показник якості (надійності) регресійної моделі. Він показує частку варіації (розкиду) залежної змінної Y, яка пояснюється впливом фактору X у межах цієї моделі.

Коефіцієнт детермінації набуває значення від 0 до 1.Чим ближче R^2 до 1, тим точніше модель підходить для прогнозування.

Для лінійної регресії коефіцієнт детермінації R^2 дорівнює квадрату коефіцієнта лінійної кореляції Пірсона r_{xy}: $R^2=r_{xy}^2$.

У загальному випадку коефіцієнт детермінації R^2 обчислюється за формулою:

$$R^2 = \frac{b_1 \cdot \mathrm{cov}(X,Y)}{\sigma_y^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - Y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \qquad (7.58)$$

Для висновку стосовно якості моделі за значенням R^2 прийнято: $R^2\in[0,9;1,0]$ – якість дуже висока (модель майже ідеальна); $R^2\in[0,7;0,9)$ – якість висока (хороша модель для прогнозу); $R^2\in[0,5;0,7)$ – якість помітна (модель описує основні тренди); $R^2<0,5$ – якість низька (моделі не вистачає важливих факторів). Тобто для гуманітарних наук $R^2>0,5$ – точність моделі прийнятна, а для технічних наук $R^2>0,8$ – відмінна.

Наприклад, $R^2=0,92$. Це означає, що 92% варіації Y пояснюється моделлю, а 8% – випадковими факторами або іншими змінними.

Якщо розрахований коефіцієнт детермінації R^2, то наступним логічним кроком буде перевірка F-критерію Фішера, щоб перевірити чи R^2 дійсно відображає закономірність або це просто випадковість.

Оцінка значущості моделі

Перевірка значущості рівняння регресії полягає у встановленні адекватності обраної математичної моделі фактичним даним. Ця процедура дозволяє визначити, чи достатньо врахованих у моделі пояснюючих змінних для того, щоб статистично значущо описати поведінку залежної змінної.

Оцінка значущості рівняння регресії проводиться для того, щоб з'ясувати, чи можна отримане рівняння регресії використовувати для практичного застосування, не є чи наш R^2 результатом вдалого збігу обставин.

Для перевірки значущості рівняння регресії загалом застосовують *F*-критерій Фішера. Для перевірки значущості окремих коефіцієнтів застосовують *t*-критерій Стьюдента.

При застосуванні *F*-критерія Фішера перевіряють гіпотези:

нульова гіпотеза H_0: "Модель статистично незначуща (коефіцієнти при факторах моделі рівні нулю, фактори ніяк не впливають на *Y*)";

альтернативна гіпотеза H_1: "Модель значуща (хоча б один коефіцієнт при факторах моделі відмінний від нуля)".

Для моделі парної регресії фактичне значення $F_{\text{факт}}$ визначається за формулою:

$$F_{\text{факт}} = \frac{R^2}{\left(1-R^2\right)} \cdot (n-2) = \frac{r_{xy}^2}{\left(1-r_{xy}^2\right)} \cdot (n-2), \qquad (7.59)$$

де n – обсяг вибірки, R^2 – коефіцієнт детермінації (показує, який % змін Y пояснюється отриманим рівнянням регресії), r_{xy}^2 – квадрат коефіцієнта лінійної кореляції Пірсона.

Критичне значення *F*-критерію Фішера $F_{\text{кр}}$ знаходимо за таблицею А.11 (див. Додаток А) для степенів свободи: k_1 (горизонталь) $k_1=1$ для моделі парної регресії, k_2 (вертикаль) $k_2=n-2$ для моделі парної регресії та рівню значущості α (зазвичай 0,05 або 0,01).

Якщо $F_{\text{факт}} > F_{\text{кр}}$ (взято з таблиці розподілу Фішера для заданого рівня значущості, наприклад, 0,05), то гіпотезу H_0 відхиляємо, модель вважається статистично значущою. Це означає, що її можна використовувати для прогнозування. Якщо $F_{\text{факт}} \le F_{\text{кр}}$, то гіпотезу H_0 не відхиляємо, модель статистично незначуща.

Якщо модель регресії виявляється значущою, то, використовуючи *t*- критерій Стьюдента, можна з'ясувати наскільки значущим є вплив незалежної змінної *X* на параметр *Y*.

Незміщеною оцінкою дисперсії σ^2 є величина S^2, яка визначається за формулою:

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - Y_i\right)^2 . \qquad (7.60)$$

Означення 7.5. Величина *S* називається *стандартною помилкою регресії* і є мірою розкиду залежної змінної біля лінії регресії.

Тоді стандартні помилки коефіцієнтів регресії b_0 і b_1 визначаються формулами:

$$S_{b_0} = S \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}, \qquad S_{b_1} = S \cdot \sqrt{\frac{1}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}. \qquad (7.61)$$

Знаменник показує варіацію самого фактору x. Чим більше розкиданий x, тим точніше зможемо визначити нахил прямої.

Величини S_{b_0} та S_{b_1} використовуються для побудови довірчих інтервалів, яким належать параметри істинної регресії та перевірки значущості коефіцієнтів регресії.

Оскільки коефіцієнти регресії b_0 і b_1 отримані за вибіркою, то вони є випадковими числами, тобто їх значення можуть бути лише випадково відмінні від нуля. Тому проводять перевірку значущості коефіцієнтів регресії, тобто перевірку того, чи значимо вони відмінні від нуля.

Для цього використовують перевірку гіпотез за t-критерієм Стьюдента. Перевіримо значущість коефіцієнта b_1. Сформулюємо гіпотези:

нульова гіпотеза H_0: "Коефіцієнт b_1=0 − коефіцієнт статистично незначимий. Це означає, що фактор X ніяк не впливає на Y, а те число, яке отримали при розрахунку, − просто випадковість через особливості вибірки";

альтернативна гіпотеза H_1: "Коефіцієнт $b_1 \neq 0$ − коефіцієнт статистично значущий. Фактор реально впливає на результат, його зв'язок з Y підтверджено математично".

Для моделі парної регресії розрахуємо фактичне значення t- статистики за формулою:

$$t_{\text{факт}} = \frac{b_1}{S_{b_1}}. \qquad (7.62)$$

По суті, перевіряємо, у скільки разів коефіцієнт b_1 більший за свою власну помилку.

Критичне значення t-критерію Стьюдента $t_{кр}$ знаходимо за таблицею А.4 (див. Додаток А) для степенів свободи $k=n-2$ та заданим рівнем значущості α (зазвичай 0,05 або 0,01).

Якщо $|t_{факт}|>t_{кр}$, то гіпотеза H_0 про рівність параметра $b_1=0$ відхиляється, параметр b_1 істотно відмінний від нуля, коефіцієнт значущий, а незалежна змінна X справді впливає на параметр Y.

Якщо $|t_{факт}|<t_{кр}$, то гіпотеза H_0 приймається, коефіцієнт b_1 незначущий і незалежна змінна X не впливає на параметр Y.

Довірчий інтервал коефіцієнта регресії b_1 визначається формулою:

$$b_1 \pm t_{кр} \cdot S_{b_1}. \tag{7.63}$$

Зауваження 7.7. Аналогічно перевіряється значущість вільного члена b_0 у рівнянні регресії. Для цього обчислюється t-статистика $t_{факт} = \dfrac{b_0}{S_{b_0}}$, значення якої порівнюється з критичним (табличним) при $k=n-2$ степенях свободи. Також будується довірчий інтервал для коефіцієнта регресії $b_0 \pm t_{кр} \cdot S_{b_0}$. Якщо нуль потрапляє в межі довірчого інтервалу (нижня межа від'ємна, а верхня – додатна), то відповідний параметр вважається статистично незначущим, оскільки отримані дані не дозволяють відкинути припущення про його рівність нулю.

Зауваження 7.8. У загальному вигляді для вибірки та таблиці при знаходженні $t_{факт}$ для коефіцієнтів регресії b_0 і b_1 використовують формули:

$$t_{факт} = \frac{b_1}{m_{b_1}}, \tag{7.64}$$

де $m_{b_1} = \sqrt{\dfrac{S_\varepsilon^2}{n \cdot \sigma_x^2}}$, $S_\varepsilon^2 = \dfrac{n \cdot \sigma_y^2 \cdot \left(1 - R^2\right)}{n-2}$ – залишкова дисперсія;

$$t_{факт} = \frac{b_0}{m_{b_0}}, \tag{7.65}$$

де $m_{b_0} = \sqrt{\dfrac{S_\varepsilon^2 \cdot \sum n_x x^2}{n \cdot \sum n_x \left(x - \bar{x}\right)^2}}$ або $m_{b_0} = \sqrt{\dfrac{S_\varepsilon^2 \cdot \overline{x^2}}{n \cdot \sigma_x^2}}$.

Справжнє значення b_0 лежить у діапазоні $b_0 \pm t_{\text{кр}} \cdot m_{b_0}$, а довірчий інтервал коефіцієнта регресії b_1: $b_1 \pm t_{\text{кр}} \cdot m_{b_1}$.

Прогнозування на основі лінійної регресії

Якщо за вибіркою обсягу n знайдено рівняння (7.39) лінійної регресії, то за допомогою його можна прогнозувати значення результату $Y_{\text{прог}}$ за певного прогнозного значення фактору $x_{\text{прог}}$. Значення $x_{\text{прог}}$ підставляють у рівняння (7.39).

Оскільки точне рівняння регресії невідоме, то точний прогноз робити не можемо. Можна тільки стверджувати, що прогнозне значення результату $Y_{\text{прог}}$ при даному $x_{\text{прог}}$ із ймовірністю γ потрапить у довірчий інтервал $Y_{\text{прог}}$. Імовірність γ називається рівнем надійності.

Помилка прогнозу визначається за формулою:

$$t_{\text{прог}} = S \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_{\text{прог}} - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}\ , \qquad (7.66)$$

де S визначається за формулою (7.60).

Формула для інтервалу прогнозу має вигляд:

$$Y_{\text{прог}} \pm t_{\text{кр}} \cdot\ t_{\text{прог}}. \qquad (7.67)$$

Оцінка точності регресійних моделей

Для оцінки точності регресійної моделі доцільно використовувати середню відносну помилку апроксимації.

Означення 7.6. *Середня помилка апроксимації* – це середнє відхилення розрахункових значень від фактичних:

$$\bar{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - Y_i}{y_i} \right| \cdot 100\% \,. \qquad (7.68)$$

Чим менше розсіювання емпіричних точок навколо теоретичної лінії регресії, тим менша середня помилка апроксимації.

В економетриці прийняті такі межі для $\overline{A}$: $\overline{A}$<10%, то модель вважається високоточною і придатною для практичного прогнозування; $\overline{A}$ є[10%;20%) – точність хороша; $\overline{A}$ є[20%;50%) – точність задовільна; $\overline{A}$ >50% – точність незадовільна.

В інженерних розрахунках прийняті такі межі $\overline{A}$: $\overline{A}$ є[1%;3%) – модель ідеально визначає технічний процес; $\overline{A}$ є[3%;7%) – висока точність (стандарт для більшості інженерних розрахунків та проектних рішень); $\overline{A}$ є[7%;15%) – прийнятна точність (допустима для попередніх ескізних розрахунків або складних систем з великою кількістю шумів); $\overline{A}$ >15% – низька точність (в інженерній таку модель зазвичай бракують та шукають невраховані фактори чи помилки у вимірах).

Коефіцієнт еластичності

У дослідженнях широке застосування знаходить *коефіцієнт еластичності*.

Коефіцієнт еластичності E показує, на скільки відсотків в середньому зміниться результативна ознака y при зміні фактору x на 1% від свого номінального значення. Для лінійної регресії $y=b_0+b_1 \cdot x$ коефіцієнт еластичності дорівнює

$$E = b_1 \cdot \frac{x}{y}, \tag{7.69}$$

Приклад 70. Для 8 співробітників підприємства дослідили вплив інвестицій в автоматизацію (X, млн грн.) на виручку від одного співробітника (Y, млн грн.). За отриманими даними склали таблицю:

i	1	2	3	4	5	6	7	8
X	2,2	3,0	3,5	4,0	4,8	5,4	6,0	9,0
Y	6,7	7,2	7,3	9,1	9,8	10,7	12,1	12,4

Треба знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості α=0,05. Побудувати рівняння регресії для прогнозування впливу інвестицій в

автоматизацію (X) на виручку від одного співробітника (Y) та оцінити якість моделі.

Розв'язання. За умовою дані несгруповані. Вибірковий коефіцієнт лінійної кореляції Пірсона знайдемо за формулою (7.15). Розрахуємо значення X^2, Y^2 , $X \cdot Y$ та їх сум.

Для зручності подальших обчислень складемо таблицю 7.15, в яку внесемо знайдені значення.

Таблиця 7.15

i	x_i	y_i	$x_i{}^2$	$y_i{}^2$	$x_i \cdot y_i$
1	2,2	6,7	4,84	44,89	14,74
2	3,0	7,2	9,00	51,84	21,60
3	3,5	7,3	12,25	53,29	25,55
4	4,0	9,1	16,00	82,81	36,40
5	4,8	9,8	23,04	96,04	47,04
6	5,4	10,7	29,16	114,49	57,78
7	6,0	12,1	36,00	146,41	72,60
8	9,0	12,4	81,00	153,76	111,60
Σ	**37,9**	**75,3**	**211,29**	**743,53**	**387,31**

Знаходимо середні значення $\bar{x}$, $\bar{y}$ за формулами (7.8):

$$\bar{x}=37{,}9/8=4{,}7375; \qquad \bar{y}=75{,}3/8=9{,}4125.$$

Для формули (7.15) розрахуємо значення чисельника и знаменника:

чисельник $8\cdot 387{,}31-(37{,}9\cdot 75{,}3)=3098{,}48-2853{,}87=$**244,61**;

знаменник (вирази під коренями)

$$8\cdot 211{,}29-(37{,}9)^2=1690{,}32-1436{,}41=\mathbf{253{,}91};$$

$$8\cdot 743{,}53-(75{,}3)^2=5948{,}24-5670{,}09=\mathbf{278{,}15}.$$

Тоді

$$r_{xy}=\frac{244{,}61}{\sqrt{253{,}91}\cdot\sqrt{278{,}15}}\approx\frac{244{,}61}{265{,}7538}\approx 0{,}9204.$$

Згідно зі шкалою Чеддока, значення 0,9204 говорить про дуже високу (сильну) пряму лінійну залежність.

Зростання інвестицій в автоматизацію на підприємстві веде до істотного збільшення виручки, що припадає на одного працівника.

Для перевірки гіпотези про значущість коефіцієнта кореляції r_{xy} розглянемо дві гіпотези:

Нульова гіпотеза H_0: "Зв'язку між інвестиціями в автоматизацію і виручкою від одного співробітника немає, результат випадковий".

Альтернативна гіпотеза H_1: "Зв'язок статистично значущий".

Обчислюємо фактичне значення t-критерію за формулою (7.23):

$$t_{\text{факт}} = \frac{0{,}9204 \cdot \sqrt{8-2}}{\sqrt{1-(0{,}9204)^2}} \approx \frac{0{,}9204 \cdot \sqrt{6}}{0{,}391} \approx 5{,}766.$$

За таблицею А.4 (див. Додаток А) для степенів свободи $k = n-2=6$ та заданим рівнем значущості $\alpha=0{,}05$ знаходимо критичне значення $t_{\text{кр}}=2{,}447$.

Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (5,766>2,447), то відхиляємо гіпотезу H_0.

З ймовірністю 95% стверджуємо, що зв'язок між інвестиціями в автоматизацію і виручкою від одного співробітника статистично значущий.

Оскільки маємо пряму лінійну залежність, то рівняння регресії згідно формули (7.39) має вигляд $Y=b_0+b_1 \cdot x$. Коефіцієнти лінійної регресії b_0 і b_1 знаходимо методом найменших квадратів (МНК).

Система рівнянь МНК за формулою (7.41) має вигляд (підставимо суми з таблиці):

$$\begin{cases} 211{,}29 \cdot b_1 + 37{,}9 \cdot b_0 = 387{,}31, \\ 37{,}9 \cdot b_1 + 8 \cdot b_0 = 75{,}3. \end{cases}$$

Розв'язок системи: $b_1 \approx 0{,}963$ (це означає, що кожен вкладений в автоматизацію мільйон приносить +0,963 млн. виручки); $b_0 \approx 4{,}849$ (базова виручка без урахування інвестицій в автоматизацію).

Рівняння регресії має вигляд:

$$Y=4{,}849+0{,}963 \cdot x.$$

Отримане рівняння регресії можна застосовувати для прогнозу про виручку підприємства при вкладенні інвестицій в автоматизацію, наприклад, 5 млн грн. Отримаємо розрахунок:

$$Y=4{,}849+0{,}963 \cdot 5=4{,}849+4{,}815=\mathbf{9{,}664} \text{ млн грн.}$$

Для оцінки якості моделі використовуємо коефіцієнт детермінації R^2. Він дорівнює квадрату коефіцієнта кореляції Пірсона r_{xy}^2, тобто $R^2=r_{xy}^2=(0{,}9204)^2 \approx 0{,}85$ (або 85%). Він показує, що 85% варіації виручки пояснюється саме рівнем автоматизації.

Відповідь: $r_{xy} \approx 0{,}9204$. Зв'язок між інвестиціями в автоматизацію на підприємстві веде до істотного збільшення виручки. Рівняння регресії

Y=4,849+0,963·x, коефіцієнт детермінації R^2≈0,85. Модель на 85% пояснює зміну виручки фактором інвестування в автоматизацію.

Приклад 71. За даними проведеного опитування 8 сімей відомі зв'язки витрат на продукти харчування Y: 0,8 1,1 1,7 2,1 2,4 2,7 3,1 3,5 (у тис. грн.) та рівнем доходів сім'ї X: 1,0 3,1 5,2 7,1 9,4 11,3 14,1 18,2. (у тис. грн.). Треба знайти рівняння парної регресії. Зробити оцінку якості рівняння регресії в цілому та оцінку статистичної значущості коефіцієнтів регресії та кореляції за t-критерієм Стьюдента при рівні значущості α=0,05. Скласти довірчий інтервал прогнозу для середнього доходу. Побудувати графік прямої лінії регресії.

Розв'язання. Рівняння регресії знайдемо за формулою (7.39). Коефіцієнти лінійної регресії b_0 і b_1 знаходимо за МНК. Спочатку знайдемо необхідні суми для системи нормальних рівнянь (7.41).

Розрахуємо значення X^2, Y^2 , $X{\cdot}Y$ та їх суми. Для зручності подальших обчислень складемо таблицю 7.16, в яку внесемо знайдені значення.

Таблиця 7.16

i	x_i	y_i	$x_i{}^2$	$y_i{}^2$	$x_i{\cdot}y_i$
1	1,0	0,8	1,0	0,64	0,8
2	3,1	1,1	9,61	1,21	3,41
3	5,2	1,7	27,04	2,89	8,84
4	7,1	2,1	50,41	4,41	14,91
5	9,4	2,4	88,36	5,76	22,56
6	11,3	2,7	127,69	7,29	30,51
7	14,1	3,1	198,81	9,61	43,71
8	18,2	3,5	331,24	12,25	63,7
Σ	**69,4**	**17,4**	**834,16**	**44.06**	**188,44**
	$\overline{x}$ ≈8,675	$\overline{y}$ ≈2,175			

Розрахуємо коефіцієнт b_1 за формулою (7.43):

$$b_1 = \frac{8\cdot 188{,}44 - 69{,}4\cdot 17{,}4}{8\cdot 834{,}16 - (69{,}4)^2} = \frac{299{,}96}{1856{,}92} \approx 0{,}162.$$

Розрахуємо коефіцієнт b_0 за формулою (7.44):

b_0=17,4/8−0,162·69,4/8=2,175−0,162·8,675≈0,77.

Рівняння регресії має вигляд: Y=0,77+0,162·x.

Додатково складемо наступну таблицю 7.17 з величинами, потрібними для подальших розрахунків.

Таблиця 7.17

y_i	0,8	1,1	1,7	2,1	2,4	2,7	3,1	3,5
x_i	1,0	3,1	5,2	7,1	9,4	11,3	14,1	18,2
$x_i-\bar{x}$	–7,675	–5,575	–3,475	–1,575	0,725	2,625	5,425	9,525
Y_i	0,932	1,272	1,612	1,92	2,292	2,60	3,054	3,718
y_i-Y_i	−0,132	−0,172	0,088	0,18	0,108	0,1	0,046	−0,218
$(y_i-Y_i)^2$	0,0174	0,0296	0,0077	0,0324	0,0117	0,01	0,0021	0,0475

Побудуємо графік прямої лінії регресії (рис. 7.2).

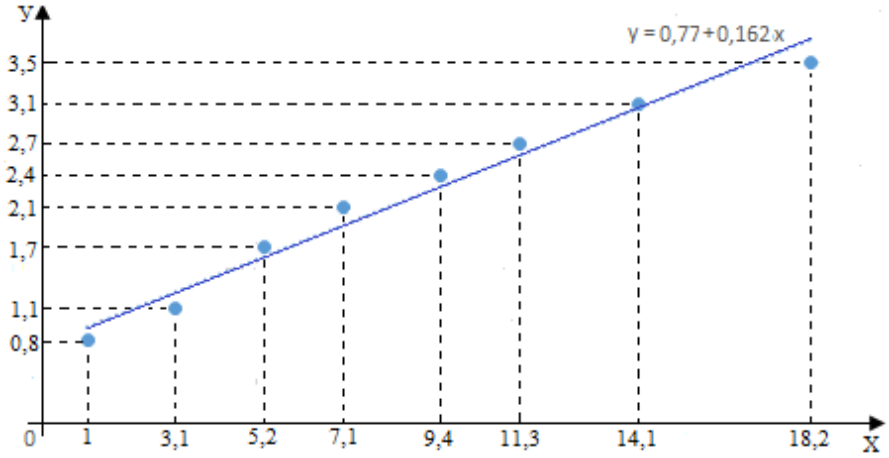


Рисунок 7.2

Для оцінки якості моделі використовуємо коефіцієнт детермінації R^2.

Вибірковий коефіцієнт лінійної кореляції Пірсона знайдемо за формулою (7.15).

$$r_{xy}=\frac{8\cdot 188{,}44-69{,}4\cdot 17{,}4}{\sqrt{8\cdot 834{,}16-(69{,}4)^2}\cdot\sqrt{8\cdot 44{,}06-(17{,}4)^2}}=\frac{299{,}96}{\sqrt{1856{,}92}\cdot\sqrt{49{,}72}}\approx 0{,}9872.$$

Зв'язок дуже високий і прямий.

$$R^2=r_{xy}{}^2=(0{,}9872)^2\approx 0{,}975.$$

Це означає, що у 97,5% сімей варіація витрат на харчування пояснюється рівнем доходу. Модель дуже якісна.

Перевіримо статистичну значущість коефіцієнта кореляції r_{xy}.

Обчислюємо фактичне значення t-критерію за формулою (7.23):

$$t_{\text{факт}} = \frac{0{,}9872 \cdot \sqrt{8-2}}{\sqrt{1-(0{,}9872)^2}} \approx \frac{0{,}9872 \cdot \sqrt{6}}{0{,}1595} \approx 15{,}16 .$$

За таблицею А.4 (див. Додаток А) для степенів свободи $k = n-2=6$ та заданим рівнем значущості $\alpha=0{,}05$ знаходимо критичне значення $t_{\text{кр}}=2{,}447$.

Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (15,16>2,447), то коефіцієнт кореляції статистично значущий.

Для повної оцінки якості рівняння регресії загалом необхідно використовувати F-критерій Фішера. Оскільки модель регресії парна, то фактичне значення $F_{\text{факт}}$ знаходимо за формулою (7.59): $F_{\text{факт}} \approx 234{,}0$.

Критичне значення F-критерію Фішера $F_{\text{кр}}$ знаходимо за таблицею А.11 (див. Додаток А) для степенів свободи: k_1 (горизонталь) $k_1=1$ для моделі парної регресії, k_2 (вертикаль) $k_2=n-2=6$ для моделі парної регресії та рівню значущості $\alpha=0{,}05$. $F_{\text{кр}}=5{,}99$.

Оскільки $F_{\text{факт}} > F_{\text{кр}}$ (234,0>5,99), рівняння регресії визнається статистично значущим в цілому. Імовірність того, що така залежність отримана випадково, дуже мала.

Крім F-критерію, якість моделі характеризують: середня помилка апроксимації A та стандартна помилка S регресії.

Середню помилку апроксимації $\overline{A}$ знаходимо за формулою (7.68). Знайдемо суми:

$$\sum_{i=1}^{8} \left| \frac{y_i - Y_i}{y_i} \right| = 0{,}165+0{,}156+0{,}052+0{,}086+0{,}045+0{,}037+0{,}015+0{,}062=0{,}618.$$

Тоді $\overline{A} = 0{,}618/8 \cdot 100\% = 7{,}7\%$.

Оскільки значення $\overline{A}$ менше 10%, модель вважається високоточною і може застосовуватись у практичних розрахунках.

Стандартну помилку S регресії знаходимо за формулою

$$S = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - Y_i)^2} .$$

Знайдемо $\frac{1}{6}\sum_{i=1}^{8}(y_i - Y_i)^2$ =(0,0174+0,0296+0,0077+0,0324+0,0117+0,01+

+0,0021+0,0475)/6=0,1584/6=0,0264.

Тоді $S = \sqrt{0,0264} \approx 0,162$ (тис. грн.).

Це середнє відхилення реальних витрат від розрахункових.

Зробимо оцінку якості коефіцієнтів регресії b_0 і b_1.

1) Розрахуємо стандартну помилку коефіцієнта b_0 за формулою (7.61). Знайдемо $\sum_{i=1}^{8}(x_i - \bar{x})^2$ =232,115 і підставимо в формулу

$$S_{b_0} = 0,162 \cdot \sqrt{\frac{834,16}{8 \cdot 232,115}} \approx 0,162 \cdot 0,67 \approx 0,1085.$$

Знаходимо фактичне значення *t*- статистики для b_0 за формулою

$$t_{\text{факт}} = \frac{b_0}{S_{b_0}} = \frac{0,77}{0,1085} \approx 7,097.$$

Порівняємо з табличним значенням $t_{\text{кр}}$=2,447 (знайдено вище). Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (7,097>2,447), то коефіцієнт b_0 статистично значущий. Це означає, що навіть за мінімального доходу існують певні базові витрати на харчування, які модель фіксує коректно.

2) Розрахуємо стандартну помилку коефіцієнта b_1 за формулою (7.61).

$$S_{b_1} = 0,162 \cdot \sqrt{\frac{1}{232,115}} \approx 0,162 \cdot 0,0656 \approx 0,0106.$$

Знаходимо фактичне значення *t*- статистики для b_1 за формулою (7.62)

$$t_{\text{факт}} = 0,162/0,0106 \approx 15,283.$$

Порівняємо з табличним значенням $t_{\text{кр}}$=2,447. Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (15,283>2,447), то коефіцієнт b_1 статистично значущий. Це означає, що дохід впливає на витрати.

Складемо довірчий інтервал прогнозу для середнього доходу $\bar{x}$ =8,675. Стандартна помилка прогнозу *S* залежить від того, наскільки далеко *x* знаходиться від середнього $\bar{x}$. Довірчий інтервал прогнозу знаходимо за формулою (7.67). Наприклад, візьмемо значення доходу

$x_{\text{прог}}$=20. Тоді $Y_{\text{прог}}$=0,77+0,162·20=4,01. Помилка прогнозу визначається за формулою (7.66):

$$t_{\text{прог}} = 0{,}162 \cdot \sqrt{1+\frac{1}{8}+\frac{(20-8{,}675)^2}{232{,}115}} \approx 0{,}162 \cdot \sqrt{1{,}6776} \approx 0{,}2098.$$

Тоді для інтервалу прогнозу матимемо

$y_{\min}$=4,01–2,447·0,2098=3,4966; $y_{\max}$=4,01+2,447·0,2098=4,5234.

З ймовірністю 95% можна стверджувати, що при рівні доходу сім'ї 20 тис. грн., витрати на харчування становитимуть приблизно від 3,50 до 4,52 тис. грн.

Відповідь: $r_{xy}\approx$0,9872, коефіцієнт кореляції статистично значущий ($t_{\text{факт}}$=15,16); $F_{\text{факт}}$=234 − рівняння регресії (Y=0,77+0,162·x) статистично значуще в цілому; коефіцієнт b_0=0,77 статистично значущий ($t_{\text{факт}}$=7,097); коефіцієнт b_1=0,162 статистично значущий ($t_{\text{факт}}$=15,283); інтервал прогнозу для середнього доходу [3,4966;4,5234].

Приклад 72. Вивчають залежність між балом за тест з математики (X) та балом за тест з фізики (Y) у групи з 50 студентів (n=50). Отримані дані подані у вигляді кореляційної таблиці. Треба знайти рівняння парної регресії. Зробити оцінку статистичної значущості коефіцієнтів регресії за t-критерієм Стьюдента при рівні значущості α=0,05.

Розв'язання. Додатково розраховуємо $n_x \cdot x$, $n_x \cdot x^2$, $n_y \cdot y$, $n_y \cdot y^2$ і вносимо їх значення у таблицю 7.18.

Таблиця 7.18

X / Y	**1(F)**	**2(FX)**	**3(D-E)**	**4(B-C)**	**5(A)**	n_y	$n_y \cdot y$	$n_y \cdot y^2$
1(F)	5	2	0	0	0	**7**	7	7
2(FX)	1	6	3	0	0	**10**	20	40
3(D-E)	0	2	10	4	0	**16**	48	144
4(B-C)	0	0	2	8	2	**12**	48	192
5(A)	0	0	0	1	4	**5**	25	125
n_x	**6**	**10**	**15**	**13**	**6**	n**=50**	**148**	**508**
$n_x \cdot x$	6	20	45	52	30	**153**		
$n_x \cdot x^2$	6	40	135	208	150	**539**		

Розраховуємо для X і Y середні за формулою (7.18):

$$\bar{x}=153/50=3{,}06;\ \bar{y}=148/50=2{,}96$$

та дисперсії за формулами $\sigma_x^2=\frac{1}{n}\cdot\overline{x^2}-\left(\bar{x}\right)^2$ і $\sigma_y^2=\frac{1}{n}\cdot\overline{y^2}-\left(\bar{y}\right)^2$:

$\overline{x^2}=539/50=10{,}78;\quad \sigma_x^2=10{,}78-(3{,}06)^2=10{,}78-9{,}36=1{,}42;$

$\overline{y^2}=508/50=10{,}16;\quad \sigma_y^2=10{,}16-(2{,}92)^2=10{,}16-8{,}76=1{,}40.$

Розраховуємо суми $\sum\sum n_{ij}\cdot x_i\cdot y_j$. Беремо кожну клітинку, де частота n_{ij} більша за нуль і утворюємо добутки:

$(1\cdot1\cdot5)+(2\cdot1\cdot2)=5+4=9;\ (1\cdot2\cdot1)+(2\cdot2\cdot6)+(3\cdot2\cdot3)=2+24+18=44;$

$(2\cdot3\cdot2)+(3\cdot3\cdot10)+(4\cdot3\cdot4)=12+90+48=150;$

$(3\cdot4\cdot2)+(4\cdot4\cdot8)+(5\cdot4\cdot2)=24+128+40=192;$

$(4\cdot5\cdot1)+(5\cdot5\cdot4)=20+100=120.$

Разом $\sum\sum n_{ij}\cdot x_i\cdot y_j=9+44+150+192+120=515$. Тоді $\overline{xy}=515/50=10{,}3$.

$$\mathrm{cov}(X,Y)=10{,}3-(3{,}06\cdot2{,}96)=10{,}3-9{,}0576=1{,}2424.$$

Коефіцієнти регресії можна знайти за формулами:

$b_1=\mathrm{cov}(X,Y)/\sigma_x^2=1{,}2424/1{,}42\approx0{,}875;\ b_0=\bar{y}-b_1\cdot\bar{x}=2{,}96-0{,}875\cdot3{,}06\approx0{,}283.$

Рівняння регресії має вигляд

$$Y=0{,}283+0{,}875\cdot x.$$

Коефіцієнт кореляції знайдемо за формулою (7.12)

$$r_{xy}=\frac{1{,}2424}{\sqrt{1{,}42}\cdot\sqrt{1{,}40}}\approx\frac{1{,}2424}{1{,}41}\approx0{,}88.$$

Зв'язок між тестами сильний ($r_{xy}=0{,}88$), точність моделі висока, оскільки $R^2=r_{xy}^2=(0{,}88)^2\approx0{,}77$ (77% варіації балів з фізики пояснюється балом з математики).

Перевіримо статистичну значущість коефіцієнта кореляції r_{xy}.

Обчислюємо фактичне значення t-критерію за формулою (7.23):

$$t_{\text{факт}}=\frac{0{,}88\cdot\sqrt{50-2}}{\sqrt{1-(0{,}88)^2}}\approx\frac{0{,}88\cdot\sqrt{48}}{\sqrt{0{,}2256}}\approx12{,}84.$$

Критичне значення t-критерію Стьюдента $t_{\text{кр}}$ знаходимо за таблицею А.4 (див. Додаток А) для степенів свободи $k=n-2=50-2=48$ та заданим рівнем значущості $\alpha=0{,}05$: $t_{\text{кр}}\approx2{,}021$.

Оскільки $t_{\text{факт}}>t_{\text{кр}}$ (12,87>2,021), то зв'язок визнається статистично достовірним.

Зробимо оцінку якості коефіцієнта регресії b_1. Розглянемо гіпотези:

H_0: "Бал по предмету X не впливає на бал з предмета Y (зв'язку немає)"; H_1: "Існує лінійна залежність між успіхами в тесті X і Y".

Знайдемо залишкову дисперсію

$$S_\varepsilon^2 = \frac{n \cdot \sigma_y^2 \cdot \left(1 - R^2\right)}{\left(n-2\right)} = \frac{50 \cdot 1{,}40 \cdot \left(1 - 0{,}77\right)}{48} \approx 0{,}335 .$$

Стандартна помилка m_{b_1} коефіцієнту регресії b_1:

$$m_{b_1} = \sqrt{\frac{S_\varepsilon^2}{n \cdot \sigma_x^2}} = \sqrt{\frac{0{,}335}{50 \cdot 1{,}42}} \approx \sqrt{0{,}0047} \approx 0{,}0686 .$$

Знаходимо фактичне значення t- статистики для b_1 за формулою (7.64)

$$t_{\text{факт}} = 0{,}875 / 0{,}0686 \approx 12{,}755.$$

Критичне значення t-критерію Стьюдента $t_{\text{кр}} \approx 2{,}021$ (знайдено вище). Оскільки $t_{\text{факт}} > t_{\text{кр}}$ (12,755>2,021), то зв'язок між предметами статистично значущий.

Зробимо оцінку якості коефіцієнта регресії b_0. Розглянемо гіпотези:

Стандартна помилка m_{b_0} коефіцієнту регресії b_0:

$$m_{b_0} = \sqrt{\frac{S_\varepsilon^2 \cdot \overline{x^2}}{n \cdot \sigma_x^2}} = \sqrt{\frac{0{,}335 \cdot 10{,}78}{50 \cdot 1{,}42}} \approx \sqrt{0{,}0509} \approx 0{,}2256 .$$

Знаходимо фактичне значення t- статистики для b_0 за формулою (7.65)

$$t_{\text{факт}} = 0{,}283 / 0{,}2256 \approx 1{,}254.$$

Оскільки $t_{\text{факт}} < t_{\text{кр}}$ (1,254<2,021), то коефіцієнт b_0 статистично не значущий. Це означає, що пряма проходить досить близько до початку координат, і не можна впевнено сказати, що вона перетинає вісь Y саме в точці 0,283, а не в 0.

Відповідь: $Y = 0{,}283 + 0{,}875 \cdot x$, коефіцієнт $b_1 = 0{,}875$ статистично значущий, коефіцієнт $b_0 = 0{,}283$ статистично не значущий.

7.2.2 Нелінійна регресія

Багато процесів по суті не є лінійними. Співвідношення між змінними X та Y не завжди можна адекватно описати за допомогою лінійних моделей, оскільки це може призвести до значного зміщення оцінок. Типовими прикладами нелінійних залежностей є зв’язок між

обсягом випуску продукції та виробничими факторами (виробничі функції), а також залежність попиту на товари від рівня цін чи доходів населення (функції еластичності) тощо. У таких випадках використовують нелінійну регресію, для якої у загальному випадку рівняння можна подати у вигляді $y=f(x)$.

Розрізняють два класи нелінійних моделей:

1) Регресії, нелінійні за пояснюючими змінними, але лінійні за оцінюваними параметрами. Прикладами таких моделей є:

- поліноми різних степенів $y=b_0+b_1\cdot x+b_2\cdot x^2$, $Y=b_0+b_1\cdot x+b_2\cdot x^2+b_3\cdot x^3$ та інші;
- рівностороння гіпербола $y=b_0+b_1/x$;
- напівлогарифмічна функція $y=b_0+b_1\cdot \ln x$.

2) Регресії, нелінійні за оцінюваними параметрами. До них належать такі функції, як:

- степенева $y = b_0 \cdot x^{b_1}$;
- показникова $y = b_0 \cdot b_1^{\,x}$;
- експоненціальна $y = \mathrm{e}^{b_0+b_1 x}$.

Для оцінювання параметрів нелінійних моделей у статистиці застосовують два основні підходи.

Перший підхід. *Метод лінеаризації* (аналітичне перетворення): за допомогою логарифмування або заміни змінних нелінійна модель зводиться до лінійного вигляду. Це дозволяє використовувати класичний метод найменших квадратів (МНК) для знаходження параметрів..

Другий підхід. *Прямі методи нелінійного оцінювання*: якщо модель неможливо лінеаризувати (наприклад, вона є істотно нелінійною за параметрами), використовують ітераційні чисельні методи (наприклад, метод Гаусса-Ньютона або метод градієнтного спуску) для безпосередньої мінімізації суми квадратів відхилень.

Метод найменших квадратів (МНК) застосовується не тільки до лінійних моделей, а й до більш загальних випадків, включаючи нелінійні моделі та моделі з кількома змінними. Принцип залишається тим самим: мінімізація суми квадратів залишків.

Розглянемо деякі парні нелінійні рівняння регресій, що відносяться до першого класу, які можна звести до парних лінійних шляхом відповідних перетворень змінних та параметрів вихідних рівнянь регресії

а) Поліном другого степеня $y=b_0+b_1\cdot x+b_2\cdot x^2$ зводиться до лінійного вигляду шляхом заміни: $x=x_1$, $x^2=x_2$. Він зазвичай застосовується в випадках, коли для певного інтервалу значень фактору змінюється характер зв'язку розглядуваних ознак: прямий зв'язок змінюється на зворотній, чи зворотній – на прямий. В результаті отримаємо двофакторне рівняння $y=b_0+b_1\cdot x_1+b_2\cdot x_2$ (знаходження коефіцієнтів цього рівняння розглянемо нижче у розділі «Множинна регресія»).

Для рівняння регресії $y=b_0+b_1\cdot x+b_2\cdot x^2$ можна застосовувати безпосередньо МНК, оскільки воно є лінійним за параметрами b_0, b_1, b_2 незважаючи на нелінійну залежність від змінної x. Розглянемо для нього МНК.

Маємо квадратну модель регресії має вигляд:

$$y=b_0+b_1\cdot x+b_2\cdot x^2, \tag{7.70}$$

де b_0, b_1, b_2 – невідомі коефіцієнти регресії.

За методом найменших квадратів необхідно мінімізувати суму квадратів залишків $Q(b_0,b_1,b_2)$:

$$Q(b_0,b_1,b_2)=\sum_{i=1}^{n}\left(y_i-\left(b_0+b_1x_i+b_2{x_i}^2\right)\right)^2\rightarrow \min\ . \tag{7.71}$$

Для знаходження b_0, b_1, b_2 знаходимо частинні похідні функції $Q(b_0,b_1,b_2)$ по кожному з цих параметрів, прирівнюємо їх до нуля, приводимо подібні і розв'язуємо систему з трьох лінійних рівнянь:

$$\begin{cases} b_0\cdot n+b_1\sum\limits_{1=1}^{n}x_i+b_2\sum\limits_{1=1}^{n}x_i^2 &=\sum\limits_{1=1}^{n}y_i,\\ b_0\sum\limits_{1=1}^{n}x_i+b_1\sum\limits_{1=1}^{n}x_i^2+b_2\sum\limits_{1=1}^{n}x_i^3=\sum\limits_{1=1}^{n}\left(y_i\cdot x_i\right),\\ b_0\sum\limits_{1=1}^{n}x_i^2+b_1\sum\limits_{1=1}^{n}x_i^3+b_2\sum\limits_{1=1}^{n}x_i^4=\sum\limits_{1=1}^{n}\left(y_i\cdot x_i^2\right). \end{cases} \tag{7.72}$$

Розв'язуючи систему рівнянь (7.72) отримаємо оцінки b_0^*, b_1^*, b_2^* параметрів b_0, b_1, b_2.

б) Гіпербола $y=b_0+b_1/x$ зводиться до лінійного вигляду шляхом заміни: $z=1/x$. Тоді матимемо $y=b_0+b_1 \cdot z$, тобто між y і z існує лінійний зв'язок. Система лінійних рівнянь при застосуванні МНК буде мати вигляд:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum\limits_{i=1}^{n} \dfrac{1}{x} = \sum\limits_{i=1}^{n} y_i, \\ b_0 \cdot \sum\limits_{i=1}^{n} \dfrac{1}{x} + b_1 \cdot \sum\limits_{i=1}^{n} \dfrac{1}{x_i^2} = \sum\limits_{i=1}^{n} \dfrac{y_i}{x_i}. \end{cases} \tag{7.73}$$

в) Напівлогарифмічна функція $y=b_0+b_1 \cdot \ln x$ зводиться до лінійного вигляду шляхом заміни: $z=\ln x$;

Також для нелінійної регресії, що відноситься до другого класу, можемо застосувати лінеаризацію моделі. Наприклад:

а) Степенева $y = b_0 \cdot x^{b_1}$. Логарифмуємо рівняння за основою e:

$$\ln y = \ln \left(b_0 \cdot x^{b_1} \right) = \ln b_0 + b_1 \ln x .$$

Зробимо заміну змінних: $X = \ln x$, $Y=\ln y$, $A_0=\ln b_0$. Маємо лінійну модель: $Y=A_0+b_1X$. Застосовуємо МНК для знаходження A_0 та b_1. Система лінійних рівнянь при застосуванні МНК буде мати вигляд:

$$\begin{cases} A_0 \cdot n + b_1 \cdot \sum\limits_{1=1}^{n} \ln x_i = \sum\limits_{1=1}^{n} \ln y_i, \\ A_0 \cdot \sum\limits_{1=1}^{n} \ln x_i + b_1 \cdot \sum\limits_{1=1}^{n} (\ln x_i)^2 = \sum\limits_{1=1}^{n} (\ln x_i \cdot \ln y_i). \end{cases} \tag{7.74}$$

Потім знаходимо $b_0 = \mathrm{e}^{A_0}$.

б) Показникова $y = b_0 \cdot b_1^{\,x}$. Логарифмуємо рівняння за основою e:

$$\ln y = \ln \left(b_0 \cdot b_1^{\,x} \right) = \ln b_0 + x \ln b_1 .$$

Зробимо заміну змінних: $Y=\ln y$, $A_1=\ln b_1$, $A_0=\ln b_0$.

Маємо лінійну модель: $Y=A_0+A_1x$. Застосовуємо МНК для знаходження A_0 та A_1. Система лінійних рівнянь при застосуванні МНК буде мати вигляд:

$$\begin{cases} A_0 \cdot n + A_1 \cdot \sum\limits_{1=1}^{n} x_i = \sum\limits_{1=1}^{n} \ln\, y_i, \\ A_0 \cdot \sum\limits_{1=1}^{n} x_i + A_1 \cdot \sum\limits_{1=1}^{n} x_i^2 = \sum\limits_{1=1}^{n} \left(x_i \cdot \ln\, y_i\right). \end{cases} \tag{7.75}$$

За допомогою зворотних перетворень знаходимо вихідні параметри $b_0=\mathrm{e}^{A_0}$, $b_1=\mathrm{e}^{A_1}$.

в) Експоненціальна модель $y=\mathrm{e}^{b_0+b_1x}$. Логарифмуємо рівняння за основою е:

$$\ln\, y=\ln\left(\mathrm{e}^{b_0+b_1x}\right)=b_0+b_1x\,.$$

Зробимо заміну змінних: $Y=\ln y$.

Маємо лінійну модель: $Y=b_0+b_1\cdot x$. Застосовуємо МНК для знаходження b_0 та b_1. Система лінійних рівнянь при застосуванні МНК буде мати вигляд:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum\limits_{1=1}^{n} x_i = \sum\limits_{1=1}^{n} \ln\, y_i, \\ b_0 \cdot \sum\limits_{1=1}^{n} x_i + b_1 \cdot \sum\limits_{1=1}^{n} x_i^2 = \sum\limits_{1=1}^{n} \left(x_i \cdot \ln\, y_i\right). \end{cases} \tag{7.76}$$

Якщо модель неможливо лінеарізувати, використовують ітераційні методи. Ці методи починаються з початкових значень параметрів і поступово коригують їх, щоб зменшити суму квадратів помилок. Прикладом такого методу є алгоритм Гауса-Ньютона чи метод Левенберга-Марквардта. Вони вимагають більше обчислювальних ресурсів та можуть не збігатися до єдиного розв'язку.

Означення 7.7. *Тісноту зв'язку між величинами* для нелінійної регресії оцінюють за величиною кореляційного відношення ρ_{xy} (*індекс кореляції*):

$$\rho_{xy}=\sqrt{1-\frac{\sum_{i=1}^{n}\left(y_i-Y_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}}=\sqrt{1-\frac{\sigma_{зал}^2}{\sigma_{заг}^2}}\,, \tag{7.77}$$

де $\sigma_{зал}^2$ – залишкова дисперсія, $\sigma_{заг}^2$ – загальна дисперсія результативної ознаки *Y*.

Інтервал зміни кореляційного відношення: $0\le\rho_{xy}\le1$. Чим ближче значення індексу кореляції до одиниці, тим тісніше зв'язок розглядуваних ознак, тим більш надійно рівняння регресії. При $\rho_{xy}=1$ зв'язок стає функціональним, тобто співвідношення $y=f(x)$ виконується для всіх спостережень.

Означення 7.8. *Індексом детермінації* називають квадрат індексу кореляції. Він характеризує частку дисперсії, що пояснюється регресією, у загальній дисперсії результативної ознаки *Y* і визначається формулою

$$R^2=\rho_{xy}^2. \tag{7.78}$$

Чим ближче коефіцієнт детермінації до 1, тим вища якість рівняння регресії. Індекс детермінації ρ_{xy}^2 можна порівнювати з коефіцієнтом детермінації r_{xy}^2 для обґрунтування можливості застосування лінійної функції. Чим більша кривизна лінії регресії, тим величина r_{xy}^2 менша за ρ_{xy}^2. А близькість цих показників вказує на те, що немає необхідності ускладнювати форму рівняння регресії і можна використовувати лінійну функцію.

Індекс детермінації використовується для перевірки значущості рівняння регресії за *F*-критерієм Фішера. Фактичне значення $F_{факт}$ для моделі визначається за формулою:

$$F_{факт}=\frac{\rho_{xy}^2}{\left(1-\rho_{xy}^2\right)}\cdot\frac{n-m-1}{m}\,, \tag{7.79}$$

де ρ_{xy}^2 – індекс детермінації, *n* – число спостережень *m* – число параметрів при змінній *X*.

Згідно з F-критерієм Фішера, висувається нульова гіпотеза H_0 про статистичну незначущість рівняння регресії.

Фактичне значення $F_{\text{факт}}$ порівнюють з критичним значенням критерію Фішера $F_{\text{кр}}$, яке знаходимо за таблицею А.11 (див. Додаток А) при рівні значущості α і числі степенів свободи $k_2=n-m-1$ (для залишкової суми квадратів) та $k_1=m$ (для факторної суми квадратів).

Якщо $F_{\text{факт}}>F_{\text{кр}}$, то визнається статистична значущість рівняння регресії у цілому.

Середню помилку апроксимації $\overline{A}$ визначаємо за формулою (7.68). Якщо помилка апроксимації менше 10%, то це свідчить про хороший підбір рівняння регресії до вихідних даних.

Якщо залежність між змінними x і y має вигляд $y=f(x)$, то коефіцієнт еластичності E обчислюється за формулою:

$$E = f'(x)\cdot\frac{x}{y}, \tag{7.80}$$

Коефіцієнт еластичності E показує, на скільки відсотків в середньому зміниться результативна ознака y при зміні фактору x на 1% від свого номінального значення.

Коефіцієнт еластичності E в загальному випадку залежить від величини x і є величиною змінної. Щоб виключити цю залежність, застосовується *середній коефіцієнт еластичності* $\overline{E}$:

$$\overline{E} = f'(\overline{x})\cdot\frac{\overline{x}}{y}, \tag{7.81}$$

який є величиною сталою.

Середній коефіцієнт еластичності $\overline{E}$ показує, на скільки відсотків у середньому за сукупністю значень фактору x зміниться результативна ознака y при зміні фактору x на 1%.

Приклад 73. За даними **прикладу 71** знайти: рівняння квадратної регресії, кореляційне відношення, індекс детермінації, середню помилку апроксимації та середній коефіцієнт еластичності та перевірити значущість моделі при рівні значущості α=0,05. Побудувати графік квадратної регресії.

Розв'язання. За умовою отримали дані опитування 8 сімей стосовно витрат на продукти харчування Y: 0,8 1,1 1,7 2,1 2,4 2,7 3,1 3,5 (у тис. грн.) та рівнем їх доходів X: 1,0 3,1 5,2 7,1 9,4 11,3 14,1 18,2. (у тис. грн.).

Рівняння поліноміальної (квадратичної) регресії знайдемо за формулою (7.70). Коефіцієнти квадратичної регресії b_0, b_1, b_2 знаходимо за МНК, розв'язавши систему (7.72), складену за даними прикладу. Для зручності подальших обчислень складемо таблицю 7.19.

Таблиця 7.19

i	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i \cdot y_i$	$x_i^2 \cdot y_i$
1	1,00	0,80	1,00	1,00	1,00	0,80	0,80
2	3,10	1,10	9,61	29,79	92,35	3,41	10,57
3	5,20	1,70	27,04	140,61	731,16	8,84	45,97
4	7,10	2,10	50,41	357,91	2541,17	14,91	105,86
5	9,40	2,40	88,36	830,58	7807,49	22,56	212,06
6	11,30	2,70	127,69	1442,90	16304,74	30,51	344,76
7	14,10	3,10	198,81	2803,22	39525,42	43,71	616,31
8	18,20	3,50	331,24	6028,57	109719,94	63,70	1159,34
Σ	**69,40**	**17,40**	**834,16**	**11634,58**	**176723,26**	**188,44**	**2495,68**
	$\overline{x} \approx 8{,}675$	$\overline{y} \approx 2{,}175$					

Підставимо суми до системи рівнянь (7.72):

$$\begin{cases} 8b_0 + 69{,}4b_1 + 834{,}16b_2 = 17{,}4 \\ 69{,}4b_0 + 834{,}16b_1 + 11634{,}58b_2 = 188{,}44 \\ 834{,}16b_0 + 11634{,}58b_1 + 176723{,}26b_2 = 2495{,}68 \end{cases}$$

Розв'язуємо систему (зазвичай методом Гауса або за формулами Крамера), отримуємо коефіцієнти:

$$b_0 \approx 0{,}501; \qquad b_1 \approx 0{,}248; \qquad b_2 \approx -0{,}0045.$$

Рівняння квадратичної регресії $Y = 0{,}501 + 0{,}248 \cdot x - 0{,}0045 \cdot x^2$.

Оскільки коефіцієнт при x^2 від'ємний ($-0{,}0045$), парабола спрямована вітками донизу. Це логічно: зі зростанням доходів частка витрат на їжу зазвичай починає сповільнюватись.

Оскільки для подальших розрахунків потрібні значення $Y(x_i)$, то складемо таблицю 7.20.

Таблиця 7.20

i	x_i	y_i	x_i^2	Y_i	$(y_i-\bar{y})^2$	$(y_i-Y_i)^2$	$(Y_i-\bar{y})^2$
1	1,00	0,80	1,00	0,7445	1,8906	0,0031	2,0463
2	3,10	1,10	9,61	1,2266	1,1556	0,0160	0,8995
3	5,20	1,70	27,04	1,6689	0,2256	0,0010	0,2561
4	7,10	2,10	50,41	2,0350	0,0056	0,0042	0,0196
5	9,40	2,40	88,36	2,4346	0,0506	0,0012	0,0674
6	11,30	2,70	127,69	2,7288	0,2756	0,0008	0,3067
7	14,10	3,10	198,81	3,1032	0,8556	0,0000	0,8615
8	18,20	3,50	331,24	3,5240	1,7556	0,0006	1,8199
Σ	**69,40**	**17,40**	**834,16**	**17,4655**	**6,2150**	**0,0269**	**6,2770**
	$\bar{y}\approx 2{,}175$						

Знайдемо кореляційне відношення за формулою (7.77):

$$\rho_{xy}=\sqrt{1-\frac{0{,}0269}{6{,}215}}=\sqrt{\frac{6{,}1881}{6{,}215}}\approx 0{,}9978\,.$$

Тоді індекс детермінації зв формулою (7.78) R^2=0,9957. Він показує, яка частка варіації y пояснюється моделлю.

Маємо, що зв'язок між доходом та витратами практично функціональний (дуже сильний).

Знайдемо середню помилку апроксимації, яка показує середнє відхилення розрахункових даних від фактичних у відсотках. За формулою (7,68)

$$\bar{A}\approx(6{,}94+11{,}51+1{,}83+3{,}10+1{,}44+1{,}07+0{,}10+0{,}69)/8=$$
$$=26{,}68/8\approx 3{,}34(\%).$$

Оскільки $\bar{A}<10\%$, то якість моделі оцінюється як «відмінна».

Знайдемо середній коефіцієнт еластичності $\bar{E}$ за формулою (7.81). Маємо Y'=0,248–2·0,0045·x, $\bar{x}$=8,675, $\bar{y}$=2,175. Тоді

$$\bar{E}=(0{,}248+2\cdot(-0{,}0045)\cdot 8{,}675)\cdot 8{,}675/2{,}175\approx 0{,}17\cdot 3{,}98\approx 0{,}677(\%).$$

Тобто при збільшенні доходу сім'ї на 1% витрати на продукти харчування в середньому зростають на 0,68%. Оскільки $\bar{E}<1$, це

підтверджує статус продуктів харчування як товарів першої необхідності (їхня частка в бюджеті падає при зростанні доходів).

Перевіримо значущість моделі. Сформулюємо гіпотези.

Нульова гіпотеза H_0: $R^2=0$. "Рівняння статистично незначне, фактор доходу X не впливає витрати Y".

Альтернативна гіпотеза H_1: $R^2\neq 0$. "Рівняння значимо, модель адекватно визначає залежність".

Знайдемо фактичне значення $F_{\text{факт}}$ для моделі за формулою (7.79), де $\rho_{xy}^2=R^2=0{,}9957$, $n=8$, $m=2$ (число факторів при x та x^2):

$$F_{\text{факт}}=\frac{0{,}9957}{(1-0{,}9957)}\cdot\frac{8-2-1}{2}=\frac{2{,}48925}{0{,}0043}\approx 578{,}9\,.$$

Критичне значення критерію Фішера $F_{\text{кр}}$ знаходимо за таблицею А.11 (див. Додаток А) при рівні значущості $\alpha=0{,}05$, степенів свободи: $k_1=2$ і $k_2=5$. $F_{\text{кр}}=5{,}79$.

Оскільки $F_{\text{факт}}>F_{\text{кр}}$ (578,9>4,79), то гіпотезу H_0 відхиляємо, модель вважається статистично значущою. Це означає, що її можна використовувати для прогнозування.

Побудуємо графік квадратичної регресії (рис. 7.3).

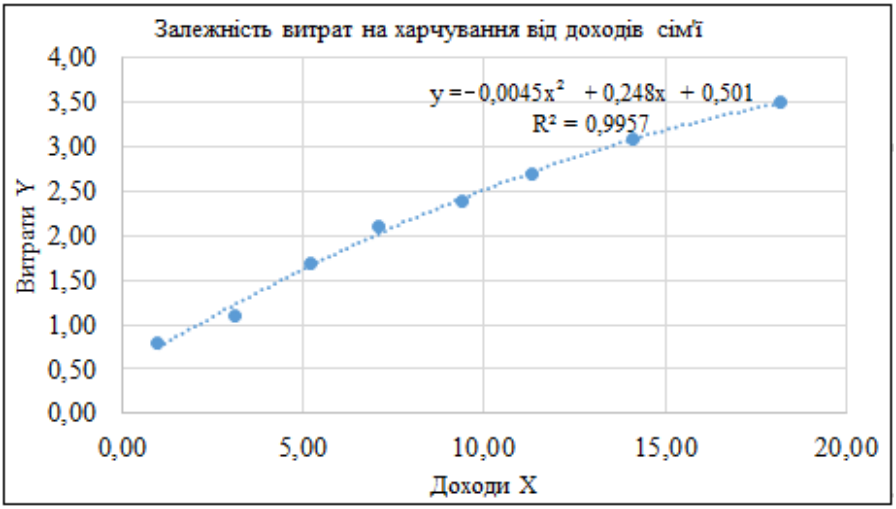


Рисунок 7.3

Відповідь: Y=0,501+0,248·x–0,0045·x^2, ρ_{xy}=0,9978, R^2=0,9957, $\overline{A}$ ≈3,34%, $\overline{E}$ ≈0,677%, $F_{факт}$=578,9 модель статистично значуща.

7.2.3 Множинна лінійна регресія

Однак у переважній більшості випадків доводиться мати справу з експериментальними даними, що стосуються впливу більш ніж одного фактору.

Означення 7.9. Прогнозування єдиної змінної Y виходячи з кількох змінних x_i (i=1,2,…,n) називається *множинною регресією*.

Головна мета множинної регресії полягає у створенні математичної моделі, яка враховує одночасний вплив декількох незалежних змінних (факторів) на результативний показник. Ключовою перевагою цього методу є можливість оцінити "чистий" внесок кожного окремого фактора за умови, що інші змінні залишаються незмінними, а також визначити загальну частку варіації результату, зумовлену всією сукупністю включених у модель факторів. У цьому випадку математична модель процесу представляється як рівняння регресії з кількома змінними величинами, тобто Y=f(x_1,x_2…,x_n). Рівняння множинної регресії зазвичай намагаються подати у формі лінійної залежності.

Множинна лінійна регресія дозволяє не тільки передбачати, але і розуміти відносний вплив різних факторів на результат, що цікавить. Це робить її незамінним інструментом для ухвалення обґрунтованих рішень.

Множинна лінійна регресія застосовується практично у всіх областях, де потрібно зрозуміти зв'язок між змінними чи зробити прогноз. Наприклад:

- *Виробництво*: Оцінка якості готової продукції в залежності від умов її виробництва; при вивченні функції витрат виробництва.
- *Економіка та фінанси*: Прогноз вартості нерухомості на основі площі, району, року будівництва та відстані до метро; прогнозування фондових індексів на основі макроекономічних показників; оцінка впливу інфляції, процентних ставок та безробіття на ВВП; моделювання кредитного ризику.

- *Маркетинг та бізнес-аналітика*: Визначення факторів, що впливають на лояльність клієнтів; прогнозування відтоку клієнтів; оптимізація рекламних бюджетів; аналіз ефективності різних каналів продажу; аналіз продажів продукту залежно від витрат на рекламу, знижки тощо.

- *Медицина та епідеміологія*: Вивчення факторів ризику захворювань (наприклад, вплив куріння, віку, спадковості на розвиток серцево-судинних захворювань); оцінка дозування ліків; прогнозування наслідків лікування.

- *Соціологія та психологія*: Аналіз факторів, що впливають на рівень освіти, доходів, моделювання вибіркової поведінки.

- *Міське планування та екологія*: Прогнозування трафіку роботи міського транспорту на основі кількості його одиниць, маршруту, часу доби; оцінка впливу забруднюючих речовин для здоров'я населення.

- *Спорт*: Аналіз чинників, які впливають на спортивні результати гравців чи команд.

Суть множинної лінійної регресії полягає у пошуку математичного рівняння у формі лінійної залежності, яка найкраще описує залежність однієї залежної змінної Y від кількох пояснюючих незалежних змінних $X_1,X_2,...,X_k$.

Модель множинної лінійної регресії можна представити у вигляді:

$$Y=b_0+b_1X_1+b_2X_2+\ldots+b_kX_k+\varepsilon \qquad (7.82)$$

де b_0 – вільний член (або зсув), це значення Y, коли всі X дорівнюють нулю; $b_1,b_2,\ldots,b_k$ – параметри (коефіцієнти регресії), які показують, як у середньому зміниться Y зі збільшенням відповідного X на одиницю (за умови, що інші незалежні змінні залишаються незмінними – це важлива відмінність від простої регресії); ε – випадкова помилка, яку модель неспроможна пояснити.

Щоб модель була адекватною, необхідно дотримуватись наступних умов:

1) *Лінійність*: Зв'язок між факторами та результатом має бути хоча б приблизно лінійним.

2) *Відсутність мультиколінеарності*: Незалежні змінні X не повинні бути сильно пов'язані одна з одною, щоб не виникало питання: який з факторів важливіший, оскільки вони будуть дублювати одна одну.

3) Нормальність залишків: Помилки моделі ε мають бути розподілені випадковим чином. Якщо в помилках видно систему, то модель щось упускає.

При аналізі рівняння множинної регресії (як у разі парної регресії) використовується таке поняття, як *помилка прогнозування* Δy, яка є різницю між розрахованим (теоретичним) значенням функції Y_i та її виміряним (дослідним) значенням y_i, тобто $\Delta y = Y_i - y_i$.

Найчастіше для знаходження параметрів лінійної моделі множинної регресії застосовують метод найменших квадратів (МНК). Він дозволяє отримати такі оцінки параметрів, щоб сума квадратів відхилень реальних даних від тих, що передбачила модель, була мінімальною:

$$Q(b_0, b_1, b_2, \ldots, b_k) = \sum_{i=1}^{n} \left(y_i - Y_i\right)^2 \rightarrow \min . \qquad (7.83)$$

Формулу (7.83) можна записати у вигляді

$$Q(b_0, b_1, b_2, \ldots, b_k) = \sum_{i=1}^{m} \left(y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}\right)\right)^2 \rightarrow \min . \qquad (7.84)$$

Пошук мінімуму функції $Q(b_0, b_1, b_2, \ldots, b_k)$ зводиться до обчислення частинних похідних по кожному коефіцієнту b_i і прирівнюванню їх до нуля. Матимемо систему із $k+1$ лінійного рівняння (їх називають нормальними рівняннями). Розв'язок цієї системи дає значення коефіцієнтів, які забезпечують найкращу апроксимацію даних.

Для того, щоб система рівнянь МНК мала розв'язок, необхідно, щоб кількість спостережень m була більшою за кількість оцінюваних параметрів $k+1$.

Систему лінійних рівнянь можна записати у матричному вигляді:

$$y = X \cdot b,$$

де X – матриця значень незалежних змінних; y – вектор значень залежної змінної; b – вектор шуканих коефіцієнтів регресії.

Матриця X має розмірність $(m \times k)$, де m – число спостережень (рядків), а k – число параметрів (стовпців, включаючи одиничний

стовпець для b_0). Оскільки даних зазвичай набагато більше, ніж факторів ($m>k$), то матриця X не є квадратною.

Щоб знайти розв'язок системи, множимо обидві частини рівняння зліва на транспоновану X^{T} матрицю. Отримаємо рівняння:

$$X^{\mathrm{T}}\cdot y=(X^{\mathrm{T}}\cdot X)\, b.$$

Розмірність добутку $(X^{\mathrm{T}}\cdot X) - (k\times k)$, оскільки розмірність $X^{\mathrm{T}} - (k\times m)$, а $X - (m\times k)$. Матриця $(X^{\mathrm{T}}\cdot X)$ квадратна, для неї існує обернена матриця $(X^{\mathrm{T}}\cdot X)^{-1}$. Помножимо обидві частини рівняння на обернену матрицю. Матимемо розв'язок системи:

$$b'=(X^{\mathrm{T}}\cdot X)^{-1}\, X^{\mathrm{T}}\cdot y. \qquad (7.85)$$

Для розрахунку рівняння множинної регресії з k факторами (аргументами) складається розширена матриця даних у вигляді таблиці 7.21:

Таблиця 7.21

Номер випробування	X_0	X_1	X_2	…	X_k	y
1	1	x_{11}	x_{12}	…	x_{1k}	y_1
2	1	x_{21}	x_{22}	…	x_{2k}	y_2
…		…	…	…	…	…
m	1	x_{m1}	x_{m2}	…	x_{mk}	y_m

Щоб знайти вільний член b_0, до таблиці додають фіктивний стовпець X_0, який завжди дорівнює 1.

Кожен рядок таблиці є результатом одного випробування. Спостереження розрізняються умовами їхнього проведення. Щоб не заплутатися в позначеннях, корисно розділяти дві величини: m – кількість точок (рядків) (чим більше m, тим надійніше будуть оцінені коефіцієнти); k – кількість факторів (стовпців X).

Знаючи рівняння лінійної регресії, необхідно перевірити адекватність отриманої моделі. У множинні регресії перевірка йде за трьома основними напрямками: перевірка значущості рівняння загалом (F-тест Фішера), перевірка значущості окремих коефіцієнтів (t- критерій Стьюдента), оцінка якості за коефіцієнтом детермінації R^2.

Коефіцієнт детермінації R^2

У множинній регресії R^2 показує частку загальної дисперсії залежної змінної Y, яку всі незалежні змінні X_i разом пояснюють. Чим ближче R^2 до 1, тим краще модель пояснює варіацію Y.

Коефіцієнт детермінації R^2 знаходиться за формулою:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - Y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{m}(Y_i - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}, \quad (7.86)$$

де $Y_i = b_0 + \sum_{j=1}^{k} b_j x_{ij}$, $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$.

Позначимо

$\sigma^2_{\text{зал}} = \sum_{i=1}^{m}(y_i - Y_i)^2$ – залишкова варіація. Вона характеризує величину суми квадратів відхилень фактичного значення результативної ознаки від її розрахункового значення.

$\sigma^2_{\text{заг}} = \sum_{i=1}^{m}(y_i - \bar{y})^2$ – загальна (повна) варіація. Вона характеризує величину суми квадратів відхилень фактичного значення результативної ознаки від загальної середньої.

$\sigma^2_{\text{регр}} = \sum_{i=1}^{m}(Y_i - \bar{y})^2$ – регресійна (пояснена) варіація. Вона характеризує величину суми квадратів відхилень розрахованого значення результативної ознаки від загальної середньої.

Вони пов'язані співвідношенням: $\sigma_{\text{заг}}^2 = \sigma_{\text{зал}}^2 + \sigma_{\text{регр}}^2$.

Якщо в модель додають абсолютно випадковий фактор, то R^2 може збільшитись або залишиться тим самим, але ніколи не зменшиться. Це створює ілюзію, що модель стає кращою, хоча насправді вона просто підганяється під випадковий шум. Для вирішення цієї проблеми використовують скоригований R^2. Сенс якого у тому, що якщо новий фактор не робить істотного вкладу в пояснення Y, то дріб збільшиться, і скоригований R^2 зменшиться. Формула скоригованого R^2 має вигляд:

$$\overline{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{m-1}{m-k-1}, \qquad (7.87)$$

де m – кількість спостережень, k – кількість факторів.

Для перевірки значущості (придатності) отриманого рівняння регресії (адекватності моделі) застосовують критерій Фішера, який перевіряє у скільки разів гірше порівняно з дослідом передбачає результат модель.

Для перевірки значущості рівняння регресії загалом (F-тест Фішера) сформулюємо гіпотези.

Нульова гіпотеза H_0: "Усі коефіцієнти $b_1,b_2,\ldots,b_k$ рівні нулю. Фактори не пояснюють варіацію залежної змінної. Модель статистично незначуща".

Альтернативна гіпотеза H_1: "$\exists b_j \neq 0 \ \left(j = \overline{1,k}\right)$, тобто хоча б один із коефіцієнтів регресії (при факторах) не дорівнює нулю. Модель статистично значуща".

Універсальна формула визначення фактичного значення $F_{\text{факт}}$ для моделі з k факторами (незалежними змінними) та m спостереженнями (обсяг вибірки) має вигляд:

$$F_{\text{факт}} = \frac{R^2}{\left(1 - R^2\right)} \cdot \frac{m-k-1}{k}, \qquad (7.88)$$

де R^2 – коефіцієнт детермінації.

Критичне значення критерію Фішера $F_{\text{кр}}$ знаходимо за таблицею А.11 (див. Додаток А) для степенів свободи: k_1 (горизонталь), для множинної моделі $k_1=k$, де k – кількість факторів, k_2 (вертикаль) – це різниця між загальним числом m спостережень і кількістю параметрів моделі, що оцінюються $k_2=m–k–1$ та рівню значущості α (зазвичай 0,05 або 0,01).

Якщо $F_{\text{факт}}>F_{\text{кр}}$ (взято з таблиці розподілу Фішера для заданого рівня значущості, наприклад, 0,05), то гіпотезу H_0 відхиляємо, модель вважається статистично значущою. Це означає, що її можна використовувати для прогнозування. Якщо $F_{\text{факт}}\leq F_{\text{кр}}$, то гіпотезу H_0 не відхиляємо, модель статистично незначуща.

Рішення про адекватність моделі може прийматись на основі коефіцієнта детермінації R^2. Для цього знайдене значення коефіцієнта

детермінації R^2 порівнюють з табличним (критичним) значенням $R_{кр}^2$, знайденим за таблицею А.13 (див. Додаток А) для заданого рівня значущості α (зазвичай 0,05 або 0,01).

Якщо $R^2 > R_{кр}^2$, то (наприклад, α=0,05) з ймовірністю 95% можна стверджувати, що аналізована регресія є значущою.

Перевірка на мультиколінеарність дозволяє зрозуміти, як саме фактори впливають на результат (розглянемо на прикладі двофакторної моделі лінійної регресії).

Після підтвердження значущості моделі в цілому, наступним етапом є оцінка ролі окремих факторів за допомогою t-критеріїв Стьюдента. Це дозволяє визначити статистичну значущість впливу кожної змінної X_j на результат Y у "чистому" вигляді. Такий аналіз дає відповідь на питання: чи є вплив конкретного фактора вагомим, якщо припустити, що всі інші змінні X_k ($j \neq k$) в моделі зафіксовані на незмінному рівні.

При застосуванні *t*-критерій Стьюдента (перевірка значущості окремих коефіцієнтів при факторах моделі) для кожного коефіцієнта b_i (при факторі X_j моделі) формулюють свою пару гіпотез H_0 і H_1:

нульова гіпотеза H_0: "Коефіцієнт b_j=0 − коефіцієнт статистично незначущий. Це означає, що фактор X_j ніяк не впливає на Y";

альтернативна гіпотеза H_1: "Коефіцієнт $b_j \neq 0$ − коефіцієнт статистично значущий. Фактор X_j реально впливає на результат, його зв'язок з Y підтверджено математично".

Незміщену оцінку залишкової дисперсії $S_{зал}^2$ визначаємо за формулою:

$$S_{зал}^2 = \sigma_{зал}^2/(m-k-1).$$

Стандартну помилку рівняння знаходиться за формулою:

$$S = \sqrt{S_{зал}^2}\ .$$

Стандартні помилки S_{b_j} *коефіцієнтів регресії* визначаються співвідношеннями:

$$S_{b_j} = S \cdot \sqrt{\left[\left(X^T \cdot X\right)^{-1}\right]_{jj}}\ , \qquad (7.89)$$

де $\left[\left(X^T \cdot X\right)^{-1}\right]_{jj}$ – діагональний елемент матриці $(X^{\mathrm{T}}\cdot X)^{-1}$. Це значення можна знайти як $\left[\left(X^T \cdot X\right)^{-1}\right]_{jj} = \dfrac{A_{jj}}{\det\left(X^T \cdot X\right)}$ (A_{jj} – алгебраїчне доповнення до елементу *jj* матриці $(X^{\mathrm{T}}\cdot X)$.

Зауваження 7.9. Для знаходження стандартних помилок S_{b_j} коефіцієнтів регресії можна використовувати

$$S_{b_j} = \frac{1}{\sqrt{m-1}} \cdot \frac{s_y}{s_{x_j}} \cdot \sqrt{\frac{1-R^2}{1-R_j^2}}, \tag{7.90}$$

де $s_y = \sqrt{\dfrac{S^2_{\text{заг}}}{m-1}}$, $s_{x_j} = \sqrt{\dfrac{\sum\limits_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2}{m-1}}$, $R_j^2 = \dfrac{\sum\limits_{i=1}^{m}\left(\tilde{x}_{ij} - \bar{x}_j\right)^2}{\sum\limits_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2}$ – коефіцієнт детермінації при регресії змінної x_j на всі інші пояснюючі змінні, $\tilde{x}_{ij}$ – значення фактору x_j, передбачені з урахуванням інших факторів. Рівняння (7.89) можна спростити і записати у вигляді:

$$S_{b_j} = \sqrt{\frac{S^2_{\text{зал}}}{\sum\limits_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2 \cdot \left(1 - R_j^2\right)}}. \tag{7.91}$$

Для кожного коефіцієнта b_j (включаючи вільний член b_0) розраховуємо особисте фактичне *t*-значення:

$$t_{b_j,\text{факт}} = \frac{b_j}{S_{b_j}}. \tag{7.92}$$

Тобто ділимо значення модуля коефіцієнта b_j на його помилку. Чим більше $t_{b_j,\text{факт}}$, тим надійніший коефіцієнт.

Критичне значення *t*-критерію Стьюдента $t_{\text{кр}}$ знаходимо за таблицею А.4 (див. Додаток А) для степенів свободи $m–k–1$ та заданим рівнем значущості α (зазвичай 0,05 або 0,01).

Якщо для коефіцієнта b_j виконується $|t_{\text{факт}}|>t_{\text{кр}}$, то гіпотеза H_0 відхиляється, коефіцієнт *bj* статистично значущий. Фактор X_j реально впливає на результат, його зв'язок з *Y* підтверджено математично.

Якщо для коефіцієнта b_j виконується $|t_{\text{факт}}|<t_{\text{кр}}$, то гіпотеза H_0 приймається, коефіцієнт b_j незначущий і фактор X_j ніяк не впливає на *Y*.

Довірчий інтервал коефіцієнта регресії b_j визначається формулою:

$$b_j \pm t_{\text{кр}} \cdot S_{b_j} . \tag{7.93}$$

Середня помилка апроксимації обчислюється за формулою

$$\overline{A} = \frac{1}{m} \cdot \sum_{i=1}^{m} \left| \frac{y_i - Y_i}{y_i} \right| \cdot 100\% . \tag{7.94}$$

Вона дає можливість перевірити, наскільки модель є зручною для практики. Якщо її значення менше 10%, то модель вважається відмінною.

Середні коефіцієнти еластичності $\overline{E}_j$ для лінійної залежності визначаємо для кожного коефіцієнта b_j регресії, окрім b_0, за формулою:

$$\overline{E}_j = b_j \cdot \frac{\overline{x}_j}{\overline{y}}, \tag{7.95}$$

Середні коефіцієнти еластичності $\overline{E}_j$ показують силу впливу конкретного фактору на результат. При зміні фактору x_j на 1% залежна змінна *Y* зміниться в середньому на $\overline{E}_j$ відсотків.

Розглянемо алгоритм побудови двофакторної моделі лінійної регресії.

Двофакторна модель лінійної регресії описує залежність однієї залежної змінної Y від двох пояснюючих незалежних змінних X_1 і X_2, які мають по m значень. Визначимо рівняння регресії наступного виду:

$$Y=b_0+b_1X_1+b_2X_2, \tag{7.96}$$

де b_0 – вільний член (значення Y, коли всі X дорівнюють нулю); b_1, b_2 – коефіцієнти регресії.

За даними складемо таблицю 7.22

Таблиця 7.22

Номер випробування	X_0	X_1	X_2	y
1	1	x_{11}	x_{12}	y_1
2	1	x_{21}	x_{22}	y_2
…		…	…	…
m	1	x_{m1}	x_{m2}	y_m

1) *Знаходимо коефіцієнти* b_j, застосовуючи МНК. Для подальших розрахунків складемо додаткову таблицю 7.23

Таблиця 7.23

№	X_1	X_2	y	X_1^2	X_2^2	$X_1 \cdot X_2$	$X_1 \cdot y$	$X_2 \cdot y$	y^2	$(X_1-\overline{X}_1)^2$	$(X_2-\overline{X}_2)^2$
1	x_{11}	x_{12}	y_1	…	…	…	…	…	…	…	…
2	x_{21}	x_{22}	y_2	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	y_m	…	…	…	…	…	…	…	…
Σ	…	…	…	…	…	…	…	…	…	…	…

Для визначення коефіцієнтів b_j двофакторного рівняння регресії необхідно розв'язати систему нормальних рівнянь:

$$\begin{cases} b_0 m + b_1 \sum X_1 + b_2 \sum X_2 = \sum y, \\ b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 = \sum X_1 y, \\ b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2 = \sum X_2 y. \end{cases} \tag{7.97}$$

Розв'язуючи систему отримаємо значення коефіцієнтів b_j.

Зауваження 7.10. Додатково для розрахунків можна скласти таблицю:

i	y_i	Y_i	$\left(y_i-\overline{y}\right)^2$	$(y_i-Y_i)^2$	$\left(Y_i-\overline{y}\right)^2$	$\left\|\frac{y_i-Y_i}{y_i}\right\|$
1	y_1	…	…	…	…	…
2	y_2	…	…	…	…	…
…	…	…	…	…	…	…
m	y_m	…	…	…	…	…
Σ	…	…	…	…	…	…

2) *Перевіряємо адекватність моделі* (F-критерій Фішера). Знаходимо коефіцієнт детермінації R^2 за формулою (7.86) і фактичне значення $F_{\text{факт}}$ за формулою

$$F_{\text{факт}}=\frac{R^2}{\left(1-R^2\right)}\cdot\frac{m-3}{2}. \qquad (7.98)$$

За таблицею А.11 (див. Додаток А) знаходимо критичне значення критерію Фішера $F_{\text{кр}}$ для степенів свободи: k_1=2 (горизонталь), k_2=m–3 (вертикаль) та заданим рівнем значущості α. Перевіряємо справедливість гіпотез H_0 і H_1 стосовно значущості отриманого рівняння моделі.

3) *Розраховуємо середню помилку апроксимації* $\overline{A}$ за формулою (7.94). Її значення дає можливість зробити висновок про якість моделі.

4) *Визначаємо середні коефіцієнти еластичності* $\overline{E}_j$ для коефіцієнта b_1 і b_2 моделі регресії за формулою (7.95), щоб бачити вплив конкретного фактору на результат.

5) *Перевірка факторів на мультиколінеарність*. Для перевірки незалежності факторів X_1 та X_2 розраховуємо парний коефіцієнт кореляції Пірсона за формулою:

$$r_{x_1x_2}=\frac{m\cdot\sum X_1X_2-\sum X_1\cdot\sum X_2}{\sqrt{m\cdot\sum X_1^2-\left(\sum X_1\right)^2}\cdot\sqrt{m\cdot\sum X_2^2-\left(\sum X_2\right)^2}}. \qquad (7.99)$$

Для оцінки наявності мультиколінеарності використовується правило: фактори вважаються колінеарними (дублюючими один одного), якщо $\left|r_{x_1x_2}\right| \geq 0{,}8$. В іншому випадку мультиколінеарність у моделі відсутня, що підтверджує коректність включення обох факторів у рівняння регресії.

Корисно знайти парні коефіцієнти кореляції: між Y та X_1 і Y та X_2, які вимірюють тісноту зв'язку, визначають вплив X_1 і X_2 на Y. Знаходимо r_{yx_1} та r_{yx_2} за формулами:

$$r_{yx_1} = \frac{m \cdot \sum X_1 y - \sum X_1 \cdot \sum y}{\sqrt{m \cdot \sum X_1^2 - \left(\sum X_1\right)^2} \cdot \sqrt{m \cdot \sum y^2 - \left(\sum y\right)^2}}, \tag{7.100}$$

$$r_{yx_2} = \frac{m \cdot \sum X_2 y - \sum X_2 \cdot \sum y}{\sqrt{m \cdot \sum X_2^2 - \left(\sum X_2\right)^2} \cdot \sqrt{m \cdot \sum y^2 - \left(\sum y\right)^2}}. \tag{7.101}$$

Для аналізу сили зв'язку використовують шкалу Чеддока.

6) *Перевіряємо значущість коефіцієнтів рівняння регресії* за *t*- критерієм Стьюдента.

Для кожного коефіцієнта b_j (при факторі X_j моделі) формулюємо гіпотези H_0 і H_1:

нульова гіпотеза H_0: "Коефіцієнт b_j=0 − статистично незначущий. Це означає, що фактор X_j ніяк не впливає на Y";

альтернативна гіпотеза H_1: "Коефіцієнт $b_j \neq 0$ − статистично значущий. Фактор X_j реально впливає на результат, його зв'язок з Y підтверджено математично".

Розраховуємо фактичне *t*-значення для b_1 і b_2 за формулою:

$$t_{b_j,\text{факт}} = \frac{b_j}{m_{b_j}}, \tag{7.102}$$

де $m_{b_j} = \sqrt{\dfrac{S^2_{\text{зал}}}{\sum\limits_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2 \cdot \left(1 - r^2_{x_1x_2}\right)}}$, $\sigma^2_{\text{зал}} = \dfrac{1}{m-3} \cdot \sum\limits_{i=1}^{m}\left(y_i - Y_i\right)^2$.

За таблицею А.4 (див. Додаток А) знаходимо критичне $t_{кр}$ значення t-критерію Стьюдента для степенів свободи m–3 та заданим рівнем значущості α.

Якщо для коефіцієнта b_j виконується $|t_{факт}|>t_{кр}$, то гіпотеза H_0 відхиляється, коефіцієнт bj статистично значущий. Фактор X_j реально впливає на результат, його зв'язок з Y підтверджено математично.

Якщо для коефіцієнта b_j виконується $|t_{факт}|<t_{кр}$, то гіпотеза H_0 приймається, коефіцієнт b_j незначущий і фактор X_j ніяк не впливає на Y.

Перевіряємо значущість коефіцієнта b_0 рівняння регресії за t- критерієм Стьюдента.

Формулюємо гіпотези H_0 і H_1: нульова гіпотеза H_0: "b_0=0 − вільний член незначущий"; альтернативна гіпотеза H_1: "$b_0\neq0$ − вільний член значущий".

Розраховуємо фактичне t-значення для b_0 за формулою (7.102).

Для двофакторної моделі регресії помилка m_{b_0} для b_0 розраховується за формулою:

$$m_{b_0}=\sqrt{S^2_{зал}\left(\frac{1}{m}+\frac{\overline{X}_1^2\cdot S_2+\overline{X}_2^2\cdot S_1-2\overline{X}_1\overline{X}_2\left(\sum X_1X_2-\frac{1}{m}\sum X_1\sum X_2\right)}{\sum X_1\cdot\sum X_2\cdot\left(1-r^2_{x_1x_2}\right)}\right)}, \tag{7.103}$$

де $S_1=\sum\left(X_1-\overline{X}_1\right)^2$, $S_2=\sum\left(X_2-\overline{X}_2\right)^2$.

Перевірку прийняття гіпотез для b_0 здійснюємо аналогічно, як для коефіцієнтів b_1 і b_2.

Середні коефіцієнти еластичності $\overline{E}_j$ для лінійної залежності визначаємо для кожного коефіцієнта b_j регресії, окрім b_0, за формулою:

$$\overline{E}_j=b_j\cdot\frac{\overline{X}_j}{\overline{y}}, \tag{7.104}$$

Середні коефіцієнти еластичності $\overline{E}_j$ показують силу впливу конкретного фактору на результат. При зміні фактору X_j на 1% залежна

змінна Y зміниться в середньому на $\overline{E}_j$ відсотків.

Приклад 74. Досліджують залежність міцності бетонної конструкції (y, МПа) від часу затвердіння (X_1, дні) та вмісту зміцнюючої добавки (X_2, %). Отримали 7 даних: y: 15,22,28,35,46,55,62; X_1: 3,7,10,14,21,28,30; X_2: 1,1,2,2,3,3,4. Скласти рівняння двофакторної моделі лінійної регресії, що описує дану залежність. Перевірити адекватність моделі (F-критерій Фішера) при рівні значущості α=0,05. Знайти середню помилку апроксимації та середні коефіцієнти еластичності для коефіцієнта моделі регресії. Перевірити наявність мультиколінеарності та значущість кожного коефіцієнта моделі регресії за t-критерій Стьюдента при рівні значущості α=0,05.

Розв'язання. Для подальших розрахунків складемо таблицю 7.24.

Таблиця 7.24

№	X_1	X_2	y	X_1^2	X_2^2	$X_1 \cdot X_2$	$X_1 \cdot y$	$X_2 \cdot y$	y^2	$(X_1-\overline{X}_1)^2$	$(X_2-\overline{X}_2)^2$
1	3	1	15	9	1	3	45	15	225	172,738	1,654
2	7	1	22	49	1	7	154	22	484	83,594	1,654
3	10	2	28	100	4	20	280	56	784	37,736	0,082
4	14	2	35	196	4	28	490	70	1225	4,592	0,082
5	21	3	46	441	9	63	966	138	2116	23,590	0,510
6	28	3	55	784	9	84	1540	165	3025	140,588	0,510
7	30	4	62	900	16	120	1860	248	3844	192,016	2,938
Σ	**113**	**16**	**263**	**2479**	**44**	**325**	**5335**	**714**	**11703**	**654,86**	**7,43**

Маємо $\overline{y} \approx 37{,}571$, $\overline{X}_1 \approx 16{,}143$, $\overline{X}_2 \approx 2{,}286$.

Рівняння двофакторної моделі лінійної регресії має вигляд:

$$Y=b_0+b_1X_1+b_2X_2.$$

Коефіцієнтів b_j знайдемо МНК, розв'язавши систему лінійних рівнянь за формулою (7.97), в яку підставимо дані з таблиці 7.24:

$$\begin{cases} 7b_0 + 113b_1 + 16b_2 = 263, \\ 113b_0 + 2479b_1 + 325b_2 = 5335, \\ 16b_0 + 325b_1 + 44b_2 = 714. \end{cases}$$

Розв'язок системи: $b_0 \approx 8{,}818$; $b_1 \approx 1{,}362$; $b_2 \approx 2{,}96$.

Рівняння двофакторної моделі лінійної регресії:

$$Y=8{,}818+1{,}362 \cdot X_1+2{,}96 \cdot X_2.$$

Додатково складемо таблицю 7.25.

Таблиця 7.25

i	y_i	Y_i	$\left(y_i-\overline{y}\right)^2$	$(y_i-Y_i)^2$	$\left(Y_i-\overline{y}\right)^2$	$\left\lvert\frac{y_i-Y_i}{y_i}\right\rvert$
1	15	15,864	509,450	0,746	471,194	0,0576
2	22	21,312	242,456	0,473	264,355	0,0313
3	28	28,358	91,604	0,128	84,879	0,0128
4	35	33,806	6,610	1,426	14,175	0,0341
5	46	46,3	71,048	0,090	76,195	0,0065
6	55	55,834	303,770	0,696	333,537	0,0152
7	62	61,518	596,776	0,232	573,459	0,0078
Σ	**263**	**262,992**	**1821,714**	**3,792**	**1817,795**	**0,1652**

Знайдемо коефіцієнт детермінації R^2 за формулою (7.86).

$$R^2=1-\frac{3{,}792}{1821{,}714}=\frac{1817{,}795}{1821{,}714}\approx 0{,}9978.$$

Визначимо фактичне значення $F_{\text{факт}}$ за формулою (7.98):

$$F_{\text{факт}}=\frac{0{,}9978}{\left(1-0{,}9978\right)}\cdot\frac{7-3}{2}\approx 907{,}01.$$

За таблицею А.11 (див. Додаток А) для степенів свободи: k_1=2 (горизонталь), k_2=4 (вертикаль) та рівнем значущості α=0,05 знаходимо критичне значення критерію Фішера $F_{\text{кр}}$=6,94.

Оскільки $F_{\text{факт}}>F_{\text{кр}}$ (907,01>6,94), то рівняння регресії статистично значуще.

Розраховуємо середню помилку апроксимації $\overline{A}$ за формулою (7.94).

$$\overline{A}=0{,}1652/7\cdot 100\%=2{,}36\%.$$

Оскільки $\overline{A}$<8%, то точність моделі оцінюється як висока. Рівняння регресії адекватно відображає реальну залежність міцності бетону від вибраних факторів.

Визначимо середні коефіцієнти еластичності $\overline{E}_j$ для коефіцієнта b_1 і b_2 моделі регресії за формулою (7.95).

За часом затвердіння X_1: $\overline{E}_1=1{,}362\cdot 16{,}143/37{,}571\approx 0{,}585(\%)$. При

збільшенні часу затвердіння на 1% міцність зростає на 0,585%.

За вмістом зміцнюючої добавки X_2: $\overline{E}_2$ =2,96·2,286/37,571≈0,18(%). При збільшенні вмісту добавки на 1% міцність зростає на 0,18%.

Зробимо перевірку факторів на мультиколінеарність: зв'язок між часом X_1 та добавкою X_2. Застосуємо формулу (7.99):

$$r_{x_1x_2} = \frac{7 \cdot 325 - 113 \cdot 16}{\sqrt{7 \cdot 2479 - (113)^2} \cdot \sqrt{7 \cdot 44 - (16)^2}} = \frac{467}{\sqrt{4584} \cdot \sqrt{52}} \approx 0{,}957 .$$

За розрахунком $r_{x_1x_2} \approx 0{,}957$ – це дуже високий показник. Оскільки коефіцієнт парної кореляції між факторами X_1 та добавкою X_2 $\left|r_{x_1x_2}\right| \geq 0{,}8$, то в моделі є сильна мультиколінеарність. Через те, що X_1 і X_2 зростають у таблиці практично синхронно, математично важко «розділити», який саме фактор впливає на успіх у зростанні міцності бетону. Це і є наслідком мультиколінеарності.

Для прогнозу Y (міцності) модель придатна і даватиме дуже точні результати тільки тоді, доки зберігається саме така залежність між X_1 і X_2, як за умовою прикладу, прогноз буде ідеальним.

Але для керування процесом модель непридатна, оскільки не можемо припустити, що буде з Y при зміні значень факторів X_1 і X_2 моделі. Модель видасть значення, але воно буде недостовірне, оскільки коефіцієнти b_1 і b_2 через колінеарність можуть бути зміщені.

Знаходимо r_{yx_1} та r_{yx_2} за формулами (7.100) та (7.101):

$$r_{yx_1} = \frac{7 \cdot 5335 - 113 \cdot 263}{\sqrt{7 \cdot 2479 - (113)^2} \cdot \sqrt{7 \cdot 11703 - (263)^2}} = \frac{7626}{\sqrt{4584} \cdot \sqrt{12752}} \approx 0{,}997 ,$$

$$r_{yx_2} = \frac{7 \cdot 714 - 16 \cdot 263}{\sqrt{7 \cdot 44 - (16)^2} \cdot \sqrt{7 \cdot 11703 - (263)^2}} = \frac{790}{\sqrt{52} \cdot \sqrt{12752}} \approx 0{,}970 .$$

Перевіряємо значущість коефіцієнтів рівняння регресії за t- критерієм Стьюдента.

Для кожного коефіцієнта b_j (при факторі X_j моделі) формулюємо гіпотези H_0 і H_1:

нульова гіпотеза H_0: "Коефіцієнт b_j=0 − статистично незначущий. Це означає, що фактор X_j ніяк не впливає на Y";

альтернативна гіпотеза H_1: "Коефіцієнт $b_j \neq 0$ – статистично значущий. Фактор X_j реально впливає на результат, його зв'язок з Y підтверджено математично".

Знайдемо: $1-r^2_{x_1x_2}=1-0{,}957^2\approx 0{,}084$; $\sigma^2_{\text{зал}}=\frac{1}{4}\cdot 3{,}792=0{,}948$;

$$m_{b_1}=\sqrt{\frac{0{,}948}{654{,}86\cdot 0{,}084}}\approx\sqrt{\frac{0{,}948}{55{,}107}}\approx 0{,}131;$$

$$m_{b_2}=\sqrt{\frac{0{,}948}{7{,}43\cdot 0{,}084}}\approx\sqrt{\frac{0{,}948}{0{,}624}}\approx 1{,}232.$$

Розраховуємо фактичне t-значення для b_1 і b_2 за формулою (7.102):

$$t_{b_1,\text{факт}}=\frac{1{,}362}{0{,}131}\approx 10{,}397,\ t_{b_2,\text{факт}}=\frac{2{,}96}{1{,}232}\approx 2{,}403.$$

За таблицею А.4 (див. Додаток А) для степенів свободи k=4 і рівню значущості α=0,05 знаходимо значення t-критерію Стьюдента $t_{\text{кр}}$=2,776.

Для b_1: оскільки $|t_{\text{факт}}|>t_{\text{кр}}$ (10,397>2,776), то гіпотезу H_0 відхиляємо, коефіцієнт b_1 статистично значущий. Фактор часу затвердіння X_1 реально впливає на результат.

Для b_2: оскільки $|t_{\text{факт}}|<t_{\text{кр}}$ (2,403<2,776), то гіпотезу H_0 приймаємо, коефіцієнт b_2 незначущий і фактор вміст зміцнюючої добавки X_2 ніяк не впливає на Y. Даний результат обумовлений сильною мультиколінеарністю факторів, що ускладнює поділ вкладу часу та добавки у підсумкову міцність бетону.

Перевіримо значущість коефіцієнта b_0 рівняння регресії. Знайдемо значення помилки m_{b_0} за формулою (7.103), підставивши суми з таблиці 24 і таблиці 25:

$$\overline{X}_1^2\approx 260{,}596,\ \overline{X}_2^2\approx 5{,}226,\ S_1=654{,}86,\ S_2=7{,}43,\ 2\overline{X}_1\overline{X}_2\approx 73{,}806,$$

$$\sum X_1X_2=325,\ \frac{1}{7}\sum X_1\sum X_2\approx 258{,}286.$$

$$m_{b_0}=\sqrt{0{,}948\cdot\left(\frac{1}{7}+\frac{260{,}596\cdot 7{,}43+5{,}226\cdot 654{,}86-73{,}806(325-258{,}286)}{113\cdot 16\cdot 0{,}084}\right)}=$$

$$=\sqrt{0{,}948\cdot\left(\frac{1}{7}+\frac{367{,}919}{151{,}872}\right)}\approx\sqrt{0{,}948\cdot 2{,}565}\approx 1{,}56.$$

Розраховуємо для b_0 фактичне значення $t_{факт}$ за формулою (7.102):

$$t_{b_0,\text{факт}} = \frac{b_0}{m_{b_0}} = \frac{8{,}818}{1{,}56} \approx 5{,}653.$$

Критичне $t_{кр}$ значення t-критерію Стьюдента для степенів свободи k=4 і рівню значущості α=0,05 знайдено вище: $t_{кр}$=2,776.

Оскільки $|t_{факт}|>t_{кр}$ (5,653>2,776), то гіпотезу H_0 відхиляємо, коефіцієнт b_0 статистично значущий. Це означає, що при нульових значеннях факторів X_1 і X_2 значення Y обґрунтовано відрізняється від нуля і дорівнює приблизно 8,818.

Відповідь: Y=8,818+1,362·X_1+2,96·X_2; $F_{факт}$=907,01 − рівняння регресії статистично значуще; середню помилка апроксимації $\overline{A}$ =2,36%ρ_{xy}=0,9978; середні коефіцієнти еластичності $\overline{E}_1$=0,585%, $\overline{E}_2$=0,18%; в моделі є сильна мультиколінеарність.

7.3 Завдання для самостійної роботи

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>1. Перевірити чи існує лінійна залежність між кількістю рекламних витрат (X, тис.ум.од) та обсягом продажів (Y, одиниць) для деякої компанії, а також оцінити тісноту зв'язку. Отримали дані для X: 10,12,15,13,18. Для Y: 100,110,130,120,150. Рівень значущості α=0,05.</td><td>r_{xy}≈0,997. Існує сильний прямий лінійний зв'язок між рекламними витратами та обсягом продажу.</td></tr>
<tr><td>2. Досліджують зв'язок між часом (X) підготовки 20 студентів до екзамену та підсумковим балом (Y). За результатами склали кореляційну таблицю:
<table>
<tr><th>Y \ X</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>n_y</th></tr>
<tr><td>90</td><td>–</td><td>–</td><td>–</td><td>–</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>80</td><td>–</td><td>–</td><td>1</td><td>2</td><td>2</td><td>1</td><td>6</td></tr>
<tr><td>70</td><td>–</td><td>1</td><td>3</td><td>1</td><td>–</td><td>–</td><td>5</td></tr>
<tr><td>60</td><td>4</td><td>2</td><td>–</td><td>–</td><td>–</td><td>–</td><td>6</td></tr>
<tr><td>n_x</td><td>4</td><td>3</td><td>4</td><td>3</td><td>3</td><td>3</td><td>20</td></tr>
</table>
Знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості α=0,05.</td><td>r_{xy}≈0,913. Зв'язок між часом підготовки та підсумковим балом статистично значущий.</td></tr>
</table>

<table>
<tr><th>Завдання</th><th>Відповідь</th></tr>
<tr><td>3. Досліджують залежність між досвідом роботи у роках (X) та кількістю оброблених заявок на день (Y) для 100 співробітників. За результатами склали кореляційну таблицю:
<table>
<tr><th>Y (y_i) \ X (x_i)</th><th>0−4 (2)</th><th>4−8 (6)</th><th>8−12 (10)</th><th>12−16 (14)</th><th>16−20 (18)</th><th>n_y</th></tr>
<tr><td>40−50 (45)</td><td>−</td><td>−</td><td>2</td><td>6</td><td>7</td><td>15</td></tr>
<tr><td>30−40 (35)</td><td>−</td><td>5</td><td>15</td><td>10</td><td>5</td><td>35</td></tr>
<tr><td>20−30 (25)</td><td>5</td><td>10</td><td>10</td><td>5</td><td>−</td><td>30</td></tr>
<tr><td>10−20 (15)</td><td>10</td><td>5</td><td>5</td><td>−</td><td>−</td><td>20</td></tr>
<tr><td>n_x</td><td>15</td><td>20</td><td>32</td><td>21</td><td>12</td><td>n=100</td></tr>
</table>
Знайти вибірковий коефіцієнт лінійної кореляції Пірсона та перевірити його значущість при рівні значущості α=0,05.</td><td>r_{xy}≈0,693.
Досвід роботи справді впливає на кількість оброблених заявок.</td></tr>
<tr><td>4. З нормальної генеральної сукупності вилучено вибірку обсягом n=122. Знайдено вибірковий коефіцієнт кореляції r_{xy}=0,4. Перевірити нульову гіпотезу H_0: "Рівність нулю генерального коефіцієнта кореляції" при рівні значущості α=0,05.</td><td>$t_{факт}$≈4,79;
$t_{кр}$=1,98.
Нульову гіпотезу відхиляємо.</td></tr>
<tr><td>5. Студенти складають іспит з двох дисциплін: математики (X) та теоретичної механіки (Y). За набраними балами 10 студентів отримали ранги:
<table>
<tr><td>X</td><td>2</td><td>4</td><td>5</td><td>1</td><td>7,5</td><td>7,5</td><td>7,5</td><td>7,5</td><td>3</td><td>10</td></tr>
<tr><td>Y</td><td>2,5</td><td>6</td><td>4</td><td>1</td><td>2,5</td><td>7</td><td>8</td><td>9,5</td><td>5</td><td>9,5</td></tr>
</table>
Обчислити ранговий коефіцієнт кореляції Спірмена та перевірити його значущість при α=0,05.</td><td>$\rho_{факт}$≈0,755;
$t_{факт}$≈3,26;
$t_{кр}$=2,31.
Зв'язок між оцінками двох дисциплін досить тісний.</td></tr>
<tr><td>6. Досліджують вплив рівня мотивації (X, бали) на ефективність вирішення складних логічних завдань (Y, кількість) для 5 студентів. Отримали дані: X={10,20,30,40,50} і Y={(10,20,30,20,11}. Обчислити ранговий коефіцієнт кореляції Спірмена та встановити характер зв'язку.</td><td>$\rho_{факт}$≈0,255;
η≈0,98.
Наявність сильного криволінійного зв'язку.</td></tr>
<tr><td>7. Для 6 робітників підприємства отримали дані про досвід роботи (X, років) та рівень доходу (Y, ум. од.). Дані для X: 1,2,3,4,5,6 і для Y: 2,4,5,7,9,11. Скласти рівняння регресії та дати оцінку якості коефіцієнту регресії.</td><td>Y=0,135+1,77·x.
$t_{факт}$≈21,3.
Коефіцієнт регресії статистично значущий</td></tr>
</table>

Завдання	Відповідь
8. Знайти параметри a і b лінійної регресії $y=ax+b$ ЗМНК для наступних даних: x_i: 1, 2, 3, 4 y_i: 3, 5, 6, 8 n_i: 2, 3, 1, 4 де x_i – значення, n_i – частота. Значення y_i є середніми для кожного x_i.	a=1,61; b=1,55.
10. Знайти оцінку параметрів a і b лінійної регресії $y=ax+b$ методом найменших квадратів для наступних даних: x_i: 1, 2, 3, 4, 5 y_i: 3, 5, 6, 8, 11	a=1,9; b=0,9.
11. Вивчають залежність витрати палива (Y, літри) від швидкості (X, км/год) по 5 вимірах. Отримали дані для X: 60,80,100,120,140 і для Y: 6,7.5,9,10.5,12. При рівні значущості α=0.05 перевірити для моделі: якість (R^2), значимість (F-критерій), точність ($\overline{A}$).	$R^2\approx$0,92. $F_{факт}\approx$32,5. $\overline{A}$ =4,5%
12. Зроблено експеримент. Отримано 10 пар даних. Час тренування (у % до номінального), x_i: 70, 120, 140, 90, 150, 100, 130, 60, 110, 80. Час безвідмовної роботи (у % до гарантійного), y_i: 78, 135, 138, 108, 142, 110, 138, 74, 98, 72. Скласти рівняння лінійної регресії. При рівні значущості α=0,05 за F-критерієм перевірити значущість рівняння моделі. За t- критерієм перевірити значущість коефіцієнта b_1 рівняння регресії.	Y=18,265+0,867·x; $F_{факт}\approx$54,02; $t_{факт}\approx$8,42. Рівняння моделі і коефіцієнт b_1 статистично значущі.
13. Провели дослідження впливу витрат на рекламу (X_1) і середньої ціни товару (X_2) на обсяг продажів (Y). За отриманими даними 6 дослідів для Y: {10,12,15,18,20,22}, X_1: {2.4.5.7.8.10}. X_2: {5.4,3,2,2,1} скласти рівняння двофакторної моделі лінійної регресії. При рівні значущості α=0,05 за F-критерієм перевірити значущість рівняння моделі. Знайти скорегований $\overline{R}^2$, Перевірити значущість кожного коефіцієнта моделі регресії за t-критерій Стьюдента при рівні значущості α=0,05.	Y=8,44+1,34·X_1− −1,31·X_2; $F_{факт}\approx$0,98: Модель статистично незначна (адекватність не підтверджена). $\overline{R}^2\approx$−0,006. b_0: $t_{факт}\approx$0,83: b_1: $t_{факт}\approx$1,16: b_2: $t_{факт}\approx$0,53: Жоден фактор не визнаний значним

Завдання	Відповідь
9. Досліджують залежність зносу деталі (Y, мм) від часу роботи (X, сотні годин). Отримали дані для X: 1, 2, 3, 4 ,5 і для Y: 0.15, 0.28, 0.42, 0.58, 0.72. Знайти коефіцієнт кореляції Пірсона r_{xy}, коефіцієнт детермінації R^2. При рівні значущості α=0,05 за t-критерієм Стьюдента перевірити значущість отриманих результатів.	$r_{xy}\approx$0,999; $R^2\approx$0,998; $t_{\text{факт}}\approx$38,7. Отримані результати статистично значущі.
14. На 10 дослідних ділянках однакового розміру отримано такі дані про врожайність X (т) та вміст білка Y (%) для деякої культури: X: 9,9 10,2 11,0 11,6 11,8 12,5 12,8 13,5 14,3 14,4; Y: 10,7 10,8 12,1 12,5 12,8 12,8 12,4 11,8 10,8 10,6. Скласти рівняння квадратної регресії. Знайти кореляційне відношення ρ_{xy} і коефіцієнт детермінації R^2.	$Y=-50{,}02+10{,}309\cdot x-0{,}4237\cdot x^2$. $\rho_{xy}\approx$0,99; $R^2\approx$0,98.

ЛІТЕРАТУРА

1. Бишевець Н. Г., Омецинська Н. Г., Юсипів Т. В. Теорія ймовірностей та математична статистика з використанням табличного процесора MS EXCEL : навч. посіб. Одеса : Гельветика, 2021. 234 с.

2. Вища математика: зб. задач : у 2 ч. Ч. 2. : Звичайні диференціальні рівняння. Операційне числення. Ряди. Рівняння мат. фізики. Стійкість за Ляпуновим. Елементи теорії ймовірностей і математичної статистики. Методи оптимізації і задачі керування. Варіаційне числення. Числові методи : навч. посіб. для студ. вищ. техн. навч. закл. / П. П. Овчинников та ін. ; за заг. ред. П. П. Овчинникова. Київ : Техніка, 2004. 376 с.

3. Волощенко А. Б., Джалладова І. А. Теорія ймовірностей та математична статистика : навч.-метод. посіб. для самост. вивч. дисц. Київ : КНЕУ, 2003. 256 с.

4. Жалдак М. І., Кузьміна Н. М, Михалін Г. О. Теорія ймовірностей і математична статистика. Полтава : Довкілля-К, 2009. 509 с.

5. Жильцов О. Б., Михалін Г. О. Теорія ймовірностей та математична статистика у прикладах і задачах : навч. посіб. для студ. вищ. навч. закл. / за ред. Г. О. Михаліна. Київ : Київ. ун-т ім. Б. Грінченка, 2015. 336 с.

6. Жлуктенко В. І., Наконечний С. І., Савіна С. С. Теорія ймовірностей і математична статистика : навч.-метод. посіб. : у 2 ч. Ч. 2 : Математична статистика. 2-ге вид., без змін. Київ : КНЕУ, 2007. 368 с.

7. Кармелюк Г. І. Теорія ймовірностей та математична статистика : посіб. з розв'яз. задач. Київ : Центр учбової літератури, 2007. 576 с.

8. Килимник І.М. Практикум з елементів теорії ймовірностей : навч. посіб. Запоріжжя : НУ «Запорізька політехніка», 2024. 256 с.

9. Методичні вказівки до виконання контрольної роботи з теорії ймовірностей та математичної статистики для студентів заочної форми навчання транспортного факультету. Частина 1 (теоретичний матеріал). / Запорізьк. нац. техн. ун-т. Каф. вищої математики ; уклад. : І. М. Килимник, Т. Г. Полякова. Запоріжжя : ЗНТУ, 2019. 110 с.

10. Овчинников П. П., Михайленко В. М. Вища математика : підручник : у 2 ч. Ч. 2 : Диференціальні рівняння. Операційне числення. Ряди та їх застосування. Стійкість за Ляпуновим. Рівняння математичної фізики. Оптимізація і керування. Теорія ймовірностей.

Числові методи / за заг. ред. П. П. Овчинникова. Київ : Техніка, 2004. 792 с.

11. Черняк І. О., Ставицький А. В., Обушна О. М. Теорія ймовірностей та математична статистика. Збірник задач : навч. посіб. Київ : Знання, 2000. 199 с.

12. Cremonezi L. Introduction statistics for market research. Ipsos Connect, 2018. 44 p. URL: https://www.ipsos.com/sites/default/files/ct/publication/documents/2018-02/intro_to_stats_2018.pdf.

13. Fein E. C., Gilmour J., Machin T., Hendry L. Statistics for Research Students. Australia : University Of Southern Queensland Toowoomba, 2021. 109 p.

14. Gupta S. C., Kapoor V. K. Fundamentals of mathematical statistics. Sultan Chand & Sons, 2020. 1303 p.

15. Montgomery D. C., Runger G. C. Applied statistics and probability for engineers. John Wiley & Sons, 2020. 836 p.

ДОДАТОК А

Таблиця А.1 – Значення функції $\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$.

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1.0	0,242	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,054	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139

Продовження Таблиці А.1

x	0	1	2	3	4	5	6	7	8	9
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблиця А.2 – Значення функції $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,11	0,0438	0,22	0,0871	0,33	0,1293
0,01	0,0040	0,12	0,0478	0,23	0,0910	0,34	0,1331
0,02	0,0080	0,13	0,0517	0,24	0,0948	0,35	0,1368
0,03	0,0120	0,14	0,0557	0,25	0,0987	0,36	0,1406
0,04	0,0160	0,15	0,0596	0,26	0,1026	0,37	0,1443
0,05	0,0199	0,16	0,0636	0,27	0,1064	0,38	0,1480
0,06	0,0239	0,17	0,0675	0,28	0,1103	0,39	0,1517
0,07	0,0279	0,18	0,0714	0,29	0,1141	0,40	0,1554
0,08	0,0319	0,19	0,0753	0,30	0,1179	0,41	0,1591
0,09	0,0359	0,20	0,0793	0,31	0,1217	0,42	0,1628
0,10	0,0398	0,21	0,0832	0,32	0,1255	0,43	0,1664

Продовження Таблиці А.2

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,44	0,1700	0,74	0,2703	1,04	0,3508	1,34	0,4099
0,45	0,1736	0,75	0,2734	1,05	0,3531	1,35	0,4115
0,46	0,1772	0,76	0,2764	1,06	0,3554	1,36	0,4131
0,47	0,1808	0,77	0,2794	1,07	0,3577	1,37	0,4147
0,48	0,1844	0,78	0,2823	1,08	0,3599	1,38	0,4162
0,49	0,1879	0,79	0,2852	1,09	0,3621	1,39	0,4177
0,50	0,1915	0,80	0,2881	1,10	0,3643	1,40	0,4192
0,51	0,1950	0,81	0,2910	1,11	0,3665	1,41	0,4207
0,52	0,1985	0,82	0,2939	1,12	0,3686	1,42	0,4222
0,53	0,2019	0,83	0,2967	1,13	0,3708	1,43	0,4236
0,54	0,2054	0,84	0,2995	1,14	0,3729	1,44	0,4251
0,55	0,2088	0,85	0,3023	1,15	0,3749	1,45	0,4265
0,56	0,2123	0,86	0,3051	1,16	0,3770	1,46	0,4279
0,57	0,2157	0,87	0,3078	1,17	0,3790	1,47	0,4292
0,58	0,2190	0,88	0,3106	1,18	0,3810	1,48	0,4306
0,59	0,2224	0,89	0,3133	1,19	0,3830	1,49	0,4319
0,60	0,2257	0,90	0,3159	1,20	0,3849	1,50	0,4332
0,61	0,2291	0,91	0,3186	1,21	0,3869	1,51	0,4345
0,62	0,2324	0,92	0,3212	1,22	0,3883	1,52	0,4357
0,63	0,2357	0,93	0,3238	1,23	0,3907	1,53	0,4370
0,64	0,2389	0,94	0,3264	1,24	0,3925	1,54	0,4382
0,65	0,2422	0,95	0,3289	1,25	0,3944	1,55	0,4394
0,66	0,2454	0,96	0,3315	1,26	0,3962	1,56	0,4406
0,67	0,2486	0,97	0,3340	1,27	0,3980	1,57	0,4418
0,68	0,2517	0,98	0,3365	1,28	0,3997	1,58	0,4429
0,69	0,2549	0,99	0,3389	1,29	0,4015	1,59	0,4441
0,70	0,2580	1,00	0,3413	1,30	0,4032	1,60	0,4452
0,71	0,2611	1,01	0,3438	1,31	0,4049	1,61	0,4463
0,72	0,2642	1,02	0,3461	1,32	0,4066	1,62	0,4474
0,73	0,2673	1,03	0,3485	1,33	0,4082	1,63	0,4484

Продовження Таблиці А.2

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,64	0,4495	1,94	0,4738	2,24	0,4875	2,54	0,4945
1,65	0,4505	1,95	0,4744	2,25	0,4878	2,55	0,4946
1,66	0,4515	1,96	0,4750	2,26	0,4881	2,56	0,4948
1,67	0,4525	1,97	0,4756	2,27	0,4884	2,57	0,4949
1,68	0,4535	1,98	0,4761	2,28	0,4887	2,58	0,4951
1,69	0,4545	1,99	0,4767	2,29	0,4890	2,59	0,4952
1,70	0,4554	2,00	0,4772	2,30	0,4893	2,60	0,4953
1,71	0,4564	2,01	0,4778	2,31	0,4896	2,61	0,4955
1,72	0,4573	2,02	0,4783	2,32	0,4898	2,62	0,4956
1,73	0,4582	2,03	0,4788	2,33	0,4901	2,63	0,4957
1,74	0,4591	2,04	0,4793	2,34	0,4904	2,64	0,4959
1,75	0,4599	2,05	0,4798	2,35	0,4906	2,65	0,4960
1,76	0,4608	2,06	0,4803	2,36	0,4909	2,66	0,4961
1,77	0,4616	2,07	0,4808	2,37	0,4911	2,67	0,4962
1,78	0,4625	2,08	0,4812	2,38	0,4913	2,68	0,4963
1,79	0,4633	2,09	0,4817	2,39	0,4916	2,69	0,4964
1,80	0,4641	2,10	0,4821	2,40	0,4918	2,70	0,4965
1,81	0,4649	2,11	0,4826	2,41	0,4920	2,71	0,4966
1,82	0,4656	2,12	0,4830	2,42	0,4922	2,72	0,4967
1,83	0,4664	2,13	0,4834	2,43	0,4925	2,73	0,4968
1,84	0,4671	2,14	0,4838	2,44	0,4927	2,74	0,4969
1,85	0,4678	2,15	0,4842	2,45	0,4929	2,75	0,4970
1,86	0,4686	2,16	0,4846	2,46	0,4931	2,76	0,4971
1,87	0,4693	2,17	0,4850	2,47	0,4932	2,77	0,4972
1,88	0,4699	2,18	0,4854	2,48	0,4934	2,78	0,4973
1,89	0,4706	2,19	0,4857	2,49	0,4936	2,79	0,4974
1,90	0,4713	2,20	0,4861	2,50	0,4938	2,80	0,4974
1,91	0,4719	2,21	0,4864	2,51	0,4940	2,81	0,4975
1,92	0,4726	2,22	0,4868	2,52	0,4941	2,82	0,4976
1,93	0,4732	2,23	0,4871	2,53	0,4943	2,83	0,4977

Продовження Таблиці А.2

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
2,84	0,4977	3,14	0,4992	3,44	0,4997	3,74	0,4999
2,85	0,4978	3,15	0,4992	3,45	0,4997	3,75	0,4999
2,86	0,4979	3,16	0,4992	3,46	0,4997	3,76	0,4999
2,87	0,4979	3,17	0,4992	3,47	0,4997	3,77	0,4999
2,88	0,4980	3,18	0,4993	3,48	0,4997	3,78	0,4999
2,89	0,4981	3,19	0,4993	3,49	0,4998	3,79	0,4999
2,90	0,4981	3,20	0,4993	3,50	0,4998	3,80	0,4999
2,91	0,4982	3,21	0,4993	3,51	0,4998	3,81	0,4999
2,92	0,4982	3,22	0,4994	3,52	0,4998	3,82	0,4999
2,93	0,4983	3,23	0,4994	3,53	0,4998	3,83	0,4999
2,94	0,4984	3,24	0,4994	3,54	0,4998	3,84	0,4999
2,95	0,4984	3,25	0,4994	3,55	0,4998	3,85	0,4999
2,96	0,4985	3,26	0,4994	3,56	0,4998	3,86	0,4999
2,97	0,4985	3,27	0,4995	3,57	0,4998	3,87	0,4999
2,98	0,4986	3,28	0,4995	3,58	0,4998	3,88	0,4999
2,99	0,4986	3,29	0,4995	3,59	0,4998	3,89	0,4999
3,00	0,49865	3,30	0,4995	3,60	0,4998	3,90	0,4999
3,01	0,4987	3,31	0,4995	3,61	0,4998	3,91	0,4999
3,02	0,4687	3,32	0,4995	3,62	0,4999	3,92	0,4999
3,03	0,4988	3,33	0,4996	3,63	0,4999	3,93	0,4999
3,04	0,4988	3,34	0,4996	3,64	0,4999	3,94	0,4999
3,05	0,4989	3,35	0,4996	3,65	0,4999	3,95	0,4999
3,06	0,4989	3,36	0,4996	3,66	0,4999	3,96	0,4999
3,07	0,4989	3,37	0,4996	3,67	0,4999	3,97	0,4999
3,08	0,4990	3,38	0,4996	3,68	0,4999	3,98	0,4999
3,09	0,4990	3,39	0,4997	3,69	0,4999	3,99	0,4999
3,10	0,4990	3,40	0,4997	3,70	0,4999	4,00	0,5
3,11	0,4991	3,41	0,4997	3,71	0,4999	$x > 4$	0,5
3,12	0,4991	3,42	0,4997	3,72	0,4999		
3,13	0,4991	3,43	0,4997	3,73	0,4999		

Таблиця А.3 – Значення $t(\gamma,k)$

γ – надійність, $k=n-1$ – число степенів свободи

γ / k	Одностороння критична область										
	0,75	**0,80**	**0,85**	**0,90**	**0,95**	**0,975**	**0,99**	**0,995**	**0,9975**	**0,999**	**0,9995**
	Двостороння критична область										
	0,50	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**0,98**	**0,99**	**0,995**	**0,998**	**0,999**
1	1,000	1,376	1,963	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,080	1,386	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	0.860	1,064	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850

Продовження Таблиці А.3

γ / k	Одностороння критична область										
	0,75	**0,80**	**0,85**	**0,90**	**0,95**	**0,975**	**0,99**	**0,995**	**0,9975**	**0,999**	**0,9995**
	Двостороння критична область										
	0,50	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**0,98**	**0,99**	**0,995**	**0,998**	**0,999**
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,767
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Таблиця А.4 – Критичні точки розподілу Стьюдента $t_{\alpha,k}$

$k=n-1$ – число степенів свободи, $\alpha=1-\gamma$,
де γ – надійність,

k	Рівень значущості α (двостороння критична область)						
	0,001	0,002	0,01	0,02	0,05	0,1	0,2
1	636,619	318,310	63,657	31,821	12,706	6,314	3,078
2	31,598	22,326	9,925	6,965	4,303	2,920	1,886
3	12,941	10,213	5,841	4,541	3,182	2,353	1,638
4	8,610	7,173	4,604	3,747	2,776	2,132	1,533
5	6,859	5,893	4,032	3,365	2,571	2,015	1,476
6	5,959	5,208	3,707	3,143	2,447	1,943	1,440
7	5,405	4,785	3,499	2,998	2,365	1,895	1,415
8	5,041	4,501	3,355	2,896	2,306	1,860	1,397
9	4,781	4,297	3,249	2,821	2,262	1,833	1,383
10	4,583	4,144	3,169	2,764	2,228	1,812	1,372
11	4,437	4,025	3,106	2,718	2,201	1,796	1,363
12	4,318	3,930	3,055	2,681	2,179	1,782	1,356
13	4,221	3,852	3,012	2,650	2,160	1,771	1,350
14	4,140	3,787	2,977	2,624	2,145	1,761	1,345
15	4,073	3,733	2,947	2,602	2,131	1,753	1,341
16	4,015	3,686	2,921	2,583	2,120	1,746	1,337
17	3,965	3,646	2,898	2,567	2,110	1,740	1,333
18	3,922	3,610	2,878	2,552	2,101	1,734	1,330
19	3,883	3,579	2,861	2,539	2,093	1,729	1,328
20	3,850	3,552	2,845	2,528	2,086	1,725	1,325
21	3,819	3,527	2,831	2,518	2,080	1,721	1,323
22	3,792	3,505	2,819	2,508	2,074	1,717	1,321
23	3,767	3,485	2,807	2,500	2,069	1,714	1,319
24	3,745	3,467	2,797	2,492	2,064	1,711	1,318
25	3,725	3,450	2,787	2,485	2,060	1,708	1,316
26	3,707	3,435	2,779	2,479	2,056	1,706	1,315
27	3,690	3,421	2,771	2,473	2,052	1,703	1,314
28	3,674	3,408	2,763	2,467	2,048	1,701	1,313
	0,0005	0,001	0,005	0,01	0,025	0,05	0,1
	Рівень значущості α (одностороння критична область)						

Продовження Таблиці А.4

k	Рівень значущості α (двостороння критична область)						
	0,001	0,002	0,01	0,02	0,05	0,1	0,2
29	3,659	3,396	2,756	2,462	2,045	1,699	1,311
30	3,646	3,385	2,750	2,457	2,042	1,697	1,310
40	3,551	3,307	2,704	2,423	2,021	1,684	1,303
60	3,460	3,232	2,660	2,390	2,000	1,671	1,296
120	3,373	3,170	2,617	2,358	1,980	1,658	1,289
∞	3,291	3,090	2,576	2,326	1,960	1,645	1,282
	0,0005	0,001	0,005	0,01	0,025	0,05	0,1
	Рівень значущості α (одностороння критична область)						

k	Рівень значущості α (двостороння критична область)						
	0,3	0,4	0,5	0,6	0,7	0,8	0,9
1	1,963	1,376	1,000	0,727	0,510	0,325	0,158
2	1,386	1,061	0,816	0,617	0,445	0,289	0,142
3	1,250	0,978	0,765	0,584	0,424	0,277	0,137
4	1,190	0,941	0,741	0,569	0,414	0,271	0,134
5	1,156	0,920	0,727	0,559	0,408	0,267	0,132
6	1,134	0,906	0,718	0,553	0,404	0,265	0,131
7	1,119	0,896	0,711	0,549	0,402	0,263	0,130
8	1,108	0,889	0,706	0,546	0,399	0,262	0,130
9	1,100	0,883	0,703	0,543	0,398	0,261	0,129
10	1,093	0,879	0,700	0,542	0,397	0,260	0,129
11	1,088	0,876	0,697	0,540	0,396	0,260	0,129
12	1,083	0,873	0,695	0,539	0,395	0,259	0,128
13	1,079	0,870	0,694	0,538	0,394	0,259	0,128
14	1,076	0,868	0,692	0,537	0,393	0,258	0,128
15	1,074	0,866	0,691	0,536	0,393	0,258	0,128
16	1,071	0,865	0,690	0,535	0,392	0,258	0,128
17	1,069	0,863	0,689	0,534	0,392	0,257	0,128
18	1,067	0,862	0,688	0,534	0,392	0,257	0,127
19	1,066	0,861	0,688	0,533	0,391	0,257	0,127
20	1,064	0,860	0,687	0,533	0,391	0,257	0,127
	0,15	0,2	0,25	0,3	0,35	0,4	0,45
	Рівень значущості α (одностороння критична область)						

Продовження Таблиці А.4

k	Рівень значущості α (двостороння критична область)						
	0,3	0,4	0,5	0,6	0,7	0,8	0,9
21	1,063	0,859	0,686	0,532	0,391	0,257	0,127
22	1,061	0,858	0,686	0,532	0,390	0,256	0,127
23	1,060	0,858	0,685	0,532	0,390	0,256	0,127
24	1,059	0,857	0,685	0,531	0,390	0,256	0,127
25	1,058	0,856	0,684	0,531	0,390	0,256	0,127
26	1,058	0,856	0,684	0,531	0,390	0,256	0,127
27	1,057	0,855	0,684	0,531	0,389	0,256	0,127
28	1,056	0,855	0,683	0,530	0,389	0,256	0,127
29	1,055	0,854	0,683	0,530	0,389	0,256	0,127
30	1,055	0,854	0,683	0,530	0,389	0,256	0,127
40	1,050	0,851	0,681	0,529	0,388	0,255	0,126
60	1,046	0,848	0,679	0,527	0,387	0,254	0,126
120	1,041	0,845	0,677	0,526	0,386	0,254	0,126
∞	1,036	0,842	0,674	0,524	0,385	0,253	0,126
	0,15	0,2	0,25	0,3	0,35	0,4	0,45
	Рівень значущості α (одностороння критична область)						

Таблиця А.5 – Критичні точки розподілу Пірсона $\chi^2_{\alpha;k}$

k – число степенів свободи

k	Рівень значущості α						
	0,999	0,995	0,99	0,98	0,975	0,95	0,9
1	0,05157	0,000039	0,00016	0,03628	0,00098	0,00393	0,0158
2	0,00200	0,0100	0,0201	0,0404	0,0506	0,1030	0,211
3	0,0243	0,0717	0,1150	0,185	0,216	0,352	0,584
4	0,0908	0,207	0,297	0,429	0,484	0,711	1,064
5	0,210	0,412	0,554	0,752	0,831	1,145	1,610
6	0,381	0,676	0,872	1,134	1,237	1,635	2,204
7	0,598	0,989	1,239	1,564	1,690	2,167	2,833
8	0,857	1,344	1,646	2,032	2,180	2,733	3,490
9	1,152	1,735	2,088	2,532	2,700	3,325	4,168
10	1,479	2,156	2,558	3,059	3,274	3,240	4,865
11	1,834	2,603	3,053	3,609	3,816	4,575	5,578

Продовження Таблиці А.5

k	Рівень значущості α						
	0,999	0,995	0,99	0,98	0,975	0,95	0,9
12	2,214	3,074	3,571	4,178	4,404	5,226	6,304
13	2,617	3,565	4,107	4,765	5,009	5,892	7,042
14	3,041	4,075	4,660	5,368	5,629	6,57	7,790
15	3,483	4,601	5,229	5,985	6,262	7,261	8,547
16	3,942	5,142	5,812	6,614	6,908	7,962	9,312
17	4,416	5,697	6,408	7,255	7,654	8,672	10,085
18	4,905	6,265	7,015	7,906	8,231	9,390	10,865
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651
20	5,921	7,434	8,620	9,237	9,591	10,871	12,433
21	6,447	8,034	8,897	9,915	10,283	11,591	13,240
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848
24	8,035	9,886	10,856	11,992	12,401	13,848	15,659
25	8,649	10,520	11,524	12,697	13,120	14,611	16,173
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114
28	10,391	12,461	13,565	14,847	15,308	16,928	18,937
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768
30	11,588	13,787	14,953	16,306	16,791	18,493	20,599

k	Рівень значущості α						
	0,80	0,75	0,70	0,50	0,30	0,25	0,20
1	0,0642	0,102	0,148	0,455	1,074	1,323	1,642
2	0,446	0,575	0,713	1,386	2,408	2,773	3,219
3	1,005	1,213	1,424	2,366	3,665	4,108	4,642
4	1,649	1,923	2,195	3,357	4,878	5,385	5,989
5	2,343	2,675	3,000	4,351	6,064	6,626	7,289
6	3,070	3,455	3,828	5,348	7,231	7,81	8,558
7	3,822	4,255	4,671	6,346	8,383	9,037	9,803
8	4,594	5,071	5,527	7,344	9,524	10,219	11,030
9	5,380	5,899	6,393	8,343	10,656	11,389	12,242
10	6,179	6,737	7,267	9,342	11,781	12,549	13,412
11	6,989	7,58	8,148	10,341	12,899	13,701	14,631
12	7,807	8,438	9,034	11,340	14,011	14,845	15,812

Продовження Таблиці А.5

k	Рівень значущості α						
	0,80	0,75	0,70	0,50	0,30	0,25	0,20
13	8,634	9,299	9,926	12,340	15,119	15,984	16,985
14	9,467	10,165	10,821	13,339	16,222	17,117	18,151
15	10,307	11,036	11,721	14,339	17,322	18,245	19,311
16	11,152	11,912	12,624	15,338	18,418	19,369	20,465
17	12,002	12,892	13,531	16,338	19,511	20,489	21,615
18	12,857	13,675	14,440	17,338	20,601	21,605	22,760
19	13,716	14,562	15,352	18,338	21,689	22,718	23,900
20	14,578	15,452	16,266	19,337	22,775	23,828	25,038
21	15,445	16,344	17,182	20,337	23,858	24,935	26,171
22	16,314	17,240	18,101	21,337	24,939	26,039	27,301
23	17,187	18,137	19,021	22,337	26,018	27,141	28,429
24	18,062	19,037	19,943	23,337	27,096	28,241	29,553
25	18,940	19,939	20,887	24,336	28,172	29,339	30,675
26	19,820	20,843	21,792	25,336	29,246	30,434	31,795
27	20,703	21,749	22,719	26,136	30,319	31,528	32,912
28	21,588	22,657	23,617	27,336	31,391	32,620	34,027
29	22,475	23,567	24,577	28,336	32,461	33,711	35,139
30	23,364	24,478	25,508	28,336	33,530	34,800	36,250

k	Рівень значущості α						
	0,10	0,05	0,025	0,02	0,01	0,005	0,001
1	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	6,251	7,815	9,348	9,837	11,345	12,838	16,268
4	7,779	9,488	11,143	11,668	13,277	14,860	18,465
5	9,236	11,070	12,839	13,388	15,086	16,750	20,517
6	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	12,017	14,067	16,013	16,622	18,475	20,278	24,322
8	13,362	15,507	17,535	18,168	20,090	21,955	26,125
9	14,648	16,919	19,023	19,679	21,666	23,589	27,877
10	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	18,549	21,026	23,337	24,054	26,217	28,300	32,909

Продовження Таблиці А.5

k	Рівень значущості α						
	0,10	0,05	0,025	0,02	0,01	0,005	0,001
13	19,812	22,362	24,736	25,472	27,688	29,819	34,528
14	21,064	23,685	26,119	26,873	29,141	31,319	36,123
15	22,307	24,996	27,488	28,259	30,578	32,801	37,697
16	23,542	26,296	28,845	29,633	32,000	34,267	39,252
17	24,769	27,587	30,191	30,995	33,409	35,718	40,790
18	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	27,204	30,144	32,852	33,678	36,191	38,582	43,820
20	28,412	31,140	34,170	35,020	37,566	39,997	45,315
21	29,615	32,671	35,479	36,343	38,932	41,401	46,797
22	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	32,007	35,172	38,076	38,968	41,638	44,181	49,768
24	33,196	36,415	39,364	40,270	42,980	45,558	51,170
25	34,382	37,652	40,046	41,566	44,314	46,928	52,620
26	35,563	38,885	41,923	42,856	45,642	48,290	54,052
27	36,741	40,113	43,194	44,140	46,963	49,645	55,476
28	37,916	41,337	44,461	45,419	48,278	50,993	56,893
29	39,087	42,557	45,722	46,693	49,588	52,336	58,302
30	40,256	43,773	46,979	47,962	50,892	53,672	59,703

Таблиця А.6 – Довірчий інтервал для σ: $\gamma_1 \cdot s < \sigma < \gamma_2 \cdot s$, де

$\gamma_1 = \sqrt{\dfrac{n-1}{\chi^2_{\alpha/2;n-1}}}$ – нижня межа і $\gamma_2 = \sqrt{\dfrac{n-1}{\chi^2_{1-\alpha/2;n-1}}}$ – верхня межа,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^2}\ ,\ k = n-1$$

γ / k	0,99		0,98		0,95		0,90	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
1	0,356	159	0,388	79,8	0,446	31,9	0,510	15,9
2	0,434	14,1	0,466	9,97	0,521	6,28	0,578	4,40
3	0,483	6,47	0,514	5,11	0,566	3,73	0,620	2,92

Продовження Таблиці А.6

γ / k	0,99		0,98		0,95		0,90	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
4	0,519	4,39	0,549	3,67	0,599	2,87	0,649	2,37
5	0,546	3,48	0,576	3,00	0,624	2,45	0,672	2,090
6	0,569	2,98	0,597	2,62	0,644	2,202	0,690	1,916
7	0,588	2,66	0,616	2,377	0,661	2,035	0,705	1,797
8	0,604	2,440	0,631	2,205	0,675	1,916	0,718	1,711
9	0,618	2,277	0,644	2,076	0,688	1,826	0,729	1,645
10	0,630	2,154	0,656	1,977	0,699	1,755	0,739	1,593
11	0,641	2,056	0,667	1,898	0,708	1,698	0,748	1,550
12	0,651	1,976	0,677	1,833	0,717	1,651	0,755	1,515
13	0,660	1,910	0,685	1,779	0,725	1,611	0,762	1,485
14	0,669	1,854	0,693	1,733	0,732	1,577	0,769	1,460
15	0,676	1,806	0,700	1,694	0,739	1,548	0,775	1,437
16	0,683	1,764	0,707	1,659	0,745	1,522	0,780	1,418
17	0,690	1,727	0,713	1,629	0,750	1,499	0,785	1,400
18	0,696	1,695	0,719	1,602	0,756	1,479	0,790	1,385
19	0,702	1,666	0,725	1,578	0,760	1,460	0,794	1,370
20	0,707	1,640	0,730	1,556	0,765	1,444	0,798	1,358
21	0,712	1,617	0,734	1,536	0,769	1,429	0,802	1,346
22	0,717	1,595	0,739	1,519	0,773	1,416	0,805	1,335
23	0,722	1,576	0,743	1,502	0,777	1,402	0,809	1,326
24	0,726	1,558	0,747	1,487	0,781	1,391	0,812	1,316
25	0,730	1,541	0,751	1,473	0,784	1,380	0,815	1,308
26	0,734	1,526	0,755	1,460	0,788	1,371	0,818	1,300
27	0,737	1,512	0,758	1,448	0,791	1,361	0,820	1,293
28	0,741	1,499	0,762	1,436	0,794	1,352	0,823	1,286
29	0,744	1,487	0,765	1,426	0,796	1,344	0,825	1,279
30	0,748	1,475	0,768	1,417	0,799	1,337	0,828	1,274
40	0,774	1,390	0,792	1,344	0,821	1,279	0,847	1,228
50	0,793	1,336	0,810	1,297	0,837	1,243	0,861	1,199
60	0,808	1,299	0,824	1,265	0,849	1,217	0,871	1,179
70	0,820	1,272	0,835	1,241	0,858	1,198	0,879	1,163
80	0,829	1,250	0,844	1,222	0,866	1,183	0,886	1,151

Продовження Таблиці А.6

γ / k	0,99		0,98		0,95		0,90	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
90	0,838	1,233	0,852	1,207	0,873	1,171	0,892	1,141
100	0,845	1,219	0,858	1,195	0,878	1,161	0,897	1,133
200	0,887	1,15	0,897	1,13	0,912	1,11	0,925	1,09

Таблиця А.7 – 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 А.8 – Таблиця критичних значень $D_{\alpha,n}$ критерію Колмогорова

n – обсяг вибірки, α – рівень значущості

n / α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377

Продовження Таблиці А.8

n \ α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
> 50	$\frac{1,94947}{\sqrt{n}}$	$\frac{1,62762}{\sqrt{n}}$	$\frac{1,51743}{\sqrt{n}}$	$\frac{1,34810}{\sqrt{n}}$	$\frac{1,22385}{\sqrt{n}}$	$\frac{1,13795}{\sqrt{n}}$	$\frac{1,07275}{\sqrt{n}}$

Таблиця А.9 – Критичні значення λ_α розподілу Колмогорова

Рівень значущості α	0,2	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 А.10 – Критичні значення D критерію Ліллієфорсу
n – обсяг вибірки, α – рівень значущості

n \ α	0,10	0,05	0,01
4	0.352	0.381	0.417
5	0.315	0.337	0.405
10	0.239	0.258	0.294
15	0.201	0.220	0.257
20	0.174	0.190	0.231
25	0.158	0.173	0.200
30	0.144	0.161	0.187
$n>30$	$\frac{0,805}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$

Таблиця А.11 – Критичні точки розподілу Фішера-Снедекора (F- розподіл)

перше значення відповідає рівню значущості α=0,05, друге – рівню значущості α=0,01; k_1 – число степенів свободи чисельника (більша дисперсія), k_2 – число степенів свободи знаменника (менша дисперсія).

k_2	k_1									
	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
	4052	4999	5403	5625	5764	5859	5928	5981	6023	5056
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40
	98,49	99,01	00,17	99,25	99,30	99,33	99,36	99,36	99,39	99,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,84	8,81	8,79
	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,27	10,16	10,05

Продовження Таблиці А.11

k_2	k_1									
	1	2	3	4	5	6	7	8	9	10
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
	11,26	8,65	7,59	7,10	6,63	6,37	6,18	6,03	5,91	5,81
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
	9,65	7,20	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	4,75	3,88	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,83	2,77	2,71	2,67
	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94
15	4,45	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	4,41	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
	8,28	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43

Продовження Таблиці А.11

k_2	k_1									
	1	2	3	4	5	6	7	8	9	10
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
	7,94	5,72	4,82	4,31	3,99	3,75	3,59	3,45	3,35	3,26
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,38	2,32	2,27
	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	4,24	3,38	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20
	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	4,19	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19
	7,64	5,54	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18
	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	4,08	3,23	2,84	2,61	2,45	3,33	2,25	2,18	2,12	2,08
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47

Продовження Таблиці А.11

∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83
	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

k_2	k_1								
	12	15	20	24	30	40	60	120	∞
1	243,91	245,95	248,01	249,05	250,09	251,14	252,20	253,25	254,32
	6106	6157	6209	6234	6261	6287	6313	6559	6366
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
	27,05	26,87	26,69	26,60	26,51	26,41	26,32	26,22	26,12
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	14,37	14,20	14,92	13,93	13,84	13,75	13,65	13,56	13,46
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,99
	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	2,69	2,62	2,54	2,50	2,47	2,43	2,38	2,34	2,30
	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
	3,96	3,82	3,66	3,59	3,59	3,51	3,43	3,34	3,16
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87

Продовження Таблиці А.11

k_2	k_1								
	12	15	20	24	30	40	60	120	∞
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
	3,45	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
	3,37	3,23	3,08	3,01	2,92	2,84	2,75	2,66	2,57
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,82
	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,30
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
	2,96	2,82	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,00	1,92	1,94	1,79	1,74	1,69	1,64	1,58	1,51
	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80

Продовження Таблиці А.11

k_2	k_1								
	12	15	20	24	30	40	60	120	∞
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,04

Таблиця А.12 – Критичні значення коефіцієнта рангової кореляції Спірмена $\rho_{кр}$

n	$\rho_{кр}$		n	$\rho_{кр}$		n	$\rho_{кр}$	
	α=0,05	α=0,01		α=0,05	α=0,01		α=0,05	α=0,01
5	0,94	–	17	0,48	0,62	29	0,37	0,48
6	0,85	–	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,39	0,51	37	0,33	0,43
14	0,54	0,68	26	0,39	0,50	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

Таблиця А.13 – Критичні значення R^2 для рівня значущості α, числа змінних X та кількості дослідів m

	α=0,1			α=0,05			α=0,01		
X \ m	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
3	0,976			0,994			1,000		
4	0,810	0,990		0,902	0,997		0,980	1,000	

Продовження Таблиці А.13

	α=0,1			α=0,05			α=0,01		
X / m	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
5	0,649	0,900	0,994	0,771	0,950	0,998	0,919	0,990	1,000
6	0,532	0,785	0,932	0,658	0,864	0,966	0,841	0,954	0,993
7	0,448	0,684	0,844	0,569	0,776	0,903	0,765	0,900	0,967
8	0,386	0,602	0,759	0,499	0,698	0,832	0,696	0,842	0,926
9	0,339	0,536	0,685	0,444	0,632	0,764	0,636	0,785	0,879
10	0,302	0,482	0,622	0,399	0,575	0,704	0,585	0,732	0,830
11	0,272	0,438	0,568	0,362	0,527	0,651	0,540	0,684	0,784
12	0,247	0,401	0,523	0,332	0,486	0,604	0,501	0,641	0,740
13	0,227	0,369	0,484	0,306	0,451	0,563	0,467	0,602	0,700
14	0,209	0,342	0,450	0,283	0,420	0,527	0,437	0,567	0,663
15	0,194	0,319	0,420	0,264	0,393	0,495	0,411	0,536	0,629
16	0,181	0,298	0,394	0,247	0,369	0,466	0,388	0,508	0,598
18	0,160	0,264	0,351	0,219	0,329	0,417	0,348	0,459	0,544
20	0,143	0,237	0,316	0,197	0,297	0,378	0,315	0,418	0,498
22	0,129	0,215	0,287	0,179	0,270	0,345	0,288	0,384	0,459
24	0,118	0,197	0,263	0,164	0,248	0,317	0,265	0,355	0,426
26	0,109	0,181	0,243	0,151	0,229	0,294	0,246	0,330	0,396
28	0,101	0,168	0,2225	0,140	0,213	0,273	0,229	0,308	0,371
30	0,094	0,157	0,210	0,130	0,199	0,256	0,214	0,289	0,349

Навчальне видання

КИЛИМНИК Ірина Михайлівна

ПРАКТИКУМ З ЕЛЕМЕНТІВ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Навчальний посібник

Комп'ютерний набір: *Килимник І.М.*
Комп'ютерна верстка: *Дяченко О.О.*

Підписано до друку 27.05.2026. Формат 60×84/16. Ум. друк. арк. 14,9.
Тираж 100 прим. Зам. № 385.

Національний університет «Запорізька політехніка»
Україна, 69063, м. Запоріжжя, вул. Університетська, 64
Тел.: (061) 769–82–96, 220–12–14

Свідоцтво суб'єкта видавничої справи ДК № 6952 від 22.10.2019.