МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ ЕКОНОМІКИ І ТЕХНОЛОГІЙ

С.В.ТКАЛІЧЕНКО

А.А.СУПРУН

СТАТИСТИКА

НАВЧАЛЬНО-МЕТОДИЧНИЙ ПОСІБНИК

КРИВИЙ РІГ 2023

УДК 311(075.8)
ББК 60.6я73

Рецензенти:

П.В. Мерзликін, к.ф-м.н, доцент катедри інформатики та прикладної математики, Криворізький державний педагогічний університет

І.В. Довгаль, к.е.н., доц., Директор представництва ПрАТ ВНЗ МАУП "Центр дистанційного навчання "Криворізький інститут""

Д.А. Медведєв, к.т.н., доцент кафедри інформатики та прикладного програмного забезпечення, Державний університет економіки і технологій

Рекомендовано до друку Вченою радою Державного університету економіки і технологій протокол № 6 від 17.01.2023р.

С.В.Ткаліченко. А.А.Супрун Статистика: Навчально-методичний посібник. –Кривий Ріг: Державний університет економіки і технологій, 2023.–160 с.

Розкриваються основи математичної статистики: предмет, методи, базові категорії, показники тенденцій і мінливості сукупностей, статистичне оцінювання, перевірка статистичних гіпотез з використанням параметричних і непараметричних критеріїв.

Розглянуто базові питання й положення організації та проведення статистичного дослідження соціально-економічних явищ і процесів, їх кореляційний, регресійний, дисперсійний аналіз.

ЗМІСТ

ВСТУП

Розвиток сучасної науки характеризується її математизацією, що виражається у використанні математичних методів і моделей не тільки у технічних та економічних дослідженнях, але й у менеджменті, соціології, педагогіці, біології, медицині.

В соціальних науках використовуються різноманітні статистичні методи для перевірки висунутих гіпотез, побудови статистичних моделей соціальних та економічних об'єктів, явищ, закономірностей і процесів, тому навчальна дисципліна «Статистика» посідає значне місце у підготовці висококваліфікованих професіоналів. Змістом навчальної дисципліни є методи кількісного аналізу; методи моделювання та аналізу впливу певних факторів на становище об'єкта або явища; способи обробки результатів соціологічного дослідження та експертного оцінювання; методи оцінки і прогнозування можливих соціальних явищ.

У першій частині посібника надано основні поняття з математичної статистики і стандартні прийоми перевірки статистичних гіпотез, які необхідні для засвоєння навчальної дисципліни.

У другій частині посібника розглянуто методи аналізу соціально-економічних явищ, кореляційного і регресійного аналізів, що є основним інструментом обробки результатів прикладних досліджень.

ЧАСТИНА1. МАТЕМАТИЧНА СТАТИСТИКА

1.1. Поняття математичної статистики

Основні поняття та визначення.

Математична статистика – розділ математики, що розробляє методи реєстрації, опису та аналізу даних спостережень та експериментів з метою побудови ймовірнісних моделей масових випадкових явищ.

Статистичний опис застосовують до таких фізичних процесів, для яких результат окремого виміру не може бути передбачений з необхідною точністю. Тим не менш, при проведенні достатньо великої кількості повторних вимірювань може бути з достатньо гарною точністю передбачена деяка величина, яка є функцією результатів вимірювань.

При побудові моделей у математичній статистиці передбачають імовірнісну природу явищ, що спостерігаються, і використовується математичний апарат теорії ймовірностей.

Хоча математична статистика спирається на методи та поняття теорії ймовірностей, але можна сказати, що в якомусь сенсі математична статистика вирішує обернені завдання.

Так, в теорії ймовірностей вважають, що імовірнісна модель явища задана, і на основі цієї моделі розраховують ймовірності подій, що нас цікавлять.

У математичній статистиці вважають, що імовірнісна модель явища невідома, і чинять так. Допустимо, що в результаті проведених експериментів (або спостережень) отримано деякі експериментальні дані (статистичні дані). З цих даних вибирають відповідну їм ймовірнісну модель. І потім вже використовують

отриману модель для опису явища, що розглядається, або процесу.

Наведемо приклад постановки завдання теорії ймовірностей і у математичної статистиці. Постановка завдання теорії ймовірностей формулюється так. Імовірність випадання цифри «шість» при підкиданні грального кістки відома і дорівнює деякому значенню p. Визначити ймовірність того, що при n підкиданнях кістки цифра «шість» випаде k разів. Тут $0 < k < n$. Типова постановка у завдання математичної статистики звучить так. Гральна кістка підкидається n разів. Цифра «шість» випала p разів. Яка ймовірність випадання цифри "шість" при одному підкиданні.

Генеральна та вибіркова сукупності.

Поняття генеральної сукупності та вибірки з неї (вибіркової сукупності) є основними поняттями математичної статистики. У теорії ймовірностей було введено поняття випадкової величини X, яка може приймати різні значення $x_1, x_2, \dots, x_n$. У математичній статистиці кажуть, що множина всіх значень випадкової величини X - це генеральна сукупність. Запроваджують також поняття випадкової вибірки. Під випадковою вибіркою розуміють деякі випадкові величини, такі, що розподіл збігається з розподілом випадкової величини X.

Розглядаючи завдання теорії ймовірностей, кажуть, що в результаті n незалежних випробувань (дослідів, спостережень) випадкова величина X прийняла значення $x_1, x_2, \dots, x_n$. Аналогічна ситуація в математичній статистиці описується так: з генеральної сукупності X отримано випадкову вибірку $X_1, X_2, \dots, X_n$.

Перш ніж вводити суворі визначення генеральної та вибіркової сукупностей, наведемо приклад, що ілюструє ці два основні поняття математичної статистики. Припустимо, деякі вимірювальні прилади випускаються певним підприємством. Нас цікавить деяка кількісна характеристика якості роботи приладу, наприклад, точність вимірювань, що забезпечується цим приладом, або час безвідмовної роботи приладу.

Очевидно, здійснення контролю параметра, що цікавить нас, може бути проведене не для всієї випущеної продукції (генеральної сукупності), а для деяких приладів (випадкової вибірки, вибіркової сукупності). Іншими прикладами генеральної сукупності є:

-всі мешканці Києва,

-усі студенти України,

-всі юридичні особи будь-якої країни,

-всі підприємства, розташовані біля певного міста, які здійснюють торгівлю книжковою продукцією.

Вибіркова сукупність складається з частини об'єктів, генеральної сукупності, відібраних для вивчення, метою якого є встановлення певних характеристик генеральної сукупності. Таким чином, прикладами вибіркових сукупностей можуть бути:

-частина населення Києва,

-деяка частина студентів України,

-деякі юридичні особи певної країни і т.д.

Щоб висновки, отримані щодо вибіркової сукупності, можна було поширити на генеральну сукупність, вибірка повинна коректна, відображати генеральну сукупність (кажуть, вибірка має бути репрезентативною). Так, вибірка, що складається з мешканців Києва, які є власниками двох автомобілів, не

репрезентує все населення міста (або не репрезентує купівельну спроможність мешканців Києва).

Розглянемо деякі визначення, прийняті у математичній статистиці.

Уся множина значень випадкової величини X називається **генеральною сукупністю** випадкової величини X. Тобто генеральна сукупність – це випадкова величина X, задана на просторі елементарних подій Ω.

Законом розподілу генеральної сукупності X називається закон розподілу випадкової величини X.

Властивості генеральної сукупності вивчають на підставі аналізу статистичних (експериментальних) даних, під якими розуміють значення випадкової величини, одержані в результаті повторень випадкового експерименту (спостережень над випадковою величиною). При цьому передбачається, що експеримент може бути проведений скільки завгодно багато разів в тих самих умовах. «Експеримент, проведений в тих самих умовах» означає, що розподіл випадкової величини Xi, i= 1, 2, ..., отриманої в і-тому експерименті, не залежить від номера випробування і збігається з розподілом генеральної сукупності X. Тобто спостереження над випадковою величиною проводяться у незалежних повторних експериментах.

Сукупність незалежних випадкових величин $X_1, X_2, \dots, X_n$, кожна з яких має той самий розподіл, що і випадкова величина X називається випадковою вибіркою з генеральної сукупності X і позначається як

$$\overrightarrow{X_n} = X_1, X_2, \dots, X_n \qquad (1.1)$$

Випадкові величини X_i називаються **елементами** випадкової вибірки, а n називають **обсягом** випадкової вибірки.

Реалізацією випадкової вибірки (або вибіркою) із генеральної сукупності X називають будь-яке можливе значення випадкової вибірки $\overrightarrow{X_n}$ та позначають

$$\overrightarrow{x_n} = x_1, x_2, \dots, x_n \quad (1.2)$$

Тут числа x_i називають елементами вибірки $\overrightarrow{x_n}$, або **варіантами**.

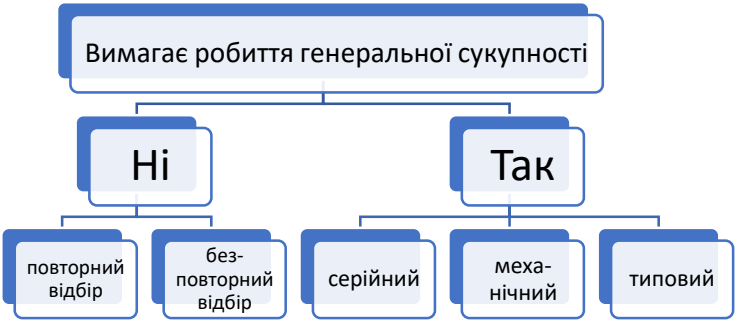
Вибірку $\overrightarrow{x_n}$ можна інтерпретувати як сукупність n чисел $x_1, x_2, \dots, x_n$, отриманих в результаті проведення n повторних незалежних спостережень (випробувань) над випадковою величиною X.

Основою будь-яких висновків про ймовірнісні властивості генеральної сукупності X, тобто статистичних висновків, є так званий **вибірковий метод**. Суть вибіркового методу полягає в тому, що властивості випадкової величини X встановлюються на підставі вивчення випадкової вибірки.

Вибірка має бути **репрезентативною** (представницькою).

При використанні у статистичних дослідженнях вибіркового методу виникають так звані помилки репрезентативності. Помилки репрезентативності обумовлені тим, що вибіркова сукупність не повністю відтворює генеральну і є розбіжністю між значеннями параметрів, отриманими з вибірки, та значеннями відповідних параметрів генеральної сукупності.

Існує кілька способів відбору випадкових величин для отримання вибіркової сукупності:



Вибірка називається **повторною**, якщо об'єкт повертається в ґенеральну сукупність перед виконанням наступного відбору. В іншому випадку вона безповторна.

Типовий відбір – це відбір, у якому неоднорідна генеральна сукупність розбивається на типові (однорідні) групи, з яких у тому числі й проводиться випадковий відбір. Так, наприклад, якщо продукція виготовляється на кількох верстатах, то для проведення контролю якості природно проводити вибірку продукції, виготовленої на кожному зі верстатів.

Механічний відбір здійснюється через певний інтервал, наприклад, вибирають кожну з 100 деталей.

При **серійному** відборі – відбирають не окрему одиницю (наприклад, вироблену деталь чи прилад), а групу, серію, наприклад, продукцію, виготовлену одним верстатом.

На практиці часто застосовують комбінований відбір.

Множину можливих значень випадкової вибірки $\overrightarrow{X_n}$ називають **вибірковим простором** χ_n

Будь-яку функцію випадкової вибірки $g\left(\overrightarrow{X_n}\right)$ називають **статистикою**, або вибірковою характеристикою. Значення

статистики $g\left(\overrightarrow{\mathrm{X}_n}\right)$ отримане при реалізації $\overrightarrow{x_n}$ випадкової вибірки $\overrightarrow{\mathrm{X}_n}$, називається її вибірковим значенням і позначається $-g\left(\overrightarrow{\mathrm{X}_n}\right)$.

Характерною рисою завдань у математичній статистиці є наявність деякої апріорної інформації.

До проведення випробувань про випадкову величину може бути відомо дуже мало, наприклад, лише те, що вона є дискретною або неперервною. З іншого боку, на практиці також зустрічаються завдання, коли відомий розподіл випадкової величини з точністю до параметра. Наприклад, генеральна сукупність має нормальний розподіл з невідомими параметрами μ (математичне сподівання) та σ (середнє квадратичне відхилення).

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.3}$$

Оскільки параметри σ і μ – невідомі, ми можемо говорити тільки про сімейство (клас) розподілів.

Вибірковий простір, на якому задано клас розподілів, називається **статистичною моделлю**. Статистична модель повністю визначається функцією розподілу. Позначатимемо статистичну модель {F(x)}, оскільки вона повністю визначена функцією розподілу F(x) генеральної сукупності.

Якщо функція розподілу задана з точністю до невідомого параметра (загалом вектор параметрів з множиною $\overrightarrow{\Theta_n} = \Theta_1, \Theta_2, \dots, \Theta_n$ то статистичну модель називають **параметричною**

$$\{\mathrm{F}(x, \vec{\Theta}), \vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_{n)}\}.$$

Статистичну модель називають **неперервною**, або **дискретною**, залежно від того, чи є випадкова величина X неперервною або дискретною. У разі неперервної статистичної

моделі розподіл задається **щільністю розподілу**. Для статистичних моделей можна використовувати також позначення $\{p(x,\vec{\Theta}),\vec{\Theta}=(\Theta_1,\Theta_2,\dots,\Theta_n)\}$.

Наведемо приклад завдання параметричної статистичної моделі. Нехай відомо, що генеральна сукупність випадкової величини X розподілена за нормальним законом із відомою дисперсією та невідомим значенням математичного сподівання Θ. Тоді статистична параметрична модель {F(x; θ)} може бути задана за допомогою щільності розподілу

$$p(x,\Theta)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\Theta)^2}{2\sigma^2}}$$

Якщо в розподілі невідомі обидва параметри - математичне сподівання і середнє квадратичне відхилення, то статистична модель набуде вигляду $\{F(x,\vec{\Theta}),\vec{\Theta}=\Theta_1,\Theta_2\}$, де щільність розподілу містить два невідомі параметри.

$$p(x,\Theta_1,\Theta_2)=\frac{1}{\Theta_2\sqrt{2\pi}}e^{-\frac{(x-\Theta_1)^2}{2\Theta_2{}^2}}$$

Все вищевикладене можна узагальнити також на багатовимірні випадкові величини.

Якщо використовуються закони розподілу випадкових величин, які мають загальноприйняті назви (нормальний розподіл, біноміальний розподіл), то й статистичні моделі називають відповідним чином: нормальна модель, біномна модель.

Основні завдання математичної статистики.

Перш за все, слід зазначити, що при вирішенні будь-якого завдання математичної статистики, у розпорядженні є два джерела інформації: результати статистичного експерименту

(тобто вибірка з генеральної сукупності випадкової величини X) і деяка апріорна інформація про властивості генеральної сукупності, що нас цікавлять, та відома на поточний момент. Апріорна інформація враховується у вибраній статистичній моделі.

Перерахуємо деякі завдання математичної статистики, що найчастіше зустрічаються:

1. Оцінка невідомих параметрів.

2. Перевірка статистичних гіпотез.

3. Встановлення форми та ступеня зв'язку між випадковими величинами.

Завдання оцінки **невідомих параметрів** формулюється так.

Функція розподілу відома з точністю до θ. Тобто задана параметрична статистична модель $\{F(x,\vec{\Theta}),\vec{\Theta}=(\Theta_1,\Theta_2,\dots,\Theta_n)\}$.

Необхідно знайти таку статистику $\widehat{\Theta}(\overrightarrow{X_n})$, вибіркове значення якої $\widehat{\Theta}(\overrightarrow{x_n})$ для поточної реалізації $\overrightarrow{x_n}$ випадкової вибірки можна було б вважати наближеним значенням параметра θ. Статистику $\widehat{\Theta}(\overrightarrow{X_n})$, вибіркове значення якої $\widehat{\Theta}(\overrightarrow{x_n}) \approx \theta$ для випадкової вибірки яка реалізується $\overrightarrow{x_n}$ випадкової вибірки, називають його **точковою** оцінкою, або просто оцінкою, а $\widehat{\Theta}(\overrightarrow{x_n})$ – значенням точкової оцінки. Однак можливе й інше формулювання поставленого завдання.

Нехай функція розподілу відома з точністю до параметра θ. Тобто задана $\{F(x,\vec{\Theta}),\vec{\Theta}=(\Theta_1,\Theta_2,\dots,\Theta_{n)}\}$ параметрична статистична модель. Необхідно знайти такі статистики та $\underline{\Theta}(\overrightarrow{X_n})$, щоб з ймовірністю γ виконувалася нерівність

$$P\left\{\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})\right\} = \gamma \qquad (1.4)$$

Говорять про **інтервальну оцінку** для оцінювання шуканого параметра θ.

Інтервал $\left\{\underline{\Theta}\left(\overrightarrow{X_n}\right),\overline{\Theta}\left(\overrightarrow{X_n}\right)\right\}$ називають довірчим інтервалом для параметра θ. Величину γ називають **довірчою ймовірністю**.

Перевірка статистичних гіпотез. Статистичною гіпотезою називають будь-яке припущення про розподіл ймовірностей випадкової величини. Йдеться або про вид розподілу (непараметрична гіпотеза), або про значення параметрів розподілу (параметрична гіпотеза).

В останньому випадку можна розглядати задачу перевірки статистичної гіпотези як обернену до завдання оцінки параметрів. Справді, під час вирішення завдання оцінки параметра справжнє його значення невідоме. Під час перевірки статистичної гіпотези ми на підставі апріорної інформації припускаємо відомим його значення та за результатами експерименту перевіряємо висунуте припущення.

Наведемо деякі приклади статистичних гіпотез.

1. Гіпотези про величину математичного сподівання (параметрична гіпотеза): $\mu = \mu_0,\ \mu = 5,\ \mu = -0{,}01$

де μ – математичне сподівання випадкової величини.

2. Гіпотеза про рівність дисперсій двох генеральних сукупностей (параметрична гіпотеза): $\sigma_1^2 = \sigma_2^2$, де σ_1^2 і σ_2^2 - дисперсії двох випадкових величин X_1 і X_2.

3. Гіпотеза про вид функції розподілу (непараметрична гіпотеза): $F(x) = F_t(x)$, де F(x) – функція розподілу випадкової величини, $F_t(x)$ – деяка шукана функція розподілу.

Встановлення форми та ступеня зв'язку між випадковими величинами. За розв'язання низки практичних завдань часто виникає необхідність встановлення залежностей

між величинами. Для пояснення такого типу завдань наведемо деякі приклади.

Нехай випадкова величина Y_1 – кількісна характеристика продуктивності будь-якої хімічної установки (хімічної реактора), випадкова величина Y_2 - описує частку браку в випущеній продукції. Припустимо, що X_1, X_2, і X_3 - випадкові величини, що характеризують технологічні параметри, наприклад, вміст домішок в сировині, вологість газової атмосфери в реакторі, температура в реакторі. Природно припустити, що технологічні параметри впливають на величини Y_i. Для оптимізації виробництва необхідно встановити ступінь впливу різних технологічних факторів X_i на величини Y_i. Після знаходження залежності Y_i від X_i можна встановити, якими повинні бути технологічні параметри для мінімізації кількості браку при заданому значенні продуктивності реактора.

Припустимо, що Y - ступінь зносу деякої конструкційної деталі в хімічному реакторі, яка використовується для очищення сировини.

X_1, X_2, і X_3 – склад матеріалів, з яких ця деталь може бути виготовлена. Z_1, Z_2, і Z_3– хімічний склад реактивних сумішей, які використовуються у технологічному процесі (наразі в очищенні). Якщо встановити ступінь впливу величин X_i та Y_i на величину X, то можна збільшити час "безперебійної" роботи реактора.

Попередня обробка результатів виміру

Використання методів математичної статистики для рішення прикладних завдань пов'язане з обробкою великої кількості даних. Тому статистичні дані, отримані в результаті вимірювань (статистичних експериментів, спостережень), повинні пройти попередню обробку, що полегшує їх аналіз. Одна з найпростіших

процедур попередньої обробки статистичних даних – це впорядкування їх за величиною.

Припустимо, що в результаті експерименту отримано вибірку $x_1, x_2, \dots, x_n$ обсягом n з генеральної сукупності X. Упорядкуємо одержану вибірку, розташувавши її елементи в неспадаючому порядку:

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(i)} \le \cdots \le x_{(n)} \qquad (1.5)$$

$x_{(1)} -$ найменший елемент вибірки, $x_{(n)} -$ найбільший.

Послідовність чисел $x_1, x_2, \dots, x_n$ що задовольняють умові (1.5) називають **варіаційним рядом** вибірки. Числа $x_{(i)}$, $i = \overline{1-n}$ називають членами варіаційного ряду. Операція впорядкування значень вибірки називається **ранжуванням** статистичних даних.

Слід зауважити, що перехід від випадкової вибірки до варіаційного ряду не призводить до втрати інформації, проте функції розподілу випадкових величин не збігаються з функцією розподілу генеральної сукупності. Серед елементів вибірки можуть бути величини, що повторюються.

Статистичним рядом для вибірки називають таблицю, яка містить елементи вибірки $x_{(i)}$ і числа n_i їх повторень (Табл. 1.1). При заданні статистичного ряду можна вказувати не тільки числа n_i, а й відношення n_i/n, де n - обсяг вибірки. При цьому числа n_i називають частотами, а відношення n_i/n – відносними частотами елемента вибірки.

Таблиця 1.1.

Статистичний ряд

$x_{(i)}$	$x_{(1)}$	⋯	$x_{(i)}$	⋯	$x_{(m)}$
n_i	n_1	⋯	n_i	⋯	n_m
n_i/n	n_1/n	⋯	n_i/n	⋯	n_m/n

Статистичні дані, подані у вигляді статистичного ряду, називаються **групованими**.

Дані можна групувати не тільки у вигляді статистичного ряду, а й у вигляді **інтервального статистичного ряду**.

Для побудови статичного інтервального ряду всі вибіркові значення розбивають на кілька інтервалів. Як правило, інтервали мають рівну довжину. Для визначення кількості інтервалів можна використати наступну оціночну формулу (формула Стерджеса)

$$m = log_2 n + 1 = 3{,}32 \ln n + 1 \qquad (1.6)$$

Кожен інтервал містить певну кількість елементів вибірки. Так, інтервал $J_j = [x_i, x_{i+1}]$ містить n_i елементів вибірки, значення яких задовольняють умовам $x_i \le x_j \le x_{i+1}$.

$$\Delta = x_{i+1} - x_i \qquad (1.7)$$

Називається **довжиною інтервалу**. Різниця між найбільшим та найменшим значенням вибірки називають **розмахом варіації**.

Отримані значення записують у вигляді таблиці (табл. 1.2), причому допускається у верхньому рядку таблиці вказувати або інтервал, або його середнє значення.

Таблиця 1.2

Інтервальний статистичний ряд

J_j	J_1	…	J_j	…	J_m
n_j	n_1	…	n_j	…	n_m
n_j/n	n_1/n	…	n_j/n	…	n_m/n

Наведемо приклад попередньої обробки даних. Нехай у результаті вимірів отримані наступні значення деякої фізичної величини (тобто, отримана наступна вибірка):

2; 3; 2; 4; 5; 2; 3; 4; 2; 5; 3; 4; 2; 4; 3; 4; 5; 3; 2; 3; 4; 3; 2; 3; 4; 3; 2; 3; 4;3; 3; 3; 3; 3; 3.

Об'єм вибірки n = 35. Побудуємо варіаційний ряд вибірки.

Перший (найменший) елемент варіаційного ряду $x1$=2, останній (найбільший) елемент варіаційного ряду $x4$=5. Варіаційний ряд виглядає наступним чином:
2; 2; 2; 2; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4; 4;4; 4; 4; 5; 5; 5.

Статистичний ряд отриманої вибірки наведено у табл. 1.3. Таким чином, статистичний ряд містить чотири елементи.

У табл. 1.4 представлено інтервальний статистичний ряд розподілу вибірки, отриманої при вимірі росту 500 дітей, підлітків та молодих людей віком до 20 років.

Таблиця 1.3

Статистичний ряд вибірки

$x_{(i)}$	2	3	4	5
n_i	8	16	8	3
$\frac{n_i}{n}$	$\frac{8}{35}$	$\frac{16}{35}$	$\frac{8}{35}$	$\frac{3}{35}$

Таблиця 1.4

Інтервальний статистичний ряд вибірки

J_j	$[135; 140)$	$[140; 145)$	$[145; 150)$	$[150; 155)$	$[155; 160)$	$[160; 165)$
n_j	21	27	35	88	123	91
n_j/n	$\frac{21}{500}$	$\frac{27}{500}$	$\frac{35}{500}$	$\frac{88}{500}$	$\frac{123}{500}$	$\frac{91}{500}$
J_j	$[165; 170)$	$[170; 175)$	$[175; 180)$	$[180; 185)$	$[185; 190)$	$[190; 195]$
n_j	60	43	7	4	0	1
n_j/n	$\frac{60}{500}$	$\frac{43}{500}$	$\frac{7}{500}$	$\frac{4}{500}$	0	$\frac{1}{500}$

1.2. Характеристики генеральної та вибіркової сукупностей

Емпірична та теоретична щільність розподілу. Гістограма розподілу.

Щільності розподілу також можна поставити у відповідність статистичний аналог, визначений на вибірці.

Емпіричною щільністю розподілу відповідною реалізації $\overrightarrow{x_n}$• випадкової вибірки $\overrightarrow{X_n}$ об'єму n з генеральної сукупності X, називають функцію, яка у всіх точках інтервалу $J_j = [x_i, x_{i+1}]$ дорівнює $n_j/n\Delta$, а поза інтервалом J_j дорівнює 0. Тут $\Delta = x_{i+1} - x_i$ – довжина інтервалів J_j, тобто інтервальна різниця.

$$p_n(x) = \begin{cases} \frac{n_j}{n\Delta}, & x \in J_j \\ 0, & x \notin J_j \end{cases} \qquad (2.3)$$

Якщо розглядаються вибірки великих обсягів та (або) неперервні статистичні моделі попередньо оброблені, зручно будувати інтервальний статистичний ряд (табл. 1.4). Як очевидно з табл. 1.4, у нижньому рядку представлені відносні частоти n_j/n. Якщо ми розділимо значення відносної частоти n_j/n на значення інтервальної різниці $\Delta = x_{i+1} - x_i$, то отримаємо значення щільності розподілу в даному інтервалі $J_j = [x_i, x_{i+1}]$. Введемо випадкову величину $n_i\left(\overrightarrow{X_n}\right)/n$. Ця величина для кожної реалізації $\overrightarrow{x_n}$ випадкової вибірки $\overrightarrow{X_n}$ дорівнює відносній частоті n_j/n. За законом великих чисел у формі теореми Бернуллі випадкова величина $n_i\left(\overrightarrow{X_n}\right)$/n сходиться по ймовірності попадання випадкової величини в проміжок $J_j = [x_i, x_{i+1}]$ при $n \to \infty$.

$$\frac{n_i\left(\overrightarrow{X_n}\right)}{n} \underset{n \to \infty}{\overset{P}{\to}} P(X \in [x_i, x_{i+1}]) = \int_{x_i}^{x_{i+1}} p(x)dx$$

Тут $p(x)$ – щільність розподілу генеральної сукупності X. Якщо довжина інтервалів Δ мала, то можна вважати, що

$$\frac{n_i}{n} \approx p(x)\Delta$$

де $\widetilde{x_i}$ – середина проміжку J_j.Тоді $\frac{n_i}{n\Delta} \approx p(\widetilde{x_i})$ $p_n(x) = p(x)$

і функцію $p_n(x)$ можна вважати статистичним аналогом теоретичної густини розподілу p(x).

Функція $p_n(x)$ – кусково-постійна. Графік функції $p_n(x)$ називається гістограмою (рис. 2.2).

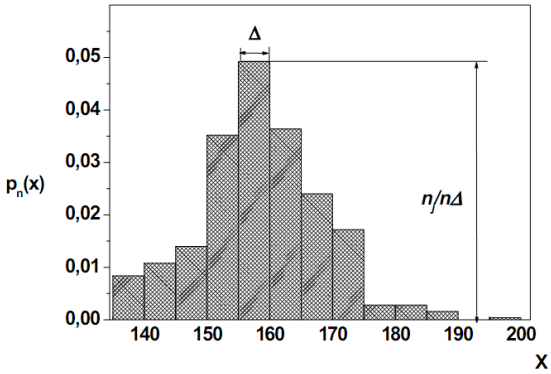


Рис. 2.2. Графік емпіричної густини розподілу

Гістограма є діаграмою, складеною з прямокутників з основою $\Delta = x_{i+1} - x_i$ та висотами n_j/n, $j = \overline{1,m}$. Сумарна площа всіх прямокутників дорівнює 1. Площа кожного прямокутника (n_j/n) – частота влучення елементів вибірки у відповідний інтервал.

Іноді замість показаної на малюнку гістограми відносних частот будують гістограму частот, відкладаючи по осі OY значення частот n_j/Δ. Поряд із гістограмою часто використовують інше графічне представлення для функції $p_n(x)$ – **полігон частот** або **полігон відносних частот**.

Полігон частот (полігон відносних частот) – ламана, відрізки якої з'єднують відкладені по осі у значення частот (відносних частот), якщо по осі OX відкладено значення елементів вибірки.

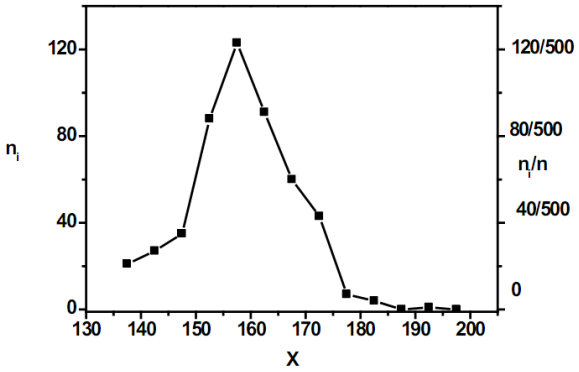


Рис. 2.3. Полігон частот (n_j) та відносних частот (n_j/n)

Вибір кількості інтервалів при потроєнні гістограм істотно залежить від обсягу даних. У літературі наводяться кілька посібників з вибору кількості інтервалів.

Наприклад, при виборі числа інтервалів можна використовувати введену в попередньому розділі формулу Стерджеса (формула (1.6)). Існують також інші методи розрахунку:

$$m = 5 \ln n \tag{2.4}$$

$$m = \sqrt{n} \tag{2.5}$$

Формули (1.6), (2.4), (2.5) слід розглядати як оцінку знизу визначення кількості інтервалів.

Вибіркові та теоретичні числові моменти

Аналогічно всім числовим характеристикам генеральної сукупності (теоретичним або генеральним) можна поставити у відповідність їх вибіркові аналоги (статистичні аналоги), визначені на вибірці.

Теоретичні початковий (m_k) та центральний ($\dot{m}_k$) моменти k-го порядку визначаються так:

$$m_k = M(X^k) \tag{2.6}$$

$$\dot{m}_k = M\left((\dot{X} - M(X)^k\right) \tag{2.7}$$

Тут $M(X)$ – математичне сподівання, яке можна знайти, знаючи щільність розподілу безперервної випадкової величини або закон розподілу дискретної випадкової величини за формулами (2.8) та (2.9) відповідно.

$$M(X) = \sum_{k=1}^{\infty} x_k p_k \tag{2.8}$$

$$M(X) = \int_{-\infty}^{+\infty} xp(x)dx \tag{2.9}$$

Визначимо вибіркові початковий $\widehat{m}_k\left(\overrightarrow{X_n}\right)$ та центральний $\widehat{\dot{m}_k}\left(\overrightarrow{X_n}\right)$моменти k-того порядку наступним чином:

$$\widehat{m}_k\left(\overrightarrow{X_n}\right) = \frac{1}{n}\sum_{i=1}^{n} X_k^i) \tag{2.10}$$

$$\widehat{\dot{m}_k}\left(\overrightarrow{X_n}\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \tag{2.11}$$

Тут $\bar{X}$ – вибіркове середнє:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ )$$

Таким чином, вибіркове середнє є вибірковим початковим моментом першого порядку:

$$\bar{X} = \widehat{m_1}\left(\overrightarrow{X_n}\right)$$

Вибіркова дисперсія – це вибірковий центральний момент 2-го порядку:

$$\widehat{\sigma_2}\left(\overrightarrow{X_n}\right) = \widehat{m_2}\left(\overrightarrow{X_n}\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.13}$$

Вибіркове середнє квадратичне відхилення визначається як

$$\widehat{\sigma}\left(\overrightarrow{X_n}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{2.14}$$

Основна властивість вибіркових моментів як початкових, так і центральних у тому, що зі збільшенням обсягу вибірки n вони

сходяться ймовірно до відповідним теоретичним (генеральним) моментам.

Відповідні величини, визначені на вибірковій сукупності (реалізації випадкової вибірки), є статистичними аналогами теоретичних величин і називаються початковим та центральним моментом k-того порядку вибірки, середнім значенням вибірки, дисперсією та середнім квадратичним відхиленням вибірки:

$$\widehat{m}_k(\overrightarrow{x_n}) = \frac{1}{n}\sum_{i=1}^{n} x_k^i \tag{2.15}$$

$$\widehat{m_k}(\overrightarrow{x_n}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k \tag{2.16}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.17}$$

$$\widehat{\sigma_2}(\overrightarrow{x_n}) = \widehat{m_2}(\overrightarrow{x_n}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.18}$$

$$\hat{\sigma}(\overrightarrow{x_n}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.19}$$

1.3. Оцінка моментів та параметрів розподілу

Види оцінок та його характеристики.

Завдання оцінки параметрів виникає під час розгляду параметричної моделі. Вважають, що закон розподілу генеральної сукупності має вигляд:

$$F(x, \overrightarrow{\Theta})$$

Тобто функція розподілу генеральної сукупності відома з точністю до параметра $\overrightarrow{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$, значення якого і необхідно оцінити за даними виміру.

У математичній статистиці розглядають два види оцінок: точкові та інтервальні.

Завдання знаходження точкової оцінки формулюється так.

Розглянемо випадкову вибірку $\overrightarrow{x_n}$ обсягу n з генеральної совкупності X, закон розподілу якої заданий з точністю до

параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Тобто задана статистична параметрична модель.

$$\{F(x, \vec{\Theta}); \vec{\Theta} \in \boldsymbol{\theta}\}$$

Необхідно знайти таку статистику $\hat{\theta}(\overrightarrow{X_n})$, вибіркове значення якої $\widehat{\Theta} = \widehat{\Theta}(\overrightarrow{X_n})$ для реалізації $(\overrightarrow{x_n})$ випадкову вибірку вважають наближеним значенням параметра $\vec{\Theta}$.

Статистику $\hat{\theta}(\overrightarrow{X_n})$, вибіркове значення $\widehat{\Theta} = \widehat{\Theta}(\overrightarrow{X_n}) \approx \vec{\Theta}$ якої для будь-якої реалізації $(\overrightarrow{x_n})$ називають **точковою оцінкою**, а вибіркове значення $\widehat{\Theta} = \widehat{\Theta}(\overrightarrow{X_n})$ - **значенням точкової оцінки**.

Завдання оцінки параметрів зводиться до знаходження невідомого параметра розподілу за результатами вимірювань (реалізації випадкової вибірки). Як оцінку вибирають деяку функцію від вимірюваних величин. Наприклад, як буде показано далі, для оцінки математичного сподівання нормального розподілу вибирається наступна статистика

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

Значення цієї статистики обчислюється за значеннями елементів вибірки x_i . Отже, точкова оцінка – це число.

Значення статистики, обчисленої щодо реалізації випадкової вибірки, може значною мірою відрізнятися від значення оцінюваного параметра. Точність точкової оцінки характеризується дисперсією.

Очевидно, що як оцінка може бути обрана будь-яка статистика. Визначаючи «оптимальну» для статистики, що оцінюється, необхідно, щоб вона задовольняла ряду умов. Властивості точкових оцінок далі будуть розглянуті докладніше. Для знаходження точкових оцінок розроблено низку методів:

метод моментів, метод максимальної правдоподібності, графічний метод, метод найменших квадратів та інші.

Завдання знаходження інтервальної оцінки формулюється наступним чином.

Розглянемо випадкову вибірку $\overrightarrow{X_n}$ обсягу n з генеральної сукупності X, закон розподілу якої заданий з точністю до параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Тобто задана статистична параметрична модель.

$$\{F(x, \vec{\Theta}); \vec{\Theta} \in \boldsymbol{\theta}\}$$

Необхідно знайти такі статистики $\overline{\Theta}(\overrightarrow{X_n})$ та $\underline{\Theta}(\overrightarrow{X_n})$, щоб з ймовірністю γ виконувалася рівність

$$P\{\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})\} = \gamma \qquad (3.1)$$

Говорять про інтервальну оцінку для параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Інтервал

$$\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n}) \qquad (3.2)$$

називають **довірчим інтервалом**, або (γ-інтервалом) для параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$, а величини $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ – верхня та нижня границі інтервальної оцінки.

Коефіцієнт γ називають коефіцієнтом довіри, довірчою ймовірністю, або рівнем довіри.

Інтервальна оцінка $\left(\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})\right)$ – це інтервал з випадковими межами, який із заданою ймовірністю покриває істинне значення параметра $\vec{\Theta}$. Отже, для різних реалізацій випадкової вибірки $\overrightarrow{X_n}$, тобто для різних елементів виборочного простору, статистики $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ можуть мати різні значення.

На відміну від точкової оцінки, інтервальна оцінка характеризується двома числами - кінцями інтервалу. Отже,

можна сказати, що інтервальні оцінки є певною мірою повнішими і надійними характеристиками порівняно з точковими.

На відміну від точкової оцінки, інтервальна оцінка дозволяє отримати ймовірнісну характеристику точності оцінювання невідомого параметра.

Імовірнісною характеристикою точності оцінювання параметра є випадкова величина, яка для будь-якої реалізації $\overrightarrow{x_n}$ випадкової вибірки $\overrightarrow{X_n}$, є довжина інтервалу $\left(\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})\right)$:

$$l(\overrightarrow{X_n}) = \overline{\Theta}(\overrightarrow{X_n}) - \underline{\Theta}(\overrightarrow{X_n})$$

Іноді параметр оцінюють лише зверху або лише знизу. Тоді відповідні статистики називають односторонніми нижніми або верхніми γ-довірчими межами.

$$P\{\underline{\Theta}(\overrightarrow{X_n}) < \Theta\} = \gamma$$
$$P\{\Theta < \overline{\Theta}(\overrightarrow{X_n})\} = \gamma$$

Тут $\overline{\Theta}(\overrightarrow{X_n})$ – одностороння верхня довірча границя, $\underline{\Theta}(\overrightarrow{X_n})$ – одностороння нижня довірча границя.

Властивості точкових оцінок

Як уже згадувалося раніше, як точкова оцінка параметр θ використовує різні статистики. Наприклад, як точкову оцінку μ=M(X) можна запропонувати такі статистики

$$\overline{\Theta}(\overrightarrow{X_n}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{\Theta}(\overrightarrow{X_n}) = \overline{X} = \frac{\mathrm{X}_1 + \mathrm{X}_n}{2}$$

$$\overline{\Theta}(\overrightarrow{X_n}) = \overline{X} = \begin{cases} \dfrac{\mathrm{X}_{\frac{n}{2}} + \mathrm{X}_{\frac{n}{2}+1}}{2}, & X - \text{парне} \\ X_{\frac{n+1}{2}}, & \mathrm{X} - \text{ непарне} \end{cases}$$

Яка із зазначених статистик є «найкращою» для отримання точкової оцінки? Вибирається необхідна функція. Розглянемо, які властивості має статистика.

Незміщена оцінка

З практичної точки зору важливо, щоб обрана як оцінка статистика не давала систематичних похибок, тобто занижених або завищених значень параметра, що оцінюється. Статистика, що задовольняє цій властивості, називається **незміщеною** оцінкою.

Статистику $\widehat{\Theta} = \widehat{\Theta}(\overrightarrow{X_n})$, називають **незміщеною** оцінкою параметра θ, якщо її математичне сподівання збігається з θ для будь-якого фіксованого n.

$$\mathrm{M}\,\widehat{\Theta}(\overrightarrow{X_n}) = \Theta$$

Оцінка є зміщеною з параметром зміщення b, якщо ця рівність не виконується.

$$b_n(\Theta) = \mathrm{M}\,\widehat{\Theta}(\overrightarrow{X_n}) - \Theta$$

Зміщення оцінки можна усунути, ввівши відповідну виправлення. Іноді досить знайти асимптотично незміщену оцінку.

Оцінка $\widehat{\Theta} = \widehat{\Theta}(\overrightarrow{X_n})$ є **асимптотично незміщеною**, якщо за $n \to \infty$ вона сходиться ймовірно до свого математичного сподівання для будь-якого $\varepsilon > 0$.

$$\lim_{n\to\infty} P(\left|\widehat{\Theta}(\overrightarrow{X_n}) - \mathrm{M}\,\widehat{\Theta}(\overrightarrow{X_n})\right|) < \varepsilon) = 1$$

Ефективна оцінка

Нехай є дві незміщені оцінки $\widehat{\Theta}(\overrightarrow{X_n})$ та $\widetilde{\Theta}(\overrightarrow{X_n})$, такі, що для дисперсій цих оцінок $D\widehat{\Theta}(\overrightarrow{X_n})$ та $D\widetilde{\Theta}(\overrightarrow{X_n})$, виконується нерівність

$$D\widehat{\Theta}(\overrightarrow{X_n}) \le D\widetilde{\Theta}(\overrightarrow{X_n}) \qquad (3.3)$$

Очевидно, що при виборі статистики, яка використовується для отримання оцінки параметра, слід віддати перевагу статистику із меншою дисперсією.

Якщо у деякому класі незміщених оцінок параметра є така $\widehat{\Theta}(\overrightarrow{X_n})$, що нерівність (3.3) виконується для всіх $\widetilde{\Theta}(\overrightarrow{X_n})$, то кажуть, що є **ефективною** в даному класі оцінок.

Таким чином, дисперсія ефективної оцінки параметра у деякому класі є мінімальною серед дисперсій усіх оцінок. Ефективну оцінку називають також незміщеною оцінкою з мінімальною дисперсією, або оптимальною оцінкою.

Достатні статистики

Оцінка є **достатньою** статистикою, якщо вся отримана з вибірки інформація щодо параметра міститься в його оцінці. Якщо відома достатня статистика, то жодна інша статистика, обчислена за тією ж вибіркою, не може дати додаткову інформацію про параметр.

Розглянемо випадкову вибірку $\overrightarrow{X_n} = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n$ обсягу n генеральної сукупності X. Нехай функція розподілу генеральної сукупності відома з точністю параметра $F(x;\ \theta)$. Тобто вид функції розподілу відомий, а параметр невідомий. Таким чином, розглядається параметрична модель

$$\{F(x;\ \theta);\ \theta \in \Theta\}$$

Розглянемо деяку статистику $T = T(\overrightarrow{X_n})$. Нехай буде відома не вся вибірка (реалізація випадкової вибірки), а лише значення статистики, отримане за результатами цієї вибірки $T(\overrightarrow{x_n}) = t$.

Введемо умовну функцію розподілу випадкової вибірки $\overrightarrow{X_n}$ за умови, що $T(\overrightarrow{x_n}) = t$.

$$F_{\underset{X_n}{\rightarrow}}(x_1, x_2, \dots, x_n / T(\overrightarrow{x_n}) = t)$$

Статистика $T(\overrightarrow{x_n})$ називається достатньою, для параметра θ, якщо умовна функція розподілу випадкової вибірки $\overrightarrow{X_n}$ з характеристикою $F\underset{X_n}{\rightarrow}(x_1, x_2, \dots, x_n / T(\overrightarrow{x_n}) = t)$ не залежить від параметра θ для будь-якого t.

З визначення достатньої статистики випливає, що для фіксованого значення t при зміні параметра θ, умовний розподіл $F\underset{X_n}{\rightarrow}(x_1, x_2, \dots, x_n / T(\overrightarrow{x_n}) = t)$ не зміниться. Тобто значення статистики t дає повну інформацію про параметр θ.

Точкові оцінки моментів випадкової величини

Точкова оцінка математичного сподівання

Статистика

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.7}$$

(вибіркова середня) є оцінкою математичного сподівання генеральної сукупності X з кінцевою $\theta = M(X) = \mu$ дисперсією. Вибіркова середня є незміщеною, спроможною та ефективною в класі всіх лінійних оцінок, тобто оцінок виду

$$\widehat{\Theta}(\overrightarrow{X_n}) = \sum_{i=1}^{n} \alpha_i X_i$$

$$\sum_{i=1}^{n} \alpha_i = 1$$

При розгляді властивостей вибіркового середнього необхідно пам'ятати таке.

Елементи $X_i, i = \overline{1,n}$ випадкової вибірки $\overrightarrow{X_n}$ є незалежними випадковими величинами і розподілені так само, як і генеральна сукупність X. Отже,

$$M(X_i) = M(X) = \mu \quad D(X_i) = D(X) = \sigma^2, i = \overline{1,n}$$

Незміщеність оцінки. Враховуючи властивості математичного сподівання, отримуємо

$$M(\bar{X}) = M\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n} n\mu = \mu$$

Це доводить незміщеність оцінки $\bar{X}$.

Спроможність оцінки. Оскільки послідовність $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ складається з незалежних однаково розподілених величин з кінцевою дисперсією, то через закон великих чисел у формі Чебишева для будь-якого ε

$$P(|\bar{X} - \mu| < \varepsilon) \xrightarrow[n \to \infty]{P} 1 .$$

Тобто оцінка сходиться ймовірно до оцінюваного параметром. Отже, оцінка спроможна.

Ефективність оцінки. Доведемо, що

$$D\left(\widetilde{\Theta}\left(\overrightarrow{X_n}\right)\right) = D\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} D(\alpha_i X_i) = \sum_{i=1}^{n} \alpha_i^2 D(X_i) = \sigma^2 \sum_{i=1}^{n} \alpha_i^2$$

досягає свого мінімального значення при $\alpha_i = 1/n$, тобто коли $\widetilde{\Theta}\left(\overrightarrow{X_n}\right) = \bar{X}$. Таким чином доведено, що оцінка є ефективною.

Знайдемо умовний мінімум функції

$$g(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{i=1}^{n} \alpha_i^2$$

при накладенні обмеження $\sum_{i=1}^{n} \alpha_i = 1$

Складемо функцію Лагранжа з множником Лагранжа λ

$$L(\alpha_1, \alpha_2, \ldots, \alpha_n; \lambda) = \sum_{i=1}^{n} \alpha_i^2 + \lambda\left(\sum_{i=1}^{n} \alpha_i - 1\right)$$

Необхідні умови існування умовного екстремуму виражаються так

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_i} = 2\alpha_i + \lambda = 0, \;\; i = \overline{1,n} \\ \dfrac{\partial L}{\partial \alpha_i} = \sum_{i=1}^{n} \alpha_i \;\; - 1 = 0, \end{cases}$$

Вирішивши систему (для вирішення треба підсумувати перші рівняння), одержуємо λ = –2/n, $\alpha_i = 1/n, \;\; i = \overline{1,n}$. Тобто за цих значень аргументу функція $g(\alpha_1, \alpha_2, \dots, \alpha_n)$ має умовний мінімум.

Оцінка дисперсії.

Статистика

$$\hat{\sigma}^2(\overrightarrow{X_n}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (3.8)$$

(вибіркова дисперсія) є оцінкою дисперсії $\Theta = D(X) = \sigma^2$ генеральної сукупності X. Тут $\overrightarrow{X_n}$ – випадкова вибірка обсягу n з генеральної сукупності X. Вибіркова дисперсія є зміщеною заможною оцінкою дисперсії генеральної сукупності.

Методи знаходження точкових оцінок параметрів розподілу

Розглянемо два методи знаходження точкових оцінок параметрів розподілу генеральної сукупності.

Метод моментів. Метод моментів був запропонований англійським статистиком Пірсоном і є одним із перших розроблених методів оцінювання.

Нехай є випадкова вибірка $\overrightarrow{X_n} = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n)$ обсягу n генеральної сукупності X. Її розподіл $p(\vec{\Theta}, x)$ відомий з точністю до вектора параметрів $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Необхідно знайти оцінку параметра Θ за випадковою вибіркою $\overrightarrow{X_n}$.

Розглянемо вибіркові моменти $\hat{m}_k(\overrightarrow{X_n})$. Вибіркові моменти є оцінками моментів генеральної сукупності. При великому обсязі вибірки генеральні моменти можуть бути замінені вибірковими.

У методі моментів як точкову оцінку $\widehat{\Theta}(\overrightarrow{X_n})$ вектору параметрів $\overrightarrow{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$ беруть статистику, значення якої для будь-якої реалізації $\overrightarrow{x_n}$ випадкової вибірки $\overrightarrow{X_n}$ одержують як розв'язання системи рівнянь

$$\widehat{m}_k = m_k(\overrightarrow{\Theta}), k = \overline{1,r} \quad (3.10)$$

Таким чином, вибіркові моменти прирівнюють до теоретичних моментів, що залежать від шуканого параметра.

Ці рівняння у багатьох випадках прості і не викликають обчислювальних складнощів. Розглянемо кілька прикладів.

Припустимо, випадкова величина має розподіл

$$p(x, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

За значеннями випадкової вибірки x_i необхідно знайти оцінку параметра розподілу σ.

Обчислимо початковий теоретичний момент першого порядку:

$$m_1(\sigma) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + t) e^{-\frac{t^2}{2}} d(\frac{t^2}{2}) = \mu$$

Вибірковий початковий момент першого порядку – це вибіркове середнє

$$\widehat{m}_1 = \overline{\mathrm{X}} = \frac{1}{n} \sum_{i=1}^{n} x_k$$

Прирівнюють теоретичний (залежний від параметра) момент та вибірковий і знаходять значення параметра:

$$m_1 = \widehat{m}_1$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_k$$

Метод максимальної правдоподібності

Метод запропонований Фішером і є найбільш універсальним. Обчислення практично бувають досить складними і вимагають застосування чисельних методів.

Нехай є випадкова вибірка $\overrightarrow{X_n} = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n)$ обсягу n генеральної сукупності X. Її розподіл $p(x, \vec{\Theta})$ відомий з точністю до вектора параметрів $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Необхідно знайти оцінку параметра Θ за випадковою вибіркою $\overrightarrow{X_n}$.

Введемо так звану функцію правдоподібності

$$L(\mathrm{X}_1, \mathrm{X}_2; \vec{\Theta}) = p(x_1, \vec{\Theta}) * p(x_2, \vec{\Theta}) * \dots * p(x_n, \vec{\Theta})$$

$$L(\mathrm{X}_1, \mathrm{X}_2; \vec{\Theta}) = \prod_{i=1}^{n} p(x_i, \vec{\Theta})$$

Тут $p(x, \vec{\Theta})$ – щільність розподілу випадкової величини (у разі неперервної випадкової величини) або ймовірність події $X = x$ (у разі дискретної випадкової величини).

Оцінкою максимальної правдоподібності параметра θ називають статистику $\widehat{\Theta}(\overrightarrow{X_n})$, значення якої для будь-якої вибірки $\overrightarrow{x_n}$ створює умову

$$L(\overrightarrow{x_n}; \widehat{\Theta}) = \max_{\vec{\Theta}} L(\overrightarrow{x_n}; \widehat{\Theta}) \qquad (3.12)$$

Якщо функція правдоподібності диференційована, то значення точкової оцінки максимальної правдоподібності для скалярного параметра задовольняють рівняння (необхідна умова екстремуму)

$$\frac{\partial L(\overrightarrow{x_n}; \Theta)}{\partial \Theta} = 0 \qquad (3.13)$$

Або, оскільки при логарифмуванні точки екстремуму залишаються ті ж, а рівняння, як правило, спрощується

$$\frac{\partial ln L(\overrightarrow{x_n}; \Theta)}{\partial \Theta} = 0 \qquad (3.14)$$

Якщо розподіл випадкової величини залежить від вектору випадкових параметрів, то останнє рівняння розпадається на систему рівнянь

$$\frac{\partial lnL(\overrightarrow{x_n};\Theta)}{\partial \Theta_k} = 0, k = \overline{1,r} \qquad (3.15)$$

Рівняння (3.12) - (3.15) називаються рівняннями правдоподібності.

Основні властивості оцінок методу максимальної правдоподібності наведемо без доказів.

1. Якщо існує ефективна оцінка для скалярного параметра, то рівняння правдоподібності має єдине рішення, яке є вибірковим значенням цієї оцінки.

2. Якщо існує достатня статистика параметра, рішення рівняння правдоподібності є функціями від вибіркового значення цієї статистики.

Оцінки, отримані методом максимальної правдоподібності, можуть бути зміщеними та неефективними. Зміщення можна усунути. У багатьох випадках неефективні оцінки є асимптотично ефективними.

Розглянемо приклад.

Для загальної нормальної моделі необхідно знайти оцінку вектору параметрів $\overrightarrow{\Theta} = (\Theta_1, \Theta_2)$ методом максимальної правдоподібності.

Функція правдоподібності для нормальної моделі має вигляд.

$$L(\overrightarrow{X_n}; \Theta_1, \Theta_2) = \frac{1}{\Theta_2\sqrt{2\pi}} e^{-\frac{\sum_{i=1}^{n}(X_i-\Theta_1)^2}{2\theta_2^2}}$$

$$ln(\overrightarrow{X_n}; \Theta_1, \Theta_2) = -nln\sqrt{2\pi} - nln\,\Theta_2 - \frac{\sum_{i=1}^{n}(X_i - \Theta_1)^2}{2\Theta_2^2}$$

Параметри знаходимо з розв'язання системи рівнянь.

$$\frac{\partial lnL(\overrightarrow{x_n};\Theta)}{\partial\Theta_1} = \frac{\sum_{i=1}^{n}(X_i - \Theta_1)}{\Theta_2^3}$$

$$\frac{\partial lnL(\overrightarrow{x_n};\Theta)}{\partial\Theta_2} = -\frac{n}{\Theta_2} + \frac{\sum_{i=1}^{n}(X_i - \Theta_1)^2}{\Theta_2^2}$$

Вирішуючи систему, отримуємо

$$\widehat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \widehat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Отже, оцінками максимальної правдоподібності для математичного сподівання $\Theta_1 = M(X)$ та дисперсії $\Theta_2^2 = D(X)$ випадкової величини, розподіленої за нормальним законом, є вибіркове середнє та вибіркова дисперсія.

Розглянемо ще один приклад.

Нехай в експерименті величина X – час роботи приладу до відмови – має експотенційний розподіл із щільністю

$$p(x,\lambda) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0\,, x < 0 \end{cases}$$

Тут - невідомий параметр. Застосовуючи метод максимального правдоподібності необхідно знайти точкову оцінку для параметра λ.

Нехай $\overrightarrow{x_n} = x_1, x_2, \dots, x_n$ – будь-яка реалізація випадкової вибірки $\overrightarrow{X_n}$ генеральної сукупності X.

У цьому випадку функція правдоподібності

$$L(x_1 \dots x_n;\ \lambda) = \prod_{i=1}^{n} p(\overrightarrow{X_n}\ ,\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\sum_{i=1}^{n} x_i}$$

$$L(x_1 \dots x_n;\ \lambda) = nln\lambda - \lambda\sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}$$

Рівняння правдоподібності мають вигляд:

$$\frac{\partial lnL(x_1 \dots x_n;\ \lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Точковою оцінкою невідомого параметра λ є величина

$$\hat{\lambda}(\overrightarrow{X_n}) = \frac{1}{\bar{X}}$$

Отримана відповідь узгоджується з тим, що $M(X) = 1/\lambda$ 1 а оцінкою математичного сподівання $M(X) = \mu$ є вибіркове середнє $\bar{X} = M(X)$.

Інтервальні оцінки

Нехай є випадкова вибірка $\overrightarrow{X_n} = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n)$ обсягу n генеральної сукупності X. Її розподіл $p(\vec{\Theta}, x)$ відомий з точністю до вектора параметрів $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Необхідно знайти оцінку параметра Θ за випадковою вибіркою $\overrightarrow{X_n}$.Таким чином, розглядається параметрична модель $\{F(x;\ \theta);\ \theta \in \Theta\}$.

Необхідно знайти такі статистики $\overline{\Theta}(\overrightarrow{X_n})$ та $\underline{\Theta}(\overrightarrow{X_n})$, щоб з ймовірністю γ виконувалася рівність

$$P\{\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})\} = \gamma$$

У цьому випадку говорять про інтервальну оцінку для параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$. Інтервал $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ називають **довірчим інтервалом**, або (γ-інтервалом) для параметра $\vec{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_n)$, а величини $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ – верхня та нижня границі інтервальної оцінки.

Коефіцієнт γ називають **коефіцієнтом довіри**, довірчою ймовірністю, або рівнем довіри.

Інтервальна оцінка $\left(\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})\right)$ – це інтервал з випадковими межами, який із заданою ймовірністю покриває істинне значення параметра $\vec{\Theta}$. Отже, для різних реалізацій випадкової вибірки $\overrightarrow{X_n}$,

тобто для різних елементів виборочного простору, статистики $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ можуть мати різні значення.

На відміну від точкової оцінки, інтервальна оцінка характеризується двома числами - кінцями інтервалу. Отже, можна сказати, що інтервальні оцінки є певною мірою повнішими і надійними характеристиками порівняно з точковими.

На відміну від точкової оцінки, інтервальна оцінка дозволяє отримати ймовірнісну характеристику точності оцінювання невідомого параметра.

Імовірнісною характеристикою точності оцінювання параметра є випадкова величина, яка для будь-якої реалізації $\overrightarrow{x_n}$ випадкової вибірки $\overrightarrow{X_n}$, є довжина інтервалу $\left(\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})\right)$:

$$l(\overrightarrow{X_n}) = \overline{\Theta}(\overrightarrow{X_n}) - \underline{\Theta}(\overrightarrow{X_n})$$

Іноді параметр оцінюють лише зверху або лише знизу. Тоді відповідні статистики називають односторонніми нижніми або верхніми γ-довірчими межами.

$$P\{\underline{\Theta}(\overrightarrow{X_n}) < \Theta\} = \gamma$$
$$P\{\Theta < \overline{\Theta}(\overrightarrow{X_n})\} = \gamma$$

Тут $\overline{\Theta}(\overrightarrow{X_n})$ – одностороння верхня довірча границя, $\underline{\Theta}(\overrightarrow{X_n})$ – одностороння нижня довірча границя.

Побудова довірчих інтервалів методом центральної статистики

Розглянемо випадкову вибірку об'єму n $\overrightarrow{X_n}$ з функцією розподілу $F(x;\theta)$, яка залежить від невідомого параметра θ. Один з найбільш поширених методів побудови інтервальних оцінок пов'язаний з використанням центральної статистики – будь-якої статистики $T(\overrightarrow{X_n}, \theta)$, функція розподілу якої

$$F_t(t) = P\{T(\overrightarrow{X_n}, \theta) < \mathrm{t}\}$$

не залежить від параметра θ.

При побудові інтервальної оцінки будемо припускати наступне:

1. Функція розподілу $F_t(t)$ неперервна та зростаюча.

2. Задано такі позитивні коефіцієнти α та β, що $\gamma = 1 - \alpha - \beta$.

3. Для будь-якої вибірки $\overrightarrow{x_n}$ з генеральної сукупності X функція $T(\overrightarrow{X_n}, \theta)$ є неперервною і зростаючою (спадаючою) функцією параметра θ.

З припущення 1 випливає, що для будь-якого q з інтервалу (0,1) існує єдиний корінь h_q рівняння $F_t(t) = q$ який є **квантиллю** рівня функції q функції $F_t(t) = q$ розподілу випадкової величини $T(\overrightarrow{X_n}, \theta)$.

Зважаючи на припущення 2, отримаємо

$$F_t(t) = P\{T(\overrightarrow{X_n}, \theta) < t\}$$

$$P\{h_\alpha < T(\overrightarrow{X_n}, \theta) < h_{1-\beta}\} = F_t(h_{1-\beta}) - F_t(h_\alpha) = 1 - \alpha - \beta \quad (3.16)$$

Рівність (3.16) справедлива для будь-якого значення параметра θ, оскільки $T(\overrightarrow{X_n}, \theta)$ – центральна статистика та її функція розподілу ня $F_t(t)$ не залежить від параметра θ.

Для побудови інтервальної оцінки необхідне рівняння (3.14) перетворити на вираз виду

$$P\{\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})\} = \gamma$$

Припустимо, що $T(\overrightarrow{X_n}, \theta)$ – зростаюча функція, тоді по допущенню 3, для кожної вибірки $\overrightarrow{x_n}$ рівняння

$$T(\overrightarrow{X_n}, \theta) = h_\alpha \quad T(\overrightarrow{X_n}, \theta) = h_{1-\beta}$$

мають єдині рішення $\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})$ а нерівності

$h_\alpha < T(\overrightarrow{X_n}, \theta) < h_{1-\beta}$ та $\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})$ рівносильними.

Таким чином

$$P\{\underline{\Theta}(\overrightarrow{X_n}) \le \Theta \le \overline{\Theta}(\overrightarrow{X_n})\} = P\{h_\alpha < T(\overrightarrow{X_n}, \theta) < h_{1-\beta}\}$$

і $\left(\underline{\Theta}(\overrightarrow{X_n}), \overline{\Theta}(\overrightarrow{X_n})\right)$ є шуканою оцінкою.

Практично побудова довірчого інтервалу зводиться до наступного:

1. Побудова центральної статистики $T(\overrightarrow{X_n}, \theta)$ та її функції розподілу $F_t(t)$.

2. Подання заданого коефіцієнта довіри у вигляді $\gamma = 1 - \alpha - \beta$.

3. Знаходження квантилів $h_\alpha, h_{1-\beta}$ рівнів α та $1 - \beta$ функції розподілу $F_t(t)$.

4. Знаходження нижньої та верхньої меж шуканої інтервальної оцінки рішенням рівнянь Розглянемо приклади побудови інтервальної оцінки для параметрів нормального розподілу.

Оцінка для математичного сподівання за відомої дисперсії

Розглянемо випадкову вибірку обсягу n $\overrightarrow{X_n}$ з генеральної сукупності X, розподілену за нормальним законом з параметрами μ та σ^2.

Визначимо оцінку для математичного сподівання за відомої дисперсії.

Розглянемо статистику

$$T\left(\overrightarrow{X_n}, \mu\right) = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \qquad (3.17)$$

Можна показати, що ця статистика має стандартний нормальний розподіл із параметрами $\mu = 0$ та $\sigma^2 = 1$, тобто є центральною статистикою. $T\left(\overrightarrow{X_n}, \mu\right)$ визначена формулою (3.15), та є функція спадною по μ.

Відповідна система рівнянь має вигляд:

$$\frac{\bar{X}-\underline{\mu}}{\sigma}\sqrt{n}=u_1-\beta$$
$$\frac{\bar{X}-\bar{\mu}}{\sigma}\sqrt{n}=u_\alpha$$

де u_q - квантиль стандартного нормального розподілу. Враховуючи, що $u_{1-\alpha}=-u_\alpha$, отримуємо нижню та верхню межі довірчого інтервалу для параметра μ при $\gamma=1-\alpha-\beta$.

$$\underline{\mu}=\bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\beta}\qquad \bar{\mu}=\bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\alpha}$$

Довірчий інтервал для математичного сподівання з відомою дисперсією записується так:

$$\left(\bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\beta},\bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right) \tag{3.18}$$

На практиці часто використовується формула:

$$\left(\bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}},\bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right) \tag{3.19}$$

Оцінка для математичного сподівання за невідомої дисперсії

Визначимо оцінку для математичного сподівання за невідомої дисперсії. Розглянемо статистику

Розглянемо статистику

$$T\left(\overrightarrow{X_n},\mu\right)=\frac{\bar{X}-\mu}{S\left(\overrightarrow{X_n}\right)}\sqrt{n} \tag{3.20}$$

Тут використовується статистика $S(\overrightarrow{X_n})$ квадрат якої є незміщеною та спроможною оцінкою дисперсії.

$$S\left(\overrightarrow{X_n}\right)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Можна показати, що ця статистика має стандартний нормальний розподіл із параметрами $\mu=0$ та $\sigma^2=1$, тобто є центральною статистикою. $T(\overrightarrow{X_n},\mu)$ визначена формулою (3.15), та є функція спадною по μ.

Статистика $T(\overrightarrow{X_n}, \mu)$ є центральною статистикою, має розподіл Стьюдента з n-1 ступенями свободи та її функція розподілу не залежить від параметрів μ та σ^2.

Щільність ймовірності розподілу Стьюдента визначається параметром n – обсягом вибірки, чи числом ступенів свободи $k = n - 1$.

$$S(t,n) = B_n(1 + \frac{t^2}{n-1})^{-\frac{n}{2}}$$

$$B_n = (1 + \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)})$$

Відповідна система рівнянь:

$$\frac{\bar{X} - \underline{\mu}}{S(\overrightarrow{X_n})}\sqrt{n} = t_{1-\beta}(n-1)$$

$$\frac{\bar{X} - \bar{\mu}}{S(\overrightarrow{X_n})}\sqrt{n} = t_{\alpha}(n-1)$$

де $t_q(n-1)$ – квантиль рівня q розподілу Стьюдента з $n - 1$ ступенями свободи. Враховуючи, що $t_{1-\alpha}(n-1) = -t_{\alpha}(n-1)$ отримуємо нижню та верхню межі довірчого інтервалу для параметра μ при $\gamma = 1 - \alpha - \beta$.

$$\underline{\mu} = \bar{x} - \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\beta}(n-1) \qquad (3.21)$$

$$\bar{\mu} = \bar{x} - \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\alpha}(n-1) \qquad (3.22)$$

Довірчий інтервал

$$\left(\bar{x} - \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\beta}(n-1), \bar{x} - \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\alpha}(n-1)\right) \qquad (3.23)$$

Сенс отриманого співвідношення: з надійністю γ можна стверджувати, що довірчий інтервал покриває невідомий параметр μ. На практиці часто застосовується формула:

$$\left(\bar{x} - \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\frac{\alpha}{2}}(n-1), \bar{x} + \frac{S(\overrightarrow{X_n})}{\sqrt{n}} u_{1-\frac{\alpha}{2}}(n-1)\right) \qquad (3.24)$$

Оцінка для середнього квадратичного відхилення

Розглянемо випадкову вибірку обсягу n $\overrightarrow{X_n}$ з генеральної сукупності X, розподілену за нормальним законом з параметрами μ та σ^2.

Визначимо оцінку для середнього відхилення квадратичного.

Розглянемо статистику

$$T\left(\overrightarrow{X_n}, \sigma\right) = \frac{(n-1)S^2\left(\overrightarrow{X_n}\right)}{\sigma^2} \tag{3.25}$$

Статистика є центральною статистикою, має розподіл «хі»-квадрат з $n-1$ ступенями свободи, не залежить від μ і σ^2, є спадною функцією σ.

Відповідна система має вигляд

$$\frac{(n-1)S^2\left(\overrightarrow{X_n}\right)}{\underline{\sigma^2}} = \chi^2_{1-\beta}(\text{n-1}) \tag{3.26}$$

$$\frac{(n-1)S^2\left(\overrightarrow{X_n}\right)}{\overline{\sigma^2}} = \chi^2_{\alpha}(\text{n-1}) \tag{3.27}$$

де χ^2_q(n-1)– квантиль рівня q розподілу «хі»-квадрат з $n-1$ ступенями свободи. Отримуємо нижню та верхню межі довірчого інтервалу для параметра при $\gamma = 1- \alpha - \beta$.

$$\underline{\sigma} = \frac{\sqrt{n-1}\,S\left(\overrightarrow{X_n}\right)}{\sqrt{\chi^2_{1-\beta}(\text{n}-1)}} \tag{3.28}$$

$$\bar{\sigma} = \frac{\sqrt{n-1}\,S\left(\overrightarrow{X_n}\right)}{\sqrt{\chi^2_{\alpha}(\text{n}-1)}} \tag{3.29}$$

Довірчий інтервал: $\left(\frac{\sqrt{n-1}\,S\left(\overrightarrow{X_n}\right)}{\sqrt{\chi^2_{1-\beta}(\text{n}-1)}}, \frac{\sqrt{n-1}\,S\left(\overrightarrow{X_n}\right)}{\sqrt{\chi^2_{\alpha}(\text{n}-1)}}\right)$

1.4. Перевірка статистичних гіпотез

Статистичні гіпотези: основні визначення

Перевірка статистичних гіпотез належить до одного з найпоширеніших завдань додатків математичної статистики.

У попередніх розділах було розглянуто завдання оцінки невідомих параметрів розподілу та побудови точкових та інтервальних оцінок параметрів розподілу.

У завданнях щодо перевірки статистичних гіпотез, розглядаємо у певному сенсі зворотну ситуацію.

На підставі тієї чи іншої апріорної інформації висувається припущення (гіпотеза) про значення параметра $\vec{\Theta}$. Після цього проводиться експеримент. В результаті експерименту отримують реалізацію $\overrightarrow{x_n} = x_1, x_2, \dots, x_n$ випадкової вибірки $\overrightarrow{X_n} = \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n$ з генеральної сукупності X, розподіл якої залежить від шуканого параметра. За даними експерименту необхідно вирішити, чи узгоджується висунуте припущення (гіпотеза) з експериментальними даними, чи гіпотезу необхідно відхилити.

Іноді висувається гіпотеза не про значення параметра, а функцію розподілу.

Статистичною називають гіпотезу про вид невідомого розподілу або параметр відомого розподілу.

Наведемо приклади статистичних гіпотез. Статистична гіпотеза може бути сформульована таким чином: «Генеральна сукупність розподілена за законом Пуассона». Іншими прикладами формулювань статистичних гіпотез є: «Дисперсія нормального розподілу не дорівнює 5», або «Математичне сподівання нормального розподілу дорівнює 7».

Нульовою (чи основною) називають висунуту гіпотезу. Нульову гіпотезу позначають $H0$. Поряд з основною гіпотезою розглядають також гіпотезу, що їй суперечить.

Альтернативною (конкуруючою) $H1$ називають гіпотезу, що суперечить нульовій.

Якщо за результатами експериментальних даних приймають рішення про відхилення висунутої нульової гіпотези, приймають альтернативну гіпотезу.

Наведемо приклади формулювань нульових та альтернативних гіпотез.

Нехай нульова гіпотеза полягає в тому, що математичне сподівання генеральної сукупності, яка підпорядковується закону нормального розподілу з відомою дисперсією, дорівнює 10, тобто $\mu = 10$. Тоді альтернативна гіпотеза сформульована як $\mu \neq 10$. Короткий запис обох гіпотез виглядає наступним чином:

$$H_0: \mu = 10; \quad H_1: \mu \neq 10;$$

Простою називається статистична гіпотеза, яка містить лише одне припущення. Наприклад, $H: \mu = 150$.

Складною називається статистична гіпотеза, що складається з кінцевого чи нескінченного числа простих гіпотез. Прикладами складних гіпотез є такі гіпотези:

$H: \mu > 150, H: \sigma > \sigma_0, H: \mu = b_i$ де b_i – будь-яке число.

Статистичні гіпотези щодо невідомого параметра θ називають **параметричними**.

У загальному випадку можна сказати, що статистичну гіпотезу називають простою, якщо вона має вигляд:

$$H: \vec{\theta} = \vec{\theta}_0$$

і статистичну гіпотезу називають складною, якщо вона має вигляд: $H: \vec{\theta} \in \boldsymbol{D}$

де D – деяка множина значень $\vec{\theta}$, що складається більш ніж з одного елемента.

Якщо параметр є скаляром, то йдеться про **однопараметричні** гіпотези, якщо вектор, то про **багатопараметричні** гіпотези. Наведемо приклади простих та складних параметричних гіпотез.

Нехай $\overrightarrow{X_n}$ випадкова вибірка з генеральної сукупності X, розподіленої за нормальним законом з невідомим математичним сподіванням μ і відомою дисперсією σ^2.

Гіпотеза H: $\mu = \mu_0$, проста, якщо μ_0 – деяке задане значення. Гіпотеза H: $\mu < \mu_0$ – складна. Гіпотези H: $\mu \neq \mu_0$ та H: $\mu_0 < \mu_1$ також складні. Статистичну гіпотезу щодо виду розподілу називають **непараметричною**.

Статистичний критерій

Нехай необхідно перевірити просту параметричну статистичну гіпотезу щодо величини скалярного параметра θ. У даному випадку будемо розглядати дві прості гіпотези: нульву $H0$ та альтернативну $H1$:

H_0: $\theta = \theta_0$, H_1: $\theta = \theta_1$, де θ_0 і θ_1 – два задані різні значення.

За даними вибірки $\overrightarrow{x_n}$ необхідно прийняти рішення про справедливість однієї з гіпотез.

Критерієм, або статистичним критерієм, перевірки гіпотези називають правило, яким за даними вибірки $\overrightarrow{x_n}$ приймається рішення про справедливість або нульової, або альтернативної гіпотези.

Критерій задають за допомогою так званої критичної множини W, що є підмножиною вибіркового простору випадкової вибірки $\overrightarrow{X_n}$.

Рішення приймають в такий спосіб.

1. Якщо вибірка $\overrightarrow{x_n}$ належить множині W, то відкидають нульову гіпотезу H_0 та приймають альтернативну гіпотезу H_1.

2. Якщо вибірка $\overrightarrow{x_n}$ не належить множині W (належить доповненням W множини W до вибіркового простору), то відкидають альтернативну гіпотезу H_1 і приймають нульову гіпотезу H_0.

При використанні критерію можливі такі помилки:

1. Помилка першого роду у тому, що прийняли гіпотезу

H_1, але вірна гіпотеза H_0. (Тобто відкинуто правильну гіпотезу).

2. Помилка другого роду полягає в тому, що прийняли гіпотезу H_0, але вірна гіпотеза H_1. (Тобто прийнято неправильну гіпотезу).

Імовірності скоєння помилок першого і другого роду виражаються через умовні ймовірності.

Імовірність вчинення помилки першого роду (позначається α) – це умовна ймовірність події $\overrightarrow{X_n} \in W$ за умови, що нульова гіпотеза H_0 вірна. Імовірність вчинення помилки другого роду (позначається β) – це умовна ймовірність події $\overrightarrow{X_n} \in W$ за умови, що альтернативна гіпотеза H_1 вірна

Імовірності помилок першого і другого роду можуть бути розраховані за формулами:

$$\alpha = \int_{\dots} \int_{W} \prod_{k=1}^{n} p(t_k, \theta_0) dt_1 \dots dt_n \qquad \beta = \int_{\dots} \int_{\overline{W}} \prod_{k=1}^{n} p(t_k, \theta_0) dt_1 \dots dt_n$$

Імовірність помилки першого роду α називають рівнем значущості критерію. Чим менше α, тим менша можливість відкинути правильну нульову гіпотезу. Допустиму помилку α

зазвичай задають заздалегідь і вибирають $\alpha = 0{,}05;\ 0{,}01;\ 0{,}005;\ 0{,}0001.$

Розмір $1 - \beta$ називається потужністю критерію. Очевидно, що потужність критерію - це ймовірність відкинути основну гіпотезу, коли вона неправильна.

Критерій Неймана-Пірсона

При побудові критерію, зазвичай, виходять із необхідності максимізації його потужності $1 - \beta$ при фіксованому рівні значимості α.

Максимізація потужності критерію означає мінімізацію ймовірності вчинення помилки 2 роду. Таким чином, умова максимізації критерію означає збільшення ймовірності відкинути основну гіпотезу в тому випадку, якщо вона невірна за фіксованого рівня значущості α.

Нехай $\overrightarrow{X_n}$ випадкова вибірка з генеральної сукупності X, розподілена з щільністю $p(t, \theta)$, де θ невідомий параметр.

Розглянемо дві прості гіпотези: нульову H_0 та альтернативну H_1: $H_0: \theta = \theta_0 \quad H_1 \theta = \theta_1$ де θ_0 і θ_1 – два задані різні значення. Введемо функцію випадкової вибірки $\varphi(\overrightarrow{X_n})$, звану відношенням правдоподібності.

$$\varphi\left(\overrightarrow{X_n}\right) = \frac{L(\overrightarrow{X_n}; \theta_1)}{L(\overrightarrow{X_n}; \theta_0)} \qquad L(\overrightarrow{X_n}; \theta\ ) = p\left(\overrightarrow{X_n}; \theta\ \right)$$

Для побудови оптимального (тобто найбільш потужного) рівня значимості α критерію Неймана-Пірсона в критичну множину W включають ті елементи $\overrightarrow{x_n}$ вибіркового простору χ_n випадкової вибірки X ,для яких виконується нерівність

$$\varphi\left(\overrightarrow{X_n}\right) \geq \mathrm{C}_\varphi$$

Величину $C\varphi$ вибирають із умови

$$P\left\{\varphi\left(\overrightarrow{X_n}\right) \geq \mathrm{C}_\varphi | H_0\right\}$$

Визначення мінімального обсягу вибірки

У попередніх задачах ми припускали, що обсяг вибірки задано. Іноді необхідно визначити, яким має бути обсяг вибірки $n*$, при якому може бути побудований критерій для перевірки двох простих гіпотез $H_0: \theta = \theta_0 \quad H_1 \theta = \theta_1$ де θ_0 і θ_1 – два задані різні значення.

В даному випадку $n*$ визначається як мінімальне ціле значення n, для якого система нерівностей може бути виконана при деякому значенні константи $\mathrm{C} = C^*$.

$$\begin{cases} P\{\varphi(\overrightarrow{X_n}) \geq \mathrm{C}_\varphi | \theta = \theta_0\} \leq \alpha \\ P\{\varphi(\overrightarrow{X_n}) \geq \mathrm{C}_\varphi | \theta = \theta_1\} \leq \beta \end{cases}$$

При цьому відповідний оптимальний критерій Неймана-Пірсона, що забезпечує задані значення α і β, матиме критичну множину, що забезпечується нерівністю

$$\varphi(\overrightarrow{X_n}) \geq \mathrm{C}_\varphi.$$

Для забезпечення заданих значень α та β – помилок першого та другого роду мінімально необхідний обсяг вибірки n^* і відповідну константу C^* можна визначити із системи рівнянь

$$1 - \Phi\left(\frac{C - n\mu_0}{\sigma\sqrt{n}}\right) \leq \alpha \quad \Phi\left(\frac{C - n\mu_0}{\sigma\sqrt{n}}\right) \leq \beta$$

Використовуючи квантилі нормального розподілу, отримуємо

$$\frac{C - n\mu_0}{\sigma\sqrt{n}} = u_{1-\alpha} \qquad \frac{C - n\mu_0}{\sigma\sqrt{n}} = u_\beta = -u_{1-\beta}$$

Вирішуючи системи рівнянь, остаточно для мінімального необхідного обсягу вибірки n^* отримуємо:

$$n^* = \frac{\sigma^2\left(u_{1-\alpha} + u_{1-\beta}\right)^2}{(\mu_1 - \mu_0)^2}$$

Складні параметричні гіпотези. Побудова критерію для перевірки складних параметричних гіпотез

Критерій перевірки складних гіпотез, як і і критерій перевірки простих гіпотез, розглянутий вище, задається за допомогою критичної множини W реалізацій випадкової вибірки X. Критерій перевірки складних гіпотез, як і і критерій перевірки простих гіпотез, розглянутий вище, задається за допомогою критичної множини W реалізацій випадкової вибірки X.

Перевірка гіпотез про математичне сподівання

1. Розглянемо перевірку простої гіпотези проти складної щодо математичного сподівання μ нормального закону розподілу з відомою дисперсією σ^2.

Причому нульову та альтернативну гіпотезу сформулюємо в такий спосіб: $H_0: \mu = \mu_0; H_1: \mu > \mu_0$.

Виберемо деякий μ_1, такий, що $\mu_1 > \mu_0$.

Для будь-якого $\mu_1 > \mu_0$ може бути сформульований оптимальний критерій Неймана Пірсона. Критична область оптимального критерію Неймана-Пірсона для фіксованого значення для простих гіпотез $H_0: \mu = \mu_0$ проти $H_1: \mu > \mu_0$, розглянутих вище, має вигляд

$$\varphi(\overrightarrow{X_n}) = \sum_{i=1}^{n} x_i \geq \mathrm{C}_\varphi$$

Константу вибирають С із умови

$$\mathrm{P}\left\{\sum_{i=1}^{n} x_i \geq \mathrm{C}_\varphi \middle| \mu = \mu_0\right\}=\alpha$$

або

$$1 - \Phi\left(\frac{C - n\mu_0}{\sigma\sqrt{n}}\right) = \alpha$$

Вочевидь, що константа С залежить від μ_1. Отже, попобудовано критерій з критичною множиною, що задається наступним виразом

$$\sum_{i=1}^{n} x_i \geq \mathrm{C} = n\mu_0 + u_{1-\alpha}\sigma\sqrt{n}$$

є рівномірно найбільш потужним критерієм розміру α для даної задачі перевірки простої гіпотези проти складної.

2. Розглянемо перевірку простої гіпотези проти складної щодо математичного сподівання μ нормального закону розподілу з відомою дисперсією σ^2.

Причому нульову та альтернативну гіпотезу сформулюємо в такий спосіб: $H_0: \mu = \mu_0$; $H_1: \mu < \mu_0$.

У цьому випадку так само, як і в попередньому, виберемо деякий μ_1, такий, що $\mu_1 < \mu_0$. Використовуємо результати побудови критерію Неймана-Пірсона для відповідного випадку та висновки попереднього прикладу.

Поступово найбільш сильний критерій розміру α для цього завдання задається критичною множиною, що визначається нерівністю

$$\sum_{i=1}^{n} x_i \geq \mathrm{C} = n\mu_0 + u_{1-\alpha}\sigma\sqrt{n}$$

3. Розглянемо перевірку простої гіпотези проти складної щодо математичного сподівання μ нормального закону розподілу з відомою дисперсією σ^2.

Причому нульову та альтернативну гіпотези сформулюємо в такий спосіб: $H_0: \mu = \mu_0$; $H_1: \mu \neq \mu_0$.

Для завдання критичної множини розглянемо статистику

$$\frac{\bar{X} - \mu_0}{\sigma} = \sqrt{n}$$

Статистика має нормальний розподіл. Критична множина перевірки гіпотези визначається

$$\frac{|\bar{X} - \mu_0|}{\sigma} = u_{1-\frac{\alpha}{2}}$$

Тут $u_1 - \alpha/2$ – квантиль нормального розподілу порядку $1 - \alpha/2$.

4. Розглянемо перевірку простої гіпотези проти складної щодо математичного сподівання μ нормального закону розподілу з невідомою дисперсією.

Нульову та альтернативну гіпотези сформулюємо наступним чином: $H_0: \mu = \mu_0; H_1: \mu > \mu_0$

Для завдання критичної множини розглядаємо статистику, яка має розподіл Стьюдента з $n - 1$ ступенями свободи

$$\frac{\bar{X} - \mu_0}{S(\overrightarrow{X_n})} = \sqrt{n}$$

Розподіл Стьюдента розглянуто у далі.

Критерій рівня значущості α задається критичним множиною

$$\frac{\bar{X} - \mu_0}{S(\overrightarrow{X_n})}\sqrt{n} \geq t_{1-\alpha}(n-1)$$

тут $t_{1-\alpha}(n-1)$– квантиль розподілу Стьюдента порядку $1 - \alpha$ та з $n - 1$ ступенями свободи.

Критерій згоди $\chi 2$ (критерій Пірсона)

Досі розглядаються методи перевірки статистичних гіпотез які належали до гіпотез з відомим видом розподілу. Тобто апріорі було зроблено деякі припущення про вид статистичної моделі. Зокрема, для всіх завдань, що розглядаються тут, передбачався нормальний закон розподілу.

Саме припущення про вид розподілу також є статистичною гіпотезою, а саме непараметричною статистичною гіпотезою, яка також може бути перевірена на основі статистичних даних. Для перевірки непараметричних гіпотез будують критерії згоди.

Критерій Пірсона, або критерій згоди χ^2 (Хі-квадрат), – найчастіше вживаний критерій згоди перевірки гіпотези про закон розподілу. Розглянемо критерій згоди для простої гіпотези.

Позначимо через X досліджувану випадкову величину. Нехай потрібно перевірити гіпотезу про те, що ця випадкова величина підпорядковується деякому закону розподілу $F(x)$. Зробимо вибірку, що складається з n незалежних спостережень над випадковою величиною X. За вибіркою побудуємо емпіричний розподіл $F^*(x)$ досліджуваної випадкової величини.

Для перевірки критерію вводимо статистику χ^2, яку визначим наступним чином:

$$\chi^2 = N \sum_{i=1}^{N} \frac{(p_i^{emp} - p_i^{theor})^2}{p_i^{theor}}$$

$p_i^{theor} = \int_{x-1}^{x} f(x)dx$ - передбачувана ймовірність влучення випадкової величини в $i-$ий інтервал.

p_i^{emp} – відповідне емпіричне значення.

n – число елементів вибірки з i-того інтервалу,

N – повний обсяг вибірки.

X - випадкова величина, отже статистика $\chi 2$ також є випадковою величиною. $\chi 2$ повинна підкорятися розподілу «хі-квадрат».

Правило критерію. Якщо отримана статистика $\chi 2$ перевищує квантиль закону розподілу $\chi 2$ заданого рівня значимості α з $l = (k - p - 1)$ ступенями свободи, де k – число спостережень, p – число оцінюваних параметрів закону розподілу, то гіпотеза відкидається. В іншому випадку гіпотеза приймається на заданому рівні значущості.

Застосування правила критерію зводиться до такого.

1. З вибіркових даних $x_1, x_2, \dots, x_n$ знаходять оцінки параметрів теоретичного розподілу.

2. Обчислюють за теоретичним розподілом ймовірності попадання випадкової величини в $i-$ті інтервали (P_i^{theor}).

3. Розраховують значення статистики $\chi 2$.

4. Визначають кількість ступенів свободи l.

5. Вибирають рівень значущості, зазвичай, 0,05 чи 0,01.

6. За таблицями знаходять квантиль розподілу «хі-квадрат» $\chi^2_{1,\alpha}$

7. Якщо статистика $\chi 2$ більша за $\chi^2_{1,\alpha}$,то гіпотеза відкидається при рівні значимості α.

Критерієм злагоди $\chi 2$ не можна користуватися при невеликих обсяги вибірки.

1.5. Деякі розподіли випадкових величин

Розподіл Стьюдента

Наведемо деякі розподіли випадкових величин, які використовуються при побудові оцінок та критеріїв перевірки.

Щільність розподілу Стьюдента визначається так:

$$S(t,n) = B_n\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} \qquad B_n = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\,\Gamma((\text{т}-1)/2)} \tag{5.1}$$

де n - обсяг вибірки. Іноді задають не n, а кількість ступенів свободи $k = n - 1$. $\Gamma(n)$ – гамма-функція.

Можна показати, що розподіл Стьюдента має випадкова величина, визначена як відношення стандартної нормальної сукупності до квадратного кореня із незалежної випадкової величини, що підпорядковується закону розподілу $\chi 2$. Так, якщо X_1 – випадкова величина, що має стандартний нормальний розподіл (тобто. нормальний розподіл з математичним сподіванням $\mu = 0$ і середнім квадратичним відхиленням $\sigma = 1$, і X_2 - випадкова величина, що підпорядковується закону розподілу $\chi 2$ зі ступенями свободи k, то випадкова величина, визначається $Y = \frac{X_1}{\sqrt{\frac{1}{k}X_2}}$ має розподіл Стьюдента, щільність ймовірностей якого задається формулою (5.1).

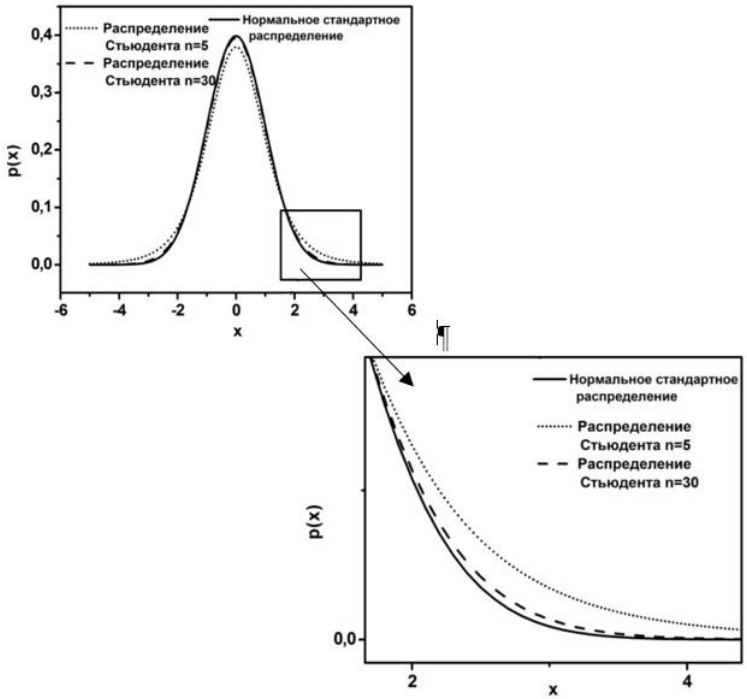


Рис. 5.1. Розподіл Стьюдента для різних значень n та нормальний стандартний розподіл (а), збільшена виділена ділянка графіка (б).

Можна помітити, що розподіл Стьюдента має більш протяжні «хвости», ніж нормальний розподіл, при $n > 30$ графіки майже зливаються.

Розподіл Стюдента використовується в математичній статистиці, зокрема, при побудові довірчих інтервалів для M(x) у разі невідомої дисперсії генеральної сукупності та при перевірці статистичних гіпотез щодо математичного сподівання за умови, що генеральна сукупність підпорядковується нормальному закону розподілу та дисперсія невідома.

Розподіл Фішера

Щільність розподілу Фішера визначається наступним чином:

$$F_{n,m}(x) = \begin{cases} C\dfrac{x^{\frac{n}{2}-1}}{\left(1+\frac{nx}{m}\right)^{\frac{n+m}{2}}}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (5.2)$$

тут n і m - ступені свободи, C - нормувальна константа, визначається як

$$C = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{B\left(\frac{n}{2},\frac{m}{2}\right)}$$

де $B(x,y)$ – бета-функція Ейлера чи інтеграл Ейлера I роду.

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

Розподіл Фішера має випадкова величина - відношення двох незалежних випадкових величин, що мають розподіл $\chi 2$ зі ступенями свободи m і n.

Так, якщо випадкова величина X_1 має розподіл $\chi 2$ з числом ступенів свободи n, а випадкова величина X_2 має розподіл $\chi 2$ з числом ступенів свободи m, величини X_1 і X_2 незалежні, то випадкова величина $X = \frac{mX_1}{nX_2}$ підпорядковується закону розподілу $F_{n,m}$, що задається формулою (5.2).

Порядок індексів n та m у записі $F_{n,m}$ суттєвий. Першим записується індекс, що відповідає числу ступенів свободи випадкової величини, що знаходиться в чисельнику (див. рис. 5.2).

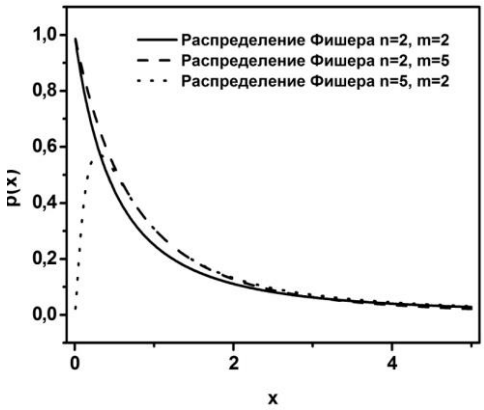


Рис. 5.2. Розподіл Фішера для різних значень n і m: $F_{2,2}$, $F_{2,5}$ та $F_{5,2}$.

Література до 1 частини

1. Математическая статистика: учеб. для вузов / В. Б. Горяинов, [и др.]; под ред. В. С. Зарубина, А. П. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. – (Сер. Математика в техническом универси- тете; Вып. XVI), С. 18–31.

2. Гмурман, В. Е. Теория вероятностей и математическая стати- стика: учеб. пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – С. 187–191.

3. Фигурин, В. В. Теория вероятностей и математическая стати- стика: учеб. пособие. / В. В. Фигурин, В. В. Оболонкин // – Минск: ООО «Новое знание», 2000. – С. 132–133.

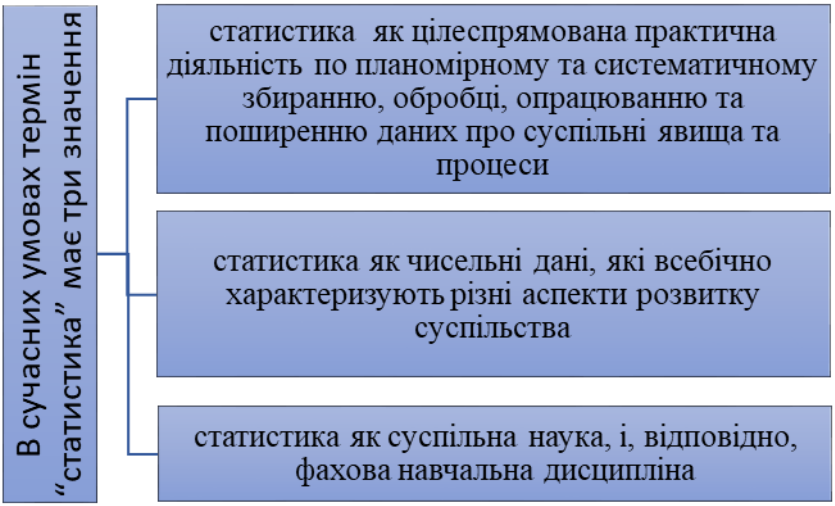
4. Гусак, А. А. Высшая математика: учебник для студентов вузов. В. 2 т. Т. 2 / А. А. Гусак. – Минск: ТетраСистемс, 2009. – С. 395–400.

5. Письменный, Д. Т. Конспект лекций по теории вероятностей, ма- тематической статистике и случайным процессам / Д. Т. Письменный. – 3-е изд. – М.: Айрис-пресс, 2008. – С. 212–218.

ЧАСТИНА 2. ПРИКЛАДНА СТАТИСТИКА

2.1. Предмет, метод і завдання статистики

Предмет і метод статистики. Основні категорії статистичної науки.

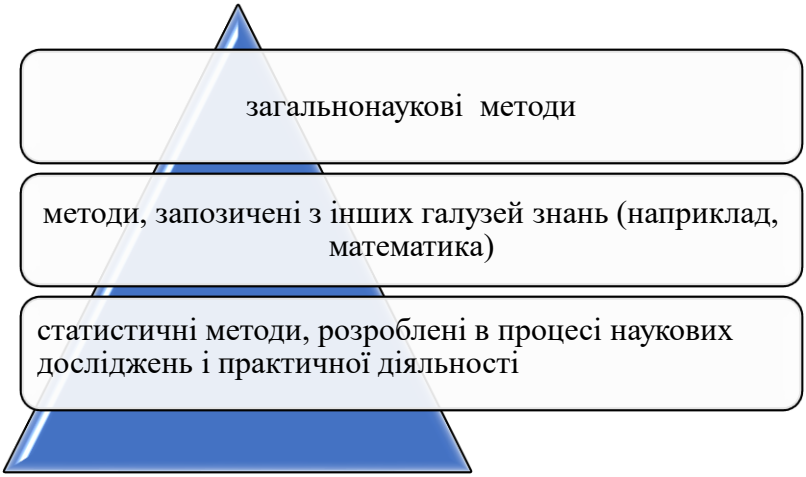


Статистика – це суспільна наука, яка вивчає кількісний бік масових суспільних явищ і процесів із врахуванням їх якісного змісту, місця і часу перебігу.

Це самостійна галузь знань об'єктом і предметом дослідження котрого виступають масові кількісні, якісно визначені параметри явищ та процесів в розрізі часу та простору.

Як і всяка дисципліна математичного характеру статистика має глибоке наукове підґрунтя, систему показників та понять на яких будується методика статистичного аналізу.

Наукові засади статистики



Категорії статистики.

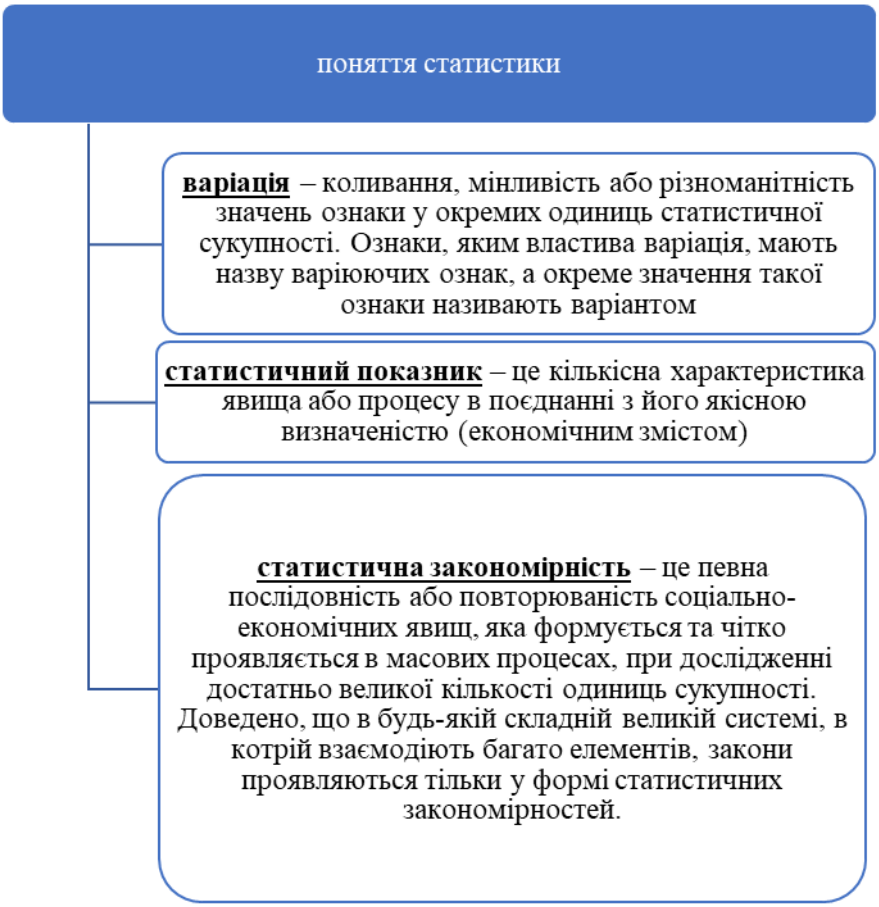
Статистикою, як будь-якою іншою галуззю знань, вироблені специфічні категорії, тобто концептуальні поняття.

- **статистична сукупність** – достатньо велика кількість елементів або явищ суспільного життя, котрі поєднуються певними зв'язками та мають як спільні (загальні), так й індивідуальні риси або властивості;
- **одиниця сукупності** - окремий елемент або явище, котрий разом з іншими утворює статистичну сукупність
- **ознака** – властивість, характерна риса або особливість одиниці сукупності, яку можна спостерігати та вимірювати (оцінювати).

Якісні ознаки не мають чисельного вираження – стать, свіжість, колір тощо. Кількісні ознаки можуть бути дискретними – тонна, дюйм, кв.метр, ціна. Можуть приймати інтервальний

вираз. Використовуються також альтернативні ознаки: так-ні, + і -.

Наступний ряд понять статистики.



Види закономірностей які можуть розглядатись:

- зміни у часі (динаміка);
- за певною ознакою;
- зміни складу та структури;
- взаємозв'язку.

Статистичний аналіз ґрунтується на законі великих чисел та понятті нормального розподілу

Сучасна організація статистичної діяльності.

Більшість країн світу організацію статистичної системи проводять за двома принципами – централізована або децентралізована. При цьому центральний статистичний орган держави може мати чисто формальний характер, як то у Великобританії, Франції та ін. В такому випадку такий орган виконує координаційні функції для регіональних органів статистики.

Неповний список органів статистики Європи:

- Австрія, Центральний статистичний офіс
- Албанія, Інститут статистики
- Андорра, Департамент національної статистики
- Бельгія, Національний інститут статистики
- Білорусь, Національний статистичний комітет
- Болгарія, Національний інститут статистики
- Босня і Герцоговина, Агенство зі статистики
- Великобританія, Національний статистичний офіс
- Гренландія, Національна статистика
- Данія, Національна статистика
- Естонія, Державний статистичний офіс
- Ірландія, Центральний статистичний офіс
- Ісландія, Національна статистика
- Іспанія, Національний інститут статистики (INE)
- Італія, Національний інститут статистики (ISTAT)
- Кіпр, Міністерство фінансів
- Латвія, Центральне статистичне бюро
- Литва, Національна статистика

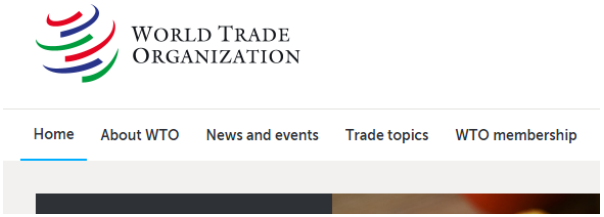
В Україні центральним статистичним органом є Державна служба статистики України:

https://www.ukrstat.gov.ua/

Існують організації міжнародної статистики:

Євростат https://ukrstat.gov.ua/work/stk_u/mo_u.htm

Статистична база даних Всесвітньої торгівельної організації (WTO) https://www.wto.org/



Статистичний відділ ООН https://unstats.un.org/UNSDWebsite/ та багато інших.

2.2. Статистичне спостереження

Статистичне спостереження

Статистичне спостереження – перша стадія статистичного дослідження, від якої в значній мірі залежить якість проведення всього дослідження.

Статистичне спостереження – планомірний, науково організований процес збирання даних щодо масових явищ та процесів, які відбуваються в економічній, соціальній та інших сферах життя України та її регіонів, шляхом їх реєстрації за спеціальною програмою, розробленою на основі статистичної методології.

Основні форми організації спостереження — звітність та спеціально організовані спостереження.

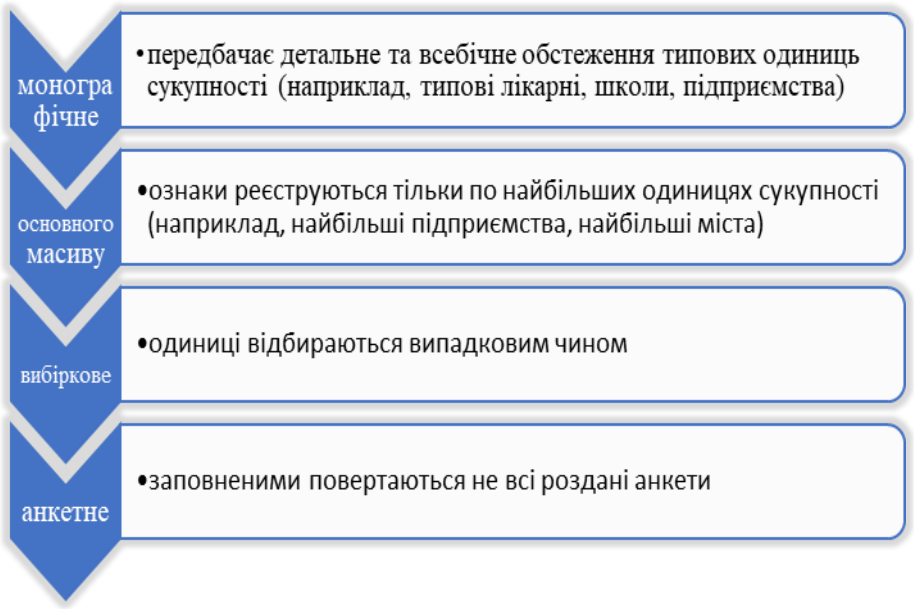
Основним завданням статистичного спостереження є одержання об'єктивної, достовірної та повної інформації, яка характеризує кожну одиницю досліджуваної сукупності. Державні статистичні спостереження проводяться органами державної статистики відповідно до затвердженого Кабінету Міністрів України плану державних статистичних спостережень, або за окремими рішеннями КМ України.

Статистичний звіт — це документ, який вміщує систему показників діяльності суб’єктів господарювання. Зміст звіту, форма та термін подання в органи державної статистики затверджуються Держкомстатом України.

Первинним джерелом даних є статистична звітність, яка охоплює підприємства, організації та установи всіх форм власності, згідно з класифікатором видів економічної діяльності та ЄДРПОУ.

Форма статистичної звітності містить такі реквізити:

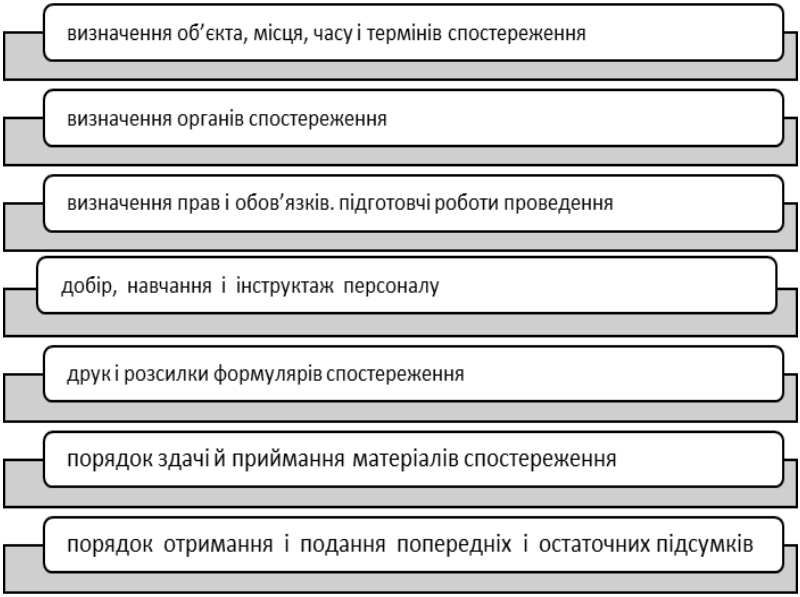
Види та способи проведення спостереження.



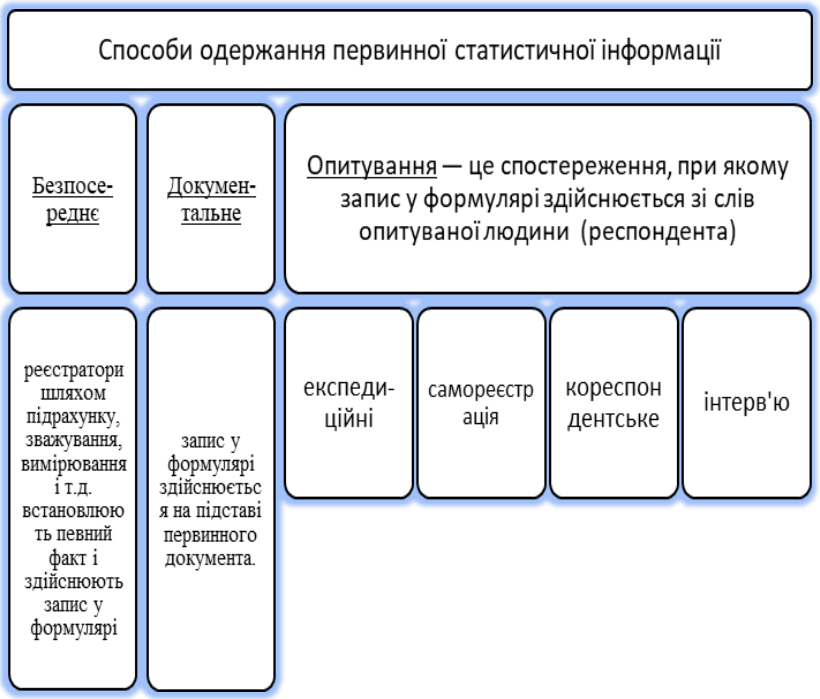
За моментом реєстрації даних спостереження поділяються на поточні, періодичні та одноразові.

У статистичній практиці використовують суцільні (щодо всіх одиниць сукупності без винятку) та несуцільні(охоплення неповне, а результати знаходяться за допомогою математичних методів) спостереження.

Організаційні питання статистичного спостереження



Статистичне спостереження де реєстрація даних проводиться по мірі їх одержання називається поточним (присвоєння ІДН, ID картка, народжуваність, смертність тощо). Вирізняють також періодичні та одноразові спостереження.



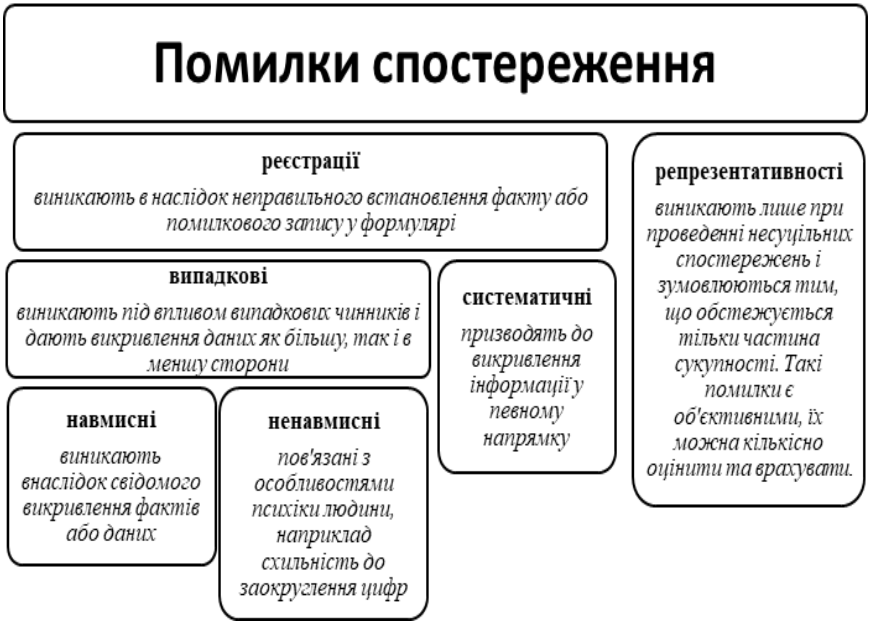
Помилки спостереження та контроль його результатів

Точністю статистичного спостереження називають ступінь відповідності значення будь-якої ознаки, визначеної за допомогою статистичного спостереження, її дійсному значенню. Чим ближчі значення ознак, отриманих в результаті статистичного спостереження, до їх фактичних значень, тим точніше проведено спостереження.

При проведенні вибіркових спостережень трапляються помилки. Помилки можуть бути зумовлені як неправильно вибраною методикою спостереження так і людським фактором.

Це призводить до невірних висновків при аналізі результатів статистичного спостереження або до їх викривлення. Після аналізу результатів, такі помилки можуть бути виправлені.

Більш небезпечними вважаються похибки пов’язані з людьми. Це можуть бути помили через некомпетентність персоналу, який проводить спостереження, або спеціальні вкиди або викривлення інформації для отримання вигоди. Таке часто трапляється під час виборів.



Наприкінці узагальнимо. При проведенні будь якого статистичного дослідження реалізується певна послідовність етапів.

ЕТАПИ СТАТИСТИЧНИХ ДОСЛІДЖЕНЬ

2.3. Зведення і групування

Зведення як друга стадія статистичного дослідження. Суть та види зведення.

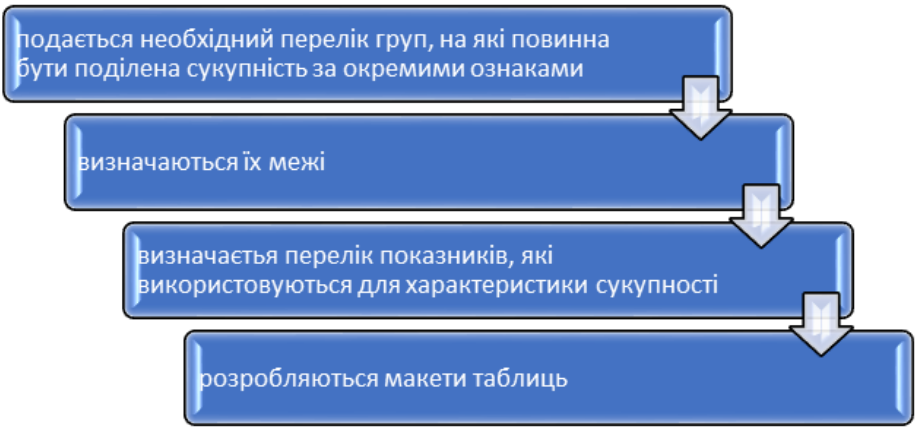
Після отримання первинної інформації завдання дослідження полягає у їх систематизації та одержанні зведених, узагальнених характеристик сукупності в цілому.

Статистичним зведенням називається спеціальна обробка первинних даних статистичного спостереження з метою отримання узагальнюючих характеристик досліджуваного явища чи процесу за рядом суттєвих для них ознак.

Фактично – зведення це етап переходу від дійсних значень показників, які ми отримали в результаті статистичного спостереження до загальних показників, тобто сукупності в цілому.

Це відповідальний етап, тому для його проведення розробляється план та програма.

Програма зведення



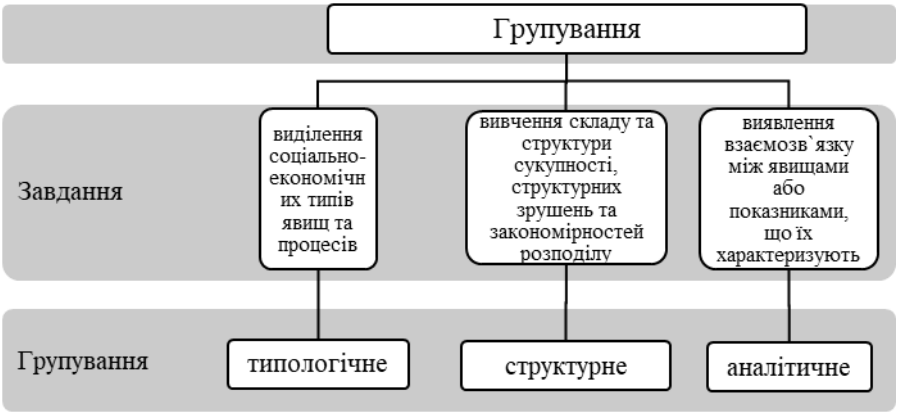
В залежності від централізованої чи децентралізованої форми зведення матеріали статистичного спостереження надходять у Держкомстат, або до територіальних органів статистики, де вони обробляються та систематизуються.

В простих зведеннях узагальнюючі показники знаходяться як підсумки таблиць, та кількість таблиць невелика, а часом взагалі одна.

Групування, його суть, завдання та види.

Статистичним групуванням називають розчленування, розподіл одиниць сукупності на класи, групи та підгрупи за суттєвими ознаками. Та ознака, що покладена в основу групування, тобто за якою утворюються групи, має назву групувальної.

З допомогою групувань вирішують різні типи завдань, при цьому використовуються різні види групувань.



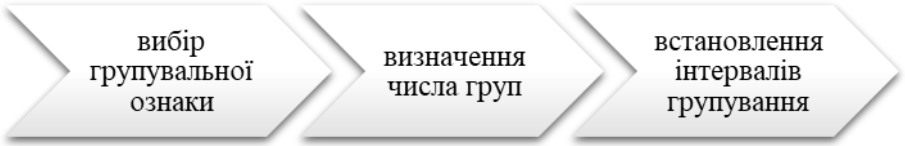
Крім цього, групування класифікуються за видами групувальних ознак, їх кількості та співвідношення.

Складніші зведення потребують розбиття генеральної сукупності на підгрупи за певною ознакою, та знаходження підсумкових показників к по підгрупах, так і по сукупності в цілому.

Вимоги до групування:

Інтервал групування.

Методологія групування



Вибір групувальної ознаки здійснюється після детального аналізу суті та закону розвитку явища, що досліджується. Головна вимога до групувальної ознаки – її значимість, вагомість.

Особливу увагу треба звернути на типологічні групування, які враховують не тільки кількісні, а також і якісні характеристики одиниці сукупності.

Загальний принцип ділення на кількість груп:

Ознака	Кількість груп
Атрибутивна	Число варіантів ознаки
Альтернативна	Дві
Дискретна	Число варіантів цієї ознаки
Інтервальна	В залежності від типу інтервалів.

Інтервал групування – це проміжок між двома значеннями групувальної ознаки, в межах яких всі одиниці сукупності відносяться до даної групи. Відповідно менше та більше число мають назву нижньої та верхньої межі інтервалу групування, а різниця між ними називається величиною інтервалу.

Види інтервалів

Прикладом спеціалізованого групування може служити групування перелітних птахів за видами.

При групуванні з рівними інтервалами, величина інтервалу визначається за формулою:

$$i = \frac{X\,max - Xmin}{m},$$

де: $Xmax$ та $Xmin$ – відповідно найбільше та найменше значення групувальної ознаки; m – число груп.

Орієнтовно число груп можна визначити за формулою Стерджесса: $m = 1 + 3{,}332\ell gn.$

Вторинні групування.

Групування виконане на основі вже існуючого називається **вторинним**.

Кількість студентів 1 курс	Кількість факультетів
25 - 50	10
51 - 75	27
76 - 100	27
101 - 125	15
126 - 150	18
151 - 175	14
176 - 200	13
201 - 225	18
226 - 250	14
Разом	156

Це можна зробити за так:

- зміною інтервалів групування;
- перегрупуванням за питомою вагою груп.

1 спосіб.

Розглянемо на прикладі перший спосіб вторинного групування. З попереднього прикладу. Необхідно утворити три групи:

Кількість студентів 1 курс	Кількість факультетів
До 100	10+27+27=64
100-200	15+18+14+13=60
Більше 200	18+14=32
Разом	156

2 спосіб.

Створити три групи по 33,3% факультетів (52шт).

Так як до першої групи, щоб обрати 52 факультети, обираємо 10+27 з перших двох інтервалів та 15 з третього.

Тоді межі інтервалів:

$1-й\ \ 76+\frac{(100-76)*15}{27}=71+13{,}3\approx 84$

$2-й\ \ 151+\frac{(175-151)*7}{52+14}=151+2{,}5\approx 154.$

Відповідно

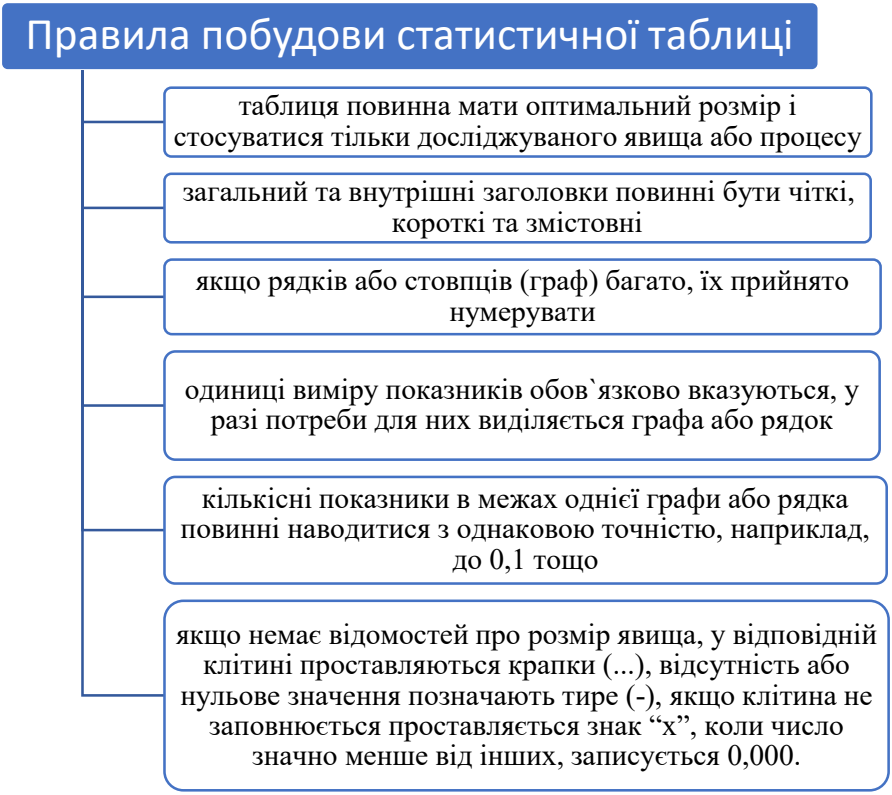
$3-й\ інтервал\ \ 155-200$

Отже, маємо нове групування:

Число студентів	Число факультетів	Питома вага, %
До 84 включно	52	33,3
84-154	52	33,3
155-200	52	33,3
Разом	**156**	**100**

Статистичні таблиці.

Статистичні таблиці застосовуються для систематизованого, раціонального та наочного викладення результатів групування. У статистичній таблиці розрізняють підмет (заголовки рядків)та присудок (заголовки стовпчиків).



2.4. Ряди розподілу.

Поняття про ряди розподілу. Види рядів розподілу

Після групування одержують цифрові показники, які характеризують розподіл сукупності за однією з ознак - ряди розподілу (РР).

Ряд розподілу складається з двох елементів:

- варіанти (x) – послідовно розміщених окремих значень групувальної ознаки;
- частоти (f) – кількості "попадань" певного значення ознаки у сукупності, або групі.

РР поділяють на атрибутивні та варіаційні (кількісні). Групувальна ознака атрибутивного ряду має нечисельний вираз. Наприклад, аналіз респондентів по відношенню гендерної рівності:

Рівні	. Кількість, чол
За	580
Індеферентні	130
Проти	90
Разом	800

Групувальна ознака варіаційного ряду має чисельний вираз. Такі ряди бувають дискретними:

Всі кімнати гуртожитку	50
кімнати, в яких живе:	
2 особи	20
3 особи	11
4 особи	9
5 і більше осіб	10

та інтервальними:

Заробітна плата, грн.	Число робітників
До 5000	30
5000 – 6000	56
6000 – 7000	77
7000 – 10000	65
Разом	260

Замінюючи інтервал в інтервальному ряду на його середину

$$X^* = \frac{X^{\text{верх межа інтерв}} + X^{\text{ниж межа інтерв}}}{2}$$

ми перетворюємо його на дискретний.

Наприклад, середина інтервалу 5000 – 6000 грн. становить 5500 грн. (5000+6000/2 =5500).

Групування за двома і більше ознаками. Комбінаційні групування дозволяють отримати більш детальну інформацію про структуру та закономірності розподілу, отримати більш якісний його аналіз. Наприклад:

Заробітна плата		Кваліфікаційний розряд						Разом
		1	2	3	4	5	6	
814	1149,125	0	1	1	0	0	0	2
1149,125	1484,25	6	0	0	0	0	1	7
1484,25	1819,375	4	0	0	0	0	0	4
1819,375	2154,5	1	11	0	0	0	0	12
2154,5	2489,625	0	0	31	0	0	0	31
2489,625	2824,75	0	0	0	45	0	0	45
2824,75	3159,875	0	0	0	0	31	0	31
3159,875	3495	0	0	0	0	0	27	27
Разом		11	12	32	45	31	28	159

Правила побудови рядів розподілу. Види частот

При побудові рядів групувальну ознаку ранжують. Дуже важливим є розмежування варіанта-інтервал.

Розподіл призовників за ростом:

170 – 174	170,0 – 174,9	До 175 см
175 – 179	175,0 – 179,9	175 - 180
180 – 184	180,0 – 184,9	180 - 185

Середина інтервалу у кожному випадку дорівнює:

$$\frac{170+174}{2} = 172 \quad \frac{170{,}0+174{,}9}{2} = 172{,}45 \quad \frac{170+175}{2} = 172{,}5$$

Розрізняють три види частот, які використовуються для аналізу рядів розподілу:

- абсолютна (кількість);
- відносна (частка);
- нагромаджена (сума часток).

Наприклад, розподіл автомобілів в автотранспортному підприємстві за строком експлуатації:

Вік, років	Число автомобілів	Частка	Питома вага
До 6	20	0,15	15,38%
6-10	34	0,26	26,15%
10-14	53	0,41	40,77%
14-18	18	0,14	13,85%
18 і більше	5	0,04	3,85%
Разом	130	1,00	100,00%

Наприклад, розподіл автомобілів в автотранспортному підприємстві за строком експлуатації:

Вік, років	Число автомобілів	Нагромаджені частоти	
		шт	%
До 6	20	20	15%
6-10	34	54	42%
10-14	53	107	82%
14-18	18	125	96%
18 і більше	5	130	100%
Разом	130		

Таким чином, 54 автомобіля або 42% мають строк експлуатації менше 10 років, а 125 автомобілів або 96% мають строк експлуатації менше 18 років.

Щільність розподілу – це кількість одиниць сукупності, що припадає на одиницю величини інтервалу

$f_d = \frac{f}{i}$, де f – частота; i – величина інтервалу.

Розрізняють абсолютну та відносну щільність розподілу.

Орендна плата квартир:

Місячна орендна плата, $.	Число квартир		Щільність	
	шт.	%	шт.	%
До 200	50	25,00	0,125	0,0625
200-300	48	24,00	0,24	0,12
300-500	32	16,00	0,08	0,04
500-1000	40	20,00	0,04	0,02
1000 і більше	30	15,00	0,075	0,0375
Разом	200		x	x

Максимальну щільність розподілу мають квартири ціною $ 200-300.

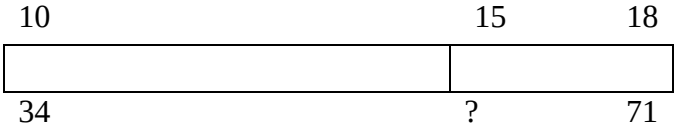
Інтерполяція в рядах розподілу

Інтерполяція в рядах розподілу визначає, скільки одиниць сукупності (або частка) мають значення ознаки, менше від заданого. Для інтерполяції використовують як абсолютні, так і відносні нагромаджені частоти.

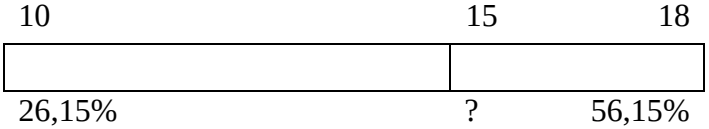
Наприклад, маємо ряд розподілу автомобілів на автотранспортному підприємстві за строком експлуатації:

Строк експлуатації, рік	Число автомобілів		Щільність розподілу	
	од.	%	од.	%
До 6	20	15,38	5,00	3,85
6-10	34	26,15	8,50	6,54
10-18	71	54,62	6,83	0,85
18 і більше	5	3,85	1,25	0,96
Разом	130	100,00	X	X

Визначимо, скільки автомобілів мають строк експлуатації менше 15 років. Графічна інтерпретація:



$$f_{\angle 15} = 34 + \frac{15-10}{18-10} * (71-34) = 57{,}13$$



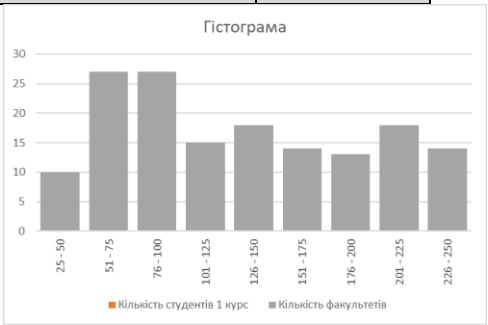
$$f_{\angle 15} = 26{,}15 + \frac{15-10}{18-10} * (56{,}15-26{,}15) = 44{,}9\%$$

Графічне зображення рядів розподілу

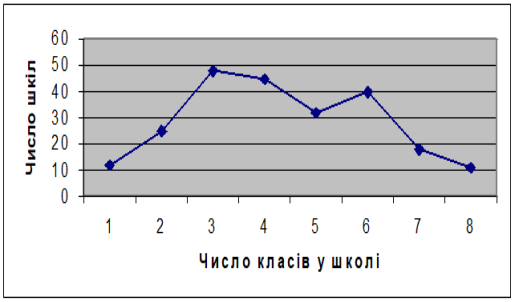
Для візуально-графічного аналізу розподілу використовують наступні графіки і діаграми .

Гістограма будується для інтервальних рядів розподілу. При цьому по осі Х відкладаються інтервали групування, а по осі У – абсолютні або відносні частоти.

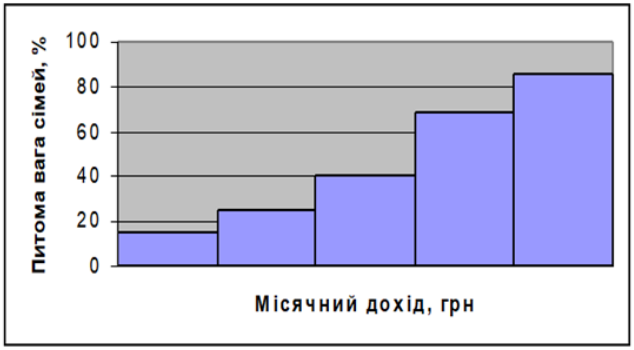
Кількість студентів 1 курс	Кількість факультетів
25 - 50	10
51 - 75	27
76 - 100	27
101 - 125	15
126 - 150	18
151 - 175	14
176 - 200	13
201 - 225	18
226 - 250	14
Разом	**156**



Полігон використовується для графічного зображення дискретних та атрибутивних рядів розподілу. Це лінійний графік, при цьому по осі Х відкладаються значення варіант, а по осі У – частоти. Гістограму можна перетворити у полігон, з`єднавши відрізками прямої середини верхівок стовпчиків.



Кумулята призначена для графічного подання рядів розподілу з нагромадженими частотами. Це може бути стовпчикова діаграма (для дискретного та атрибутивного рядів розподілу – лінійний графік). Будується вона аналогічно попереднім графікам, тільки по осі У подаються нагромаджені частоти.

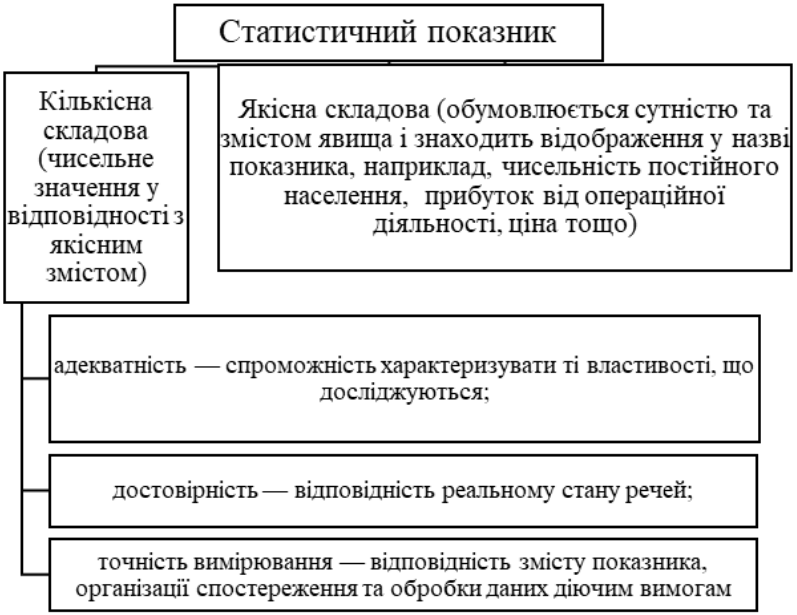


2.5. Абсолютні та відносні величини

Статистичні показники, їх суть та види

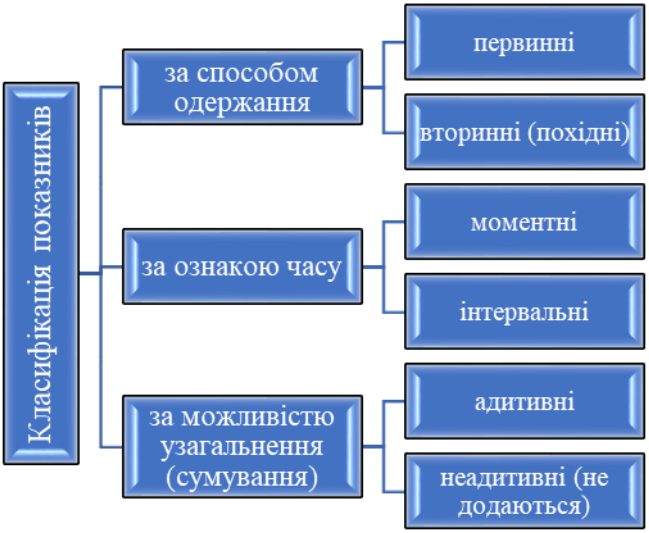
Статистична інформація, яку одержують у процесі статистичного дослідження, являє собою сукупність статистичних показників.

Статистичний показник — це одна з основний категорій статистики, яка характеризує суспільні явища та процеси у поєднанні кількісної та якісної визначеності.



Існують різні класифікації статистичних показників.

Однією з найбільш повних на наш погляд є класифікація наведена нижче.



Абсолютні величини, їх види та одиниці виразу

Умовно-натуральні одиниці застосовуються для підсумовування натуральних одиниць, різновидів об'єктів одного і того ж типу. Наприклад, різні види палива (нафта, газ, вугілля тощо) перетворюються на тони умовного палива зі стандартною теплотою згоряння 29 мДж/кг, алкогольні напої – у літри умовного 100% спирту, продукція консервної промисловості – в умовні консервні банки зі стандартною ємністю 0,33 л, різні види шкільних зошитів – в умовні шкільні зошити стандартним обсягом 12 аркушів тощо.

Грузопідйомність автомобіля, тон	Кількість автомобілів, шт.	Перехідний коефіцієнт	Кількість в уно
12	10	1	10
18	20	1,5	30
30	30	2,5	75
Разом	**60**	**x**	**105**

Складні одиниці виміру – це комбінація простих як кінські сили, кВт-год тощо. Перехід від натуральних до умовно-натуральних показників за допомогою перехідних коефіцієнтів.

$$K_1 = \frac{12}{12} = 1{,}0; \quad K_2 = \frac{18}{12} = 1{,}5; \quad K_3 = \frac{30}{12} = 2{,}5. \qquad K = K_1 \cdot 10 + K_2 \cdot 20 + K_3 \cdot 30 = 105$$

Отже, загальна грузопідйомність дорівнює 105 автомобілів в перерахунку на автомобілі з навантаженням 12 тон.

Вартісні одиниці (гривні, рублі, долари, євро та інша валюта) застосовуються для обчислення шляхом грошової оцінки об'єктів різного типу (наприклад, продукції або витрат на її виробництво всіх чи кількох видів), оскільки вони не піддаються підсумовуванню у натуральних одиницях.Наприклад, немає сенсу підсумовувати кілограми риби і м'яса, оскільки в результаті вийде "ні риба, ні м'ясо".

Відносні величини

В залежності від поставленої задачі відносні показники знаходяться або коефіцієнтом, або відсотками.

$$ВВ = \frac{А}{Б} \qquad ВВ = \frac{А}{Б} \cdot 100\%$$

Відносними величинами називають такі показники, які виражають кількісне співвідношення між ознаками, що характеризують досліджувані явища та процеси.

Відносна величина використовується в тому випадку, коли треба охарактеризувати:

- в скільки разів одна ознака більша або менша за іншу;
- яку частину становить одна ознака відносно іншої;
- скільки одиниць однієї ознаки припадає на 1, 1000, 10000 одиниць іншої ознаки.

Чисельник (А), називається порівнюваною величиною, а знаменник (Б) — базою порівняння.

Співставність	Формула	Приклад
А ≅ Б	Відсотки $ВВ = \frac{А}{Б} \times 100$	У січні реалізовано продукції на 123000 грн., а у лютому на 184000 грн. $\text{ВВ} = \frac{123000}{184000} \times 100 = 66{,}8\%$. Таким чином, у січні обсяг реалізації продукції становив 66,8% порівняно із лютим.
А ≫ Б А ≪ Б	Проміле $ВВ = \frac{А}{Б} \times 1000$	на підприємстві 7500 робітників та 320 службовців. $ВВ = \frac{320}{7500} \times 1000 = 42{,}4^0/_{00}$ Отже 1000 робітників обслуговують 42,4 службовців.
А ⋙ Б А ⋘ Б	Продецеміле $ВВ = \frac{А}{Б} \times 10000$	Ковідом захворіло 1230 осіб., а в відповідних лікарнях працює 28,5 тис. лікарів. $\text{ВВ} = \frac{28{,}5}{1230} \times 10000 = 234^0/_{000}$ – 234 лікаря на 10000 захворівших
	Просантиміле $ВВ = \frac{А}{Б} \times 100000$	аналогічно

В економічних розрахунках використовуються три основних відносних показника.

Відносна величина планового завдання показує, у скільки разів або на скільки процентів запланований рівень показника більший чи менший фактично досягнутого рівня:

$$ВВ_{пз} = \frac{П_1}{Ф_0} \text{ або } ВВ_{пз} = \frac{П_1}{Ф_0} x100,$$

де $П_1$ та $Ф_0$ — відповідно планове значення показника у наступному та фактичне значення у попередньому періоді.

Відносна величина виконання плану характеризує, у скільки разів або на скільки процентів фактичне значення показника більше або менше запланованого:

$$ВВ_{пз} = \frac{Ф_1}{П_1} \text{ або } ВВ_{пз} = \frac{Ф_1}{П_1} x100.$$

де $П_1$ та $Ф_1$ — відповідно планове значення показника у наступному та фактичне значення у наступному періоді.

Відносна величина динаміки характеризує зміну показника у часі і визначається як відношення значення у наступному періоді до величини у попередньому періоді:

$$ВВ_{д} = \frac{Ф_1}{Ф_0} \text{ або } ВВ_{д} = \frac{Ф_1}{Ф_0} x100.$$

де $Ф_1$ та $Ф_0$ — відповідно фактичне значення показника у наступному та фактичне значення у попередньому періоді.

Між названими видами відносних величин, якщо вони підраховані за одними даними, існує даними, існує взаємозв'язок:

$$ВВ_{д} = ВВ_{пз} * ВВ_{зп} = \frac{П_1}{Ф_0} \times \frac{Ф_1}{П_1} = \frac{Ф_1}{Ф_0}.$$

Рік	Виробництво, тис.грн			Відхилення, тис.грн	ВПЗ	ВВП	ВД
	1 півріччя	2 півріччя					
		План	Факт				
2011	1387	1473	1430	-43	106,20%	97,08%	103,10%
2012	1443	1304	1373,5	69,5	90,37%	105,33%	95,18%
2013	1453	1335	1394	59	91,88%	104,42%	95,94%
2014	1314	1457	1385,5	-71,5	110,88%	95,09%	105,44%
2015	1378	1390	1384	-6	100,87%	99,57%	100,44%
2016	1485	1617	1551	-66	108,89%	95,92%	104,44%
2017	1371	1274	1322,5	48,5	92,92%	103,81%	96,46%
2018	1500	1458	1479	21	97,20%	101,44%	98,60%
2019	1353	1518	1435,5	-82,5	112,20%	94,57%	106,10%
2020	1348	1474	1411	-63	109,35%	95,73%	104,67%
2021	1426	1570	1498	-72	110,10%	95,41%	105,05%

Дещо рідше використовуються такі відносні величини:

Відносна величина структури характеризує співвідношення частини та цілого. Вона показує, яку частину або скільки процентів становить частина від загального підсумку. Якщо ця відносна величина визначається у вигляді коефіцієнту, вона називається часткою, а якщо у процентах — питомою вагою.

$$ВВ_c = \frac{Частина}{Ціле(Сумачастин)}$$

Наприклад, підприємство виготовило 5600 чоловічих костюмів та 3400 жіночих. Тоді

$$ВВ_{стр} = \frac{5600}{5600+3400} = \frac{5600}{9000} = 0{,}622 \text{ або } 60{,}2\%$$

$$BB_{стр} = \frac{3400}{9000} = 0{,}398 \text{ або } 39{,}8\%.$$

Відносна величина координації показує співвідношення між окремими частинами одного цілого, при цьому одна частина приймається за базу порівняння. Вона може визначатися на 100, 1000 або 10000 одиниць знаменника.

$$BB_{к} = \frac{Частина\ 1}{Частина\ 2} \times 100(1000,10000).$$

Відносна величина порівняння — це співвідношення однойменних показників, що обчислені по різних об'єктах або територіях за однаковий час.

$$BB_{п} = \frac{Показник\ по\ А}{Показник\ по\ Б}.$$

Відносна величина інтенсивності визначається як відношення двох різних показників і переважно характеризує ступінь поширення чи розвитку явища у певному середовищі. Ця відносна величина може мати одиниці виміру вихідних показників. Широко застосовуються для аналізу економічного розвитку, демографічних, соціальних та політичних процесів.

$$BB_{інт} = \frac{Показник1}{Показник2}.$$

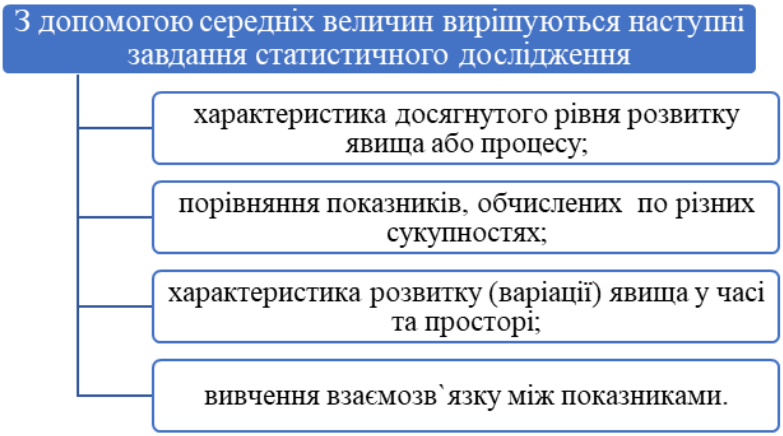
2.6. Середні величини

Суть та умови використання середніх величин

Необхідність розрахунку середньої величини виникає у порівнянні між собою кількох сукупностей (підприємств, працівників, студентів і т.д.) з різними числовими значеннями досліджуваної ознаки (наприклад, заробітної плати працівників, успішності студентів тощо).

Середня величина – це узагальнена характеристика однорідної сукупності за варіюючою ознакою, що показує типовий рівень цієї ознаки у одиниці сукупності.

Характерний, типовий рівень ознаки формується під впливом так званих систематичних (невипадкових, постійних) факторів, а відхилення індивідуальних значень від типового рівня зумовлені дією випадкових факторів, котрі впливають по-різному на окремі одиниці сукупності. Таким чином, середня величина відображає те спільне, загальне, що є характерним для усіх одиниць досліджуваної сукупності.



Вимоги при визначенні середніх величин- однорідність сукупності, та масовість її якісного показника.

Види середніх

Загальна формула знаходження середньої величини:

$$\overline{x} = \sqrt[m]{\frac{\sum x^m}{n}}$$

де x – індивідуальні значення ознаки;

n – чисельність сукупності (кількість одиниць).

В залежності від значення показника степеня m розрізняють такі середні величини:

m = -1 - середня гармонійна;

m = 0 - середня геометрична;

m = 1 - середня арифметична;

m = 2 - середня квадратична.

Середня арифметична в загальному випадку:

$$\overline{x} = \frac{Обсяг\_ознаки}{Чисельність\_сукупності}.$$

На основі індивідуальних показників

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n} = \frac{1}{n}\sum x.$$

Для згрупованих даних(дискретний ряд)

$$\bar{x} = \frac{\sum x_1 f_1 + x_2 f_2 + \ldots + x_m f_m}{f_1 + f_2 + \ldots + f_m} = \frac{\sum xf}{\sum f} = \frac{1}{\sum f}\sum xf,$$

В інтервальному ряду (x^*- середина відповідного інтервалу)

$$\overline{x} = \frac{\sum x^* f}{\sum f}$$

де x-значення ознаки,

f-частота (кількість її появ в відповідному інтервалі)

Наприклад, ряд платежів:

286, 378, 183, 295, 363, 280, 276, 292, 358, 265, 275, 373 грн.

Середній рівень:

$$\bar{x} = \frac{286 + 378 + 183 + \ldots + 275 + 373}{12} = \frac{3624}{12} = 302 \ грн$$

Приклад. Дискретний ряд

$$x = \frac{\sum x_1 f_1 + x_2 f_2 + \ldots + x_m f_m}{f_1 + f_2 + \ldots + f_m} = \frac{\sum xf}{\sum f} = \frac{1}{\sum f} \sum xf,$$

де x – варіанти; f – частоти; m – число груп.

Днів відпустки (*x*)	**Кількість робітників (*f*)**	***xf***
8	4	32
10	5	50
14	13	182
20	12	240
24	32	768
30	14	420
56	2	112
Разом	**82**	**1804**

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{1804}{82} = 22 дня.$$

Для знаходження середнього значення в інтервальних рядах використовується та ж формула середньої зваженої. Але перед цим ряд перетворюється в дискретний, де в ролі варіанти виступає середина інтервалу. В відкритих інтервалах ми «відступаємо вперед чи назад» половину закритого інтервалу.

В наступному прикладі перший і останній інтервал відкриті.

Ціна за обід в кафе:

Ціна обіду, грн.	Число (f)	Середина інтервалу (x*)	xf
До 30	4	15	60
30-60	38	45	1710
60-90	44	75	3300
90-120	34	105	3570
120-150	23	135	3105
150-180	5	165	825
180-210	2	195	390
Більше 210	1	225	225
Разом	151	x	13185

Середня ціна:

$$\overline{x} = \frac{\sum xf}{\sum f} = \frac{13185}{151} = 87{,}32грн.$$

Якщо відомі групові середні, то розрахунок загальної середньої здійснюється за формулою:

$$x = \frac{\sum \bar{X}_i f_i}{\sum f_i}$$

де $\overline{X_i}$ - групові середні величини; f_i –число у ***i***-тій групі.

Групи за зростом	**Середній ріст, см.**	**Число, чол.**	$\overline{\boldsymbol{X_i}} * \boldsymbol{f_i}$
Невисокі	162	78	12636
Середні	176	650	114400
Високі	185	97	17945
Разом	**x**	**825**	**144981**

Загальна середня дорівнює:

$$\overline{x} = \frac{\sum \bar{x}_i f_i}{\sum f_i} = \frac{144981}{825} = 175{,}73грн.$$

Властивості середньої арифметичної:

При збільшенні або зменшенні кожної частоти в к разів, середня не зміниться	• $x = \frac{\sum xf}{\sum f} = \frac{\sum \frac{xf}{k}}{\sum \frac{f}{k}}$
При збільшенні або зменшенні кожної варіанти в к разів середня зміниться в стільки ж разів	• $k \cdot \overline{x} = \frac{\sum k \cdot x \cdot f}{\sum f}$ або • $\frac{\overline{x}}{k} = \frac{\sum \frac{x}{k} \cdot f}{\sum f}$.
При збільшенні або зменшенні кожної варіанти та сталу величину А, середня зміниться на цю ж величину	• $\dot{x} - A = \frac{\Sigma(x-A)f}{\Sigma f}$ або • $+A = \frac{\Sigma(x+A)f}{\Sigma f}$
Сума відхилень значень ознаки (варіант) від середньої арифметичної дорівнює нулю	• $\Sigma(x - \dot{x}) = 0$ або • $\Sigma(x - \dot{x})f = 0.$
Середня арифметична, що помножена на чисельність сукупності, дорівнює обсягу ознаки	• $\dot{x} \cdot n = \Sigma x$ або • $\dot{x} \cdot \Sigma f = \Sigma xf.$
Сума квадратів відхилень варіант від середньої арифметичної є мінімальною величиною із всіх можливих	• $\Sigma(x - \dot{x})^2 = min$ або • $\Sigma(x - \dot{x})^2 f = min.$

В певних випадках для підрахунку середньої арифметичної в інтервальних рядах розподілу використовується метод «моментів». Підрахунок виконується так:

$$\bar{x} = m_1 \cdot i + A, m_1 = \frac{\Sigma\left(\frac{x-A}{i}\right)f}{\Sigma f},$$

де: m_1 — момент першого порядку;
i — величина інтервалу;
A — середина інтервалу з найбільшою частотою.

Розглянемо приклад розрахунку середньої арифметичної зваженої методом «моментів».

Ціна, *грн.*	Кількість, *шт.*	Середина інтервалу	$x - A$, $A = 350$	$\frac{x-A}{i}$, $i = 50$	$\left(\frac{x-A}{i}\right)f$
До 100	11	50	-300	-7,14	-78,57
100–200	22	150	-200	-4,76	-104,76
200–300	36	250	-100	-2,38	-85,71
300–400	42	350	0	0,00	0,00
400–500	19	450	100	2,38	45,24
Разом	**130**	**x**	**x**	**x**	**-223,81**

Момент першого порядку: $m_1 = \frac{\Sigma\left(\frac{x-A}{i}\right)f}{\Sigma f} = \frac{-223,81}{130} = -1,72$

Середня величина: $\bar{x} = m_1 \cdot i + A = -1,72 \cdot 100 + 350 = 177,8$ *грн.*

ромір взуття	кількість	сер.інт	$x - A$	$\frac{x-A}{h}$	$\frac{x-A}{h}f$
до 25	5	20	3	0,3	1,5
25-35	12	30	13	1,3	15,6
35-45	17	40	23	2,3	39,1
45-55	15	50	33	3,3	49,5
більше 55	1	60	43	4,3	4,3
Итого	50				

$$m_1 = \frac{110}{50} = 2,2$$

$$x = m_1 h + A = 2,210 + 17 = 22 + 17 = 39$$

Середня гармонійна величина використовується у тому випадку, якщо відомі обернені значення осереднюваного показника. У цьому разі $\bar{x} = 1/x'$, де х — значення прямого (осереднюваного) показника, x' — значення оберненого показника. Для індивідуальних (незгрупованих) даних використовується середня гармонійна проста. Для рядів розподілу зважена.

$$\bar{x} = \frac{n}{\Sigma\frac{1}{x}}. \qquad - \qquad x = \frac{\Sigma f}{\Sigma\frac{f}{x}}.$$

Частіше при розрахунках середньої величини використовується середня гармонійна у вигляді:

$$\bar{x} = \frac{\Sigma W}{\Sigma\frac{W}{x}}.$$

де: $W = xf$ — значення об'ємного показника; х — значення осереднюваного показника.

1. Якщо є дані, що відносяться до чисельника осереднюваного показника, то застосовується формула середньої гармонійної.

2. Якщо є дані, що належать до знаменника осереднюваного показника, то застосовується формула середньої арифметичної.

Цех	Зарплата працівника, *грн.*(*x*)	Фонд зарплати цеху, *грн.*
1	282	118900
2	364	53120
3	258	17980
Разом	**x**	**190000**

В такому випадку середня зарплата:

$$\bar{x} = \frac{\Sigma W}{\Sigma \frac{W}{x}} = \frac{190000}{\frac{118900}{282} + \frac{53120}{364} + \frac{17980}{258}} = \frac{190000}{637} = 298 грн.$$

Середня геометрична величина застосовується тоді, коли обсяг ознаки дорівнює не сумі, а добутку варіант. Її формула має вигляд:

$$\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}.$$

За наведеною формулою підраховується середній коефіцієнт росту, при цьому х – ланцюгові коефіцієнти росту.

Середній ріст продуктивності праці по рокам складав 1,01 1,03 0,99 0,95 1,02. Знайти середній ріст продуктивності праці.

$$\bar{X} = \sqrt[5]{1{,}01 \cdot 1{,}03 \cdot 0{,}99 \cdot 0{,}95 \cdot 1{,}02} = 0{,}9995$$

Середня характеристика ряду динаміки обчислюється як середня хронологічна. Для інтервального ряду динаміки (відображає розміри явищ за певні проміжки часу) з рівними періодами середня хронологічна це середня арифметична, а з нерівними відрізками часу – середня арифметична зважена.

Для моментного ряду динаміки (характеризує розміри економічних і суспільних явищ станом на якийсь момент, здебільшого на певну дату) середня хронологічна:

$$\bar{X} = \frac{\frac{x_1 + x_n}{2} + x_2 + x_3 + \ldots + x_{n-1}}{n - 1}$$

Надої характеризуються даними: 1.01-300л, 1.04-320л, 1.08-310л, 1.11-330л. Середнє значення:

$$\overline{X} = \frac{\frac{300+330}{2} + 320 + 310}{4-1} = \frac{945}{3} = 315л.$$

Структурні середні – мода і медіана

Мода (М0) — це значення ознаки, що найчастіше зустрічається у сукупності. Таким чином, у дискретному ряді розподілу - це варіанта, що має найбільшу частоту.

Ряд розподілу: 23,25,11,23,54,54,23,45. $Mo = 23$

Зазначимо, що модальних значень може бути декілька.

В дискретному ряду:

X	10	20	30	40
f	11	10	45	22

$Mo = 30$ (зустрічається найбільшу – 45, кількість разів).

В інтервальному ряді розподілу мода знаходиться за формулою: $Mo = x_{mo} + i\frac{f_2 - f_2}{(f_2 - f_1)+(f_2 - f_3)}$,

де: $x_{мо}$ — нижня межа модального інтервалу;
i — величина модального інтервалу;
f_2, f_1, f_3 — відповідно частота модального, передмодального та після модального інтервалів.

Зазначимо, що в рядах з нерівномірними інтервалами модальний інтервал вибирається по щільності розподілу, а мода це його середина.

Медіана (Me) — це значення ознаки, що ділить ранжований ряд значень показника на дві рівні частини. У першої половини одиниць значення ознаки менше медіани, а у другої — більше. Тобто, медіана — це серединне значення.

Вихідний ряд:

X	10	20	30	40	11	10	45	22	33

Ранжований ряд:

X	10	10	11	20	22	30	33	40	45

Потім визначають номер (місце) медіани:

$$N_{Me} = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

Me=X(5)=22. (при непарній кількості елементів).

При парній кількості елементів Me це напівсума середніх елементів.

X	10	10	11	20	22	30	33	40	45	50

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} = \frac{22+30}{2} = 26$$

В інтервальному ряді розподілу медіана визначається за формулою: $Me = x_{me} + i\frac{\frac{\Sigma f}{2} - f_н}{f_{ме}}$

де $x_{ме}$ — нижня межа медіанного інтервалу;
i — величина інтервалу;
$f_н$ — нагромаджена частота передмедіанного інтервалу;
$f_{ме}$ — частота медіанного інтервалу.

X, грн	f	Нагромаджена частота ($f_н$)
До 100	4	4
100-200	20	24
200-300	26	50
300-400	15	65
400-500	8	73
500-600	3	76
600-700	2	78
700 і більше	2	80

Мода дорівнює:

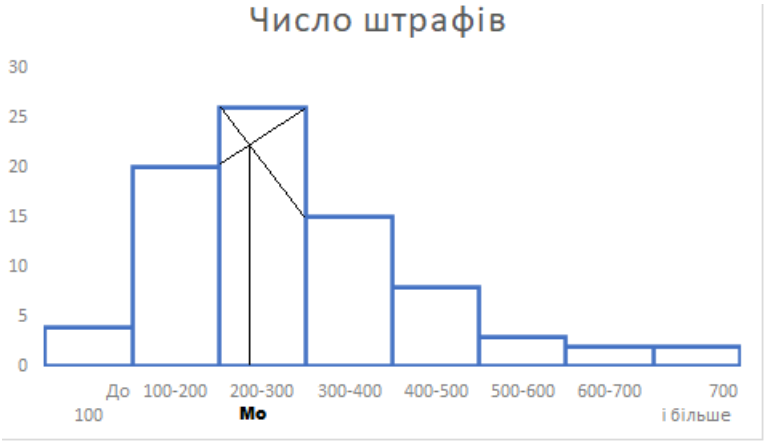
$$Mo = x_{mo} + i\frac{f_2 - f_1}{(f_2 - f_1) + (f_2 - f_3)} = 200 + 100\frac{26 - 20}{(26 - 20) + (26 - 15)} - 235{,}3\,грн.$$

Медіана становить:

$$Me = x_{me} + i\frac{\Sigma f / 2 - f_p}{f_{by}} = 200 + 100\frac{80/2 - 24}{26} = 261{,}5\,грн.$$

Таким чином, найчастіше розмір X становить 235,3 грн, половина значень X менше 261,5 грн, а половина — більше.

Графічний спосіб знаходження моди.

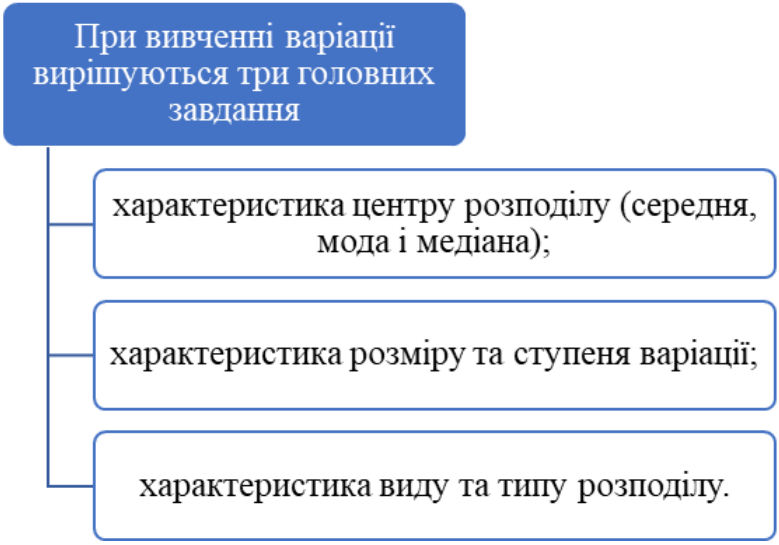


2.7. Показники варіації, аналіз рядів розподілу

Абсолютні показники варіації

У статистиці під варіацією розуміють мінливість, коливання значень ознак у одиниць сукупності. В результаті зведення та групування одержують ряди розподілу, які характеризують склад або структуру сукупності за певною варіюючою ознакою.

Варіація зумовлена дією багатьох факторів, які поділяються на систематичні та випадкові.



Вивчення варіації має велике значення з точки зору аналізу диференціації соціально-економічних явищ та процесів. Показники варіації покладено в основу вивчення взаємозв'язку між ознаками (дисперсійний аналіз), а також вибіркового спостереження. Варіація є також характеристикою однорідності сукупності за певною ознакою: чим менше є варіації, тим більш однорідною є сукупність.

Основні формули:

Показник	Вираз	Примітки
Розмах варіації (R)	$R = x_{max} - x_{min}$, де x_{max}, x_{min} — відповідно найбільше та найменше значення ознаки сукупності	В інтервальних рядах розподілу розмах варіації визначається як різниця між верхньою межею останнього та нижньою межею першого інтервалу.
Середнє лінійне відхилення (l)	$l = \sum \frac{\lvert X - \bar{X} \rvert}{n}$ – просте $l = \sum \frac{\lvert X - \bar{X} \rvert \cdot f}{\sum f}$ – зважене де X-значення ознаки, $\bar{X}$-середнє значення, n-кількість одиниць сукупності, f-частоти.	Просте середнє лінійне відхилення визначається по індивідуальних даних, а зважене — в рядах розподілу.
Дисперсія (σ^2)-це середній квадрат відхилень значень ознаки від середнього рівня	$\sigma^2 = \sum \frac{(X - \bar{X})^2}{n}$ – проста $\sigma^2 = \sum \frac{(X - \bar{X})^2 \cdot f}{\sum f}$ – зважена або $\sigma^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2$ проста $\sigma^2 = \frac{\sum X^2 f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2$ – зважена і — величина інтервалу	В інтервальних рядах розподілу для знаходження дисперсії спочатку визначається середина кожного інтервалу
Середнє квадратичне відхилення (σ) — показує, на скільки в середньому відхиляються значення ознаки від середнього рівня	$\sigma = \sqrt{\sum \frac{(X - \bar{X})^2}{n}}$ – просте $\sigma = \sqrt{\sum \frac{(X - \bar{X})^2 \cdot f}{\sum f}}$ –зважене	Показує, на скільки в середньому відхиляються значення ознаки від середнього рівня. Чим меншою є його величина, тим слабкішою є варіація і більш однорідною - статистична сукупність.

Пеня (штраф) за несплачену електроенергію

Розмір штрафу, *грн.*	Число штрафів	Середина інтервалу	*xf*	*/x–x/*	*/x–x/f*
До 100	4	50	200	385	1540
100–200	20	150	3000	285	5700
200–400	26	300	7800	135	3510
400–600	15	500	7500	65	975
600–800	8	700	5600	265	2120
800–1000	3	900	2700	465	1395
1000–2000	2	1500	3000	1065	2130
2000–3000	2	2500	5000	2065	4130
Разом	**80**	**x**	**34800**	**x**	**21500**

Середній розмір штрафу:

$$\bar{X} = \frac{\sum xf}{\sum f} = \frac{34800}{80} = 435\text{грн}$$

Середнє лінійне відхилення:

$$l = \sum \frac{|X - \bar{X}| \cdot f}{\sum f} = \frac{21500}{80} = 269\text{грн}$$

Ціна, *грн.*	**Кіль кість *шт.***	**Сере дина інтер валу**	***xf***	***(x–x)²***	***(x–x)²f***	***x²***	***x²f***
До 100	20	85	1700	10816	216320	7225	144500
100–130	24	115	2760	5476	131424	13225	317400
130–150	32	140	4480	2401	76832	19600	657200
150–200	56	175	9800	196	10976	30625	1715000
200–300	48	250	12000	3721	178608	62500	3000000
300–400	20	350	7000	25921	518420	422500	2450000
Разом	**200**	**x**	**37740**	**x**	**1132580**	**x**	**8254100**

Середнє значення:

$$\bar{X} = \frac{\sum xf}{\sum f} = \frac{37740}{200} = 189\text{грн}$$

Дисперсія:

$$\sigma^2 = \sum \frac{(X - \bar{X})^2 \cdot f}{\sum f} = \frac{1132580}{200} = 5663$$

або

$$\sigma^2 = \frac{\sum X^2 f}{\sum f} - \left(\frac{\sum Xf}{\sum f}\right)^2 = 5663$$

Другий спосіб знаходження дисперсії- через моменти.

$$\sigma^2 = i^2(m_2 - m_1^2)$$

$$m_2 = \frac{\sum\left(\frac{X-A}{i}\right)^2 f}{\sum f} \qquad m_1 = \frac{\sum\left(\frac{X-A}{i}\right) f}{\sum f}$$

i — величина інтервалу.

Місячний дохід, *грн.*	Число сімей	Серед. інтер.	*x–A,*	(x–A)/i	$\sum\left(\frac{X-A}{i}\right)$	$\sum\left(\frac{X-A}{i}\right) f$	$\sum\left(\frac{X-A}{i}\right)^2 f$
			A = 325	*i* = 50			
100–150	5	125	–200	–4	–20	16	80
150–200	15	175	–150	–3	–45	9	135
200–250	10	225	–100	–2	–20	4	40
250–300	20	275	–50	–1	–20	1	20
300–350	17	325	0	0	0	0	0
350–400	23	375	50	1	23	1	23
400–450	8	425	100	2	16	4	32
450 і більше	2	475	150	3	6	9	18
Разом	**100**	**x**	**x**	**x**	**–60**	**x**	**348**

$$m_1 = \frac{\sum\left(\frac{X-A}{i}\right) f}{\sum f} = \frac{-60}{100} = -0{,}6$$

$$m_2 = \frac{\sum\left(\frac{X-A}{i}\right)^2 f}{\sum f} = \frac{348}{100} = 3{,}48$$

$$\sigma^2 = i^2(m_2 - m_1^2) = 50^2(3{,}48\text{-}(-0{,}6)^2) = 7800$$

Середнє квадратичне відхилення (σ)

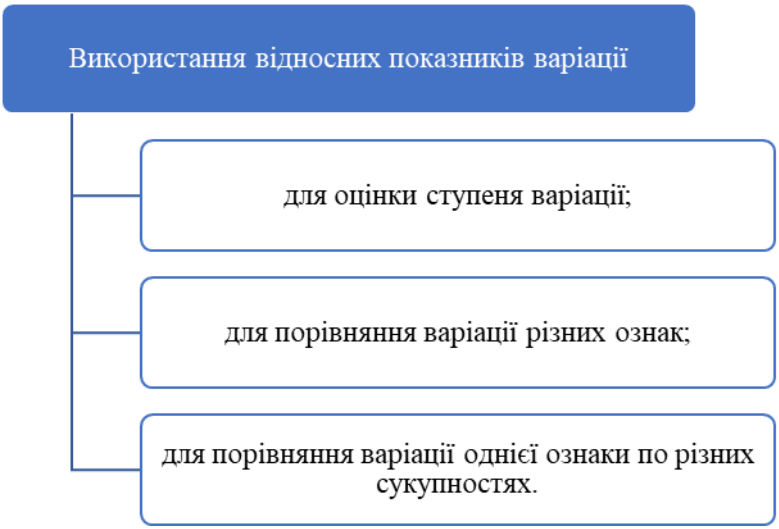
$$\sigma = \sqrt{\sum \frac{(X-\bar{X})^2}{n}} - \text{просте} \quad \sigma = \sqrt{\sum \frac{(X-\bar{X})^2 \cdot f}{\sum f}} - \text{зважене}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{7800} = 88{,}3\text{грн}$$

Відносні показники варіації

У загальному вигляді відносні показники варіації визначаються за формулою:

$$\text{Кв} = \frac{\text{Абсолютний показник варіації}}{\text{Середня величина}}$$



У статистичному аналізі найчастіше використовується коефіцієнт варіації у вигляді:

$$V = \frac{\sigma}{\bar{X}} \cdot 100$$

Вважається, що сукупність є однорідною, якщо V ≤ 33%. Крім цього, наведений коефіцієнт варіації застосовують для оцінки ступеня варіації: V < 15% — слабка; 15 ≤ V ≤ 25% — середня; V > 25% — сильна.

Наприклад, для порівняння успішності у двох групах:

Оцінка на іспиті	Кількість студентів		$xf1$	$xf2$	x^2	x^2f_1	x^2f_2
	1 група, $f1$	2 група, $f2$					
2	4	2	8	4	4	16	8
3	13	3	39	9	9	117	27
4	6	10	24	40	16	96	160
5	7	10	35	50	25	175	250
Разом	**30**	**25**	**106**	**103**		**404**	**445**

Середній бал:

$$\bar{X}_1 = \frac{\sum x_1 f_1}{\sum f_1} = \frac{106}{30} = 3{,}53 \quad \bar{X}_2 = \frac{\sum x_2 f_2}{\sum f_2} = \frac{103}{25} = 4{,}12$$

$$\sigma_1 = \sqrt{\frac{404}{30} - 3{,}53^2} = 1{,}005 \quad V_1 = \frac{1{,}005}{3{,}53} \cdot 100 = 28{,}5\%$$

$$\sigma_2 = \sqrt{\frac{445}{25} - 4{,}04^2} = 1{,}472 \quad V_2 = \frac{1{,}472}{4{,}04} \cdot 100 = 36{,}3\%$$

Досить сильна варіація у кожній групі. У другій вища.

Міжгрупова та внутрішньогрупова дисперсії

Розмір систематичної варіації, яка обумовлюється впливом групувальної ознаки, характеризує міжгрупова дисперсія. Це — середній квадрат відхилень групових середніх значень результативної ознаки $\bar{y}_i$ від його загальної середньої $\bar{y}_{\text{заг}}$. Таким чином, міжгрупова дисперсія визначається за формулою:

де f_i— число одиниць у кожній групі.

$$\sigma_{\text{м}}^2 = \frac{\sum(\bar{y}_i - \bar{y}_{\text{заг}})^2 f_i}{\sum f_i}$$

Випадкова варіація обумовлена дією випадкових факторів і проявляється у коливанні значень результативної ознаки в межах однієї групи. Розмір цієї варіації характеризується показником внутрішньогрупової дисперсії. Вона показує середній розмір відхилень значень результативної ознаки (y) від групової середньої (y_i) і визначається за формулою:

$$\sigma_i^2 = \frac{\sum(y - \bar{y}_i)^2}{f_i}$$

Мірою систематичної варіації є міжгрупова дисперсія ($\sigma_{\text{м}}^2$), а випадкової — середня із внутрішньогрупових дисперсій ($\sigma_{\text{і}}^2$).

Внутрішньогрупова дисперсія знаходиться окремо для кожної групи, тому для одержання її значення по сукупності в цілому підраховують середню величину:

$$\bar{\sigma}_i^2 = \frac{\sum \sigma_i{}^2 f_i}{\sum f_i}$$

Дисперсія результативної ознаки дорівнює сумі міжгрупової дисперсії та середньої з внутрішньогрупових дисперсій:

$$\sigma^2 = \sigma_м^2 + \bar{\sigma}_i^2$$

Це правило має назву правила додавання дисперсій. Воно використовується для того, щоб розкласти загальну варіацію результативної ознаки на систематичну та випадкову.

Для одержання розрахункових показників використаємо робочу таблицю:

Стать	Бали, одержані на екзамені			$(y-\bar{y}_i)^2$		
Чоловіки	3	5	5	1,77	0,45	0,45
	5	4	4	0,45	0,11	0,11
	4	3	5	0,11	1,77	0,45
	4	5	5	0,11	0,45	0,45
Разом по 1-й групі $\bar{y}_1 = \frac{\sum y}{n} = \frac{52}{12} = 4{,}33$				$\sigma_1^2 = \frac{\sum (y-\bar{y}_i)^2}{f_i} = \frac{6{,}68}{12} = 0{,}557$		
Жінки	3	4	4	0,64	0,04	0,04
	3	5	3	0,64	1,44	0,64
	4	5	3	0,04	1,44	0,64
	4			0,04		
Разом по 2-й групі $\bar{y}_1 = \frac{\sum y}{n} = \frac{38}{10} = 3{,}8$				$\sigma_1^2 = \frac{\sum (y-\bar{y}_i)^2}{f_i} = \frac{5{,}6}{10} = 0{,}560$		
По сукупності $\bar{y}_1 = \frac{\sum y}{n} = \frac{90}{22} = 4{,}09$				$\sigma_1^2 = \frac{\sum (y-\bar{y}_i)^2}{f_i} = \frac{13{,}86}{22} = 0{,}63$		

$$\sigma_м^2 = \frac{\sum (\bar{y}_i - \bar{y}_{заг})^2 f_i}{\sum f_i} = \frac{(4{,}33-4{,}09)\cdot 12 + (3{,}8-4{,}09)\cdot 10}{22} = 0{,}07$$

$$\bar{\sigma}_i^2 = \frac{\sum \sigma_i{}^2 f_i}{\sum f_i} = \frac{0{,}557\cdot 12 + 0{,}560\cdot 10}{22} = 0{,}56$$

$$\sigma^2 = \sigma_м^2 + \bar{\sigma}_i^2 = 0{,}07 + 0{,}56 = 0{,}63$$

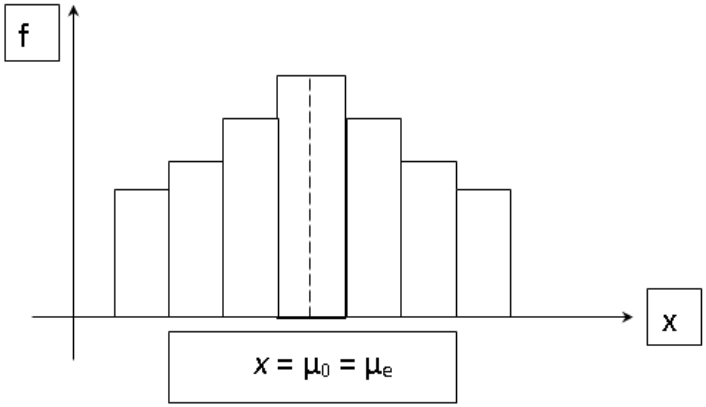
Варіація формується переважно за рахунок випадкових факторів.

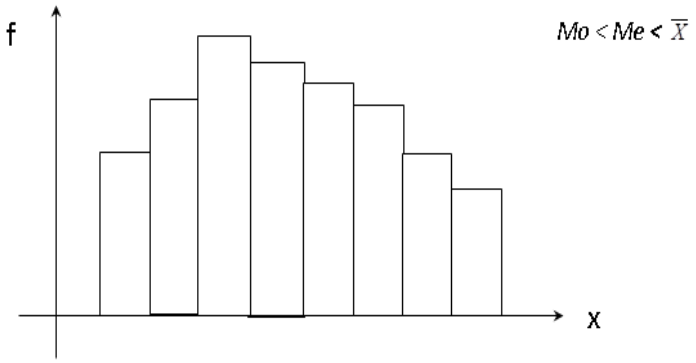
Характеристики форми розподілу

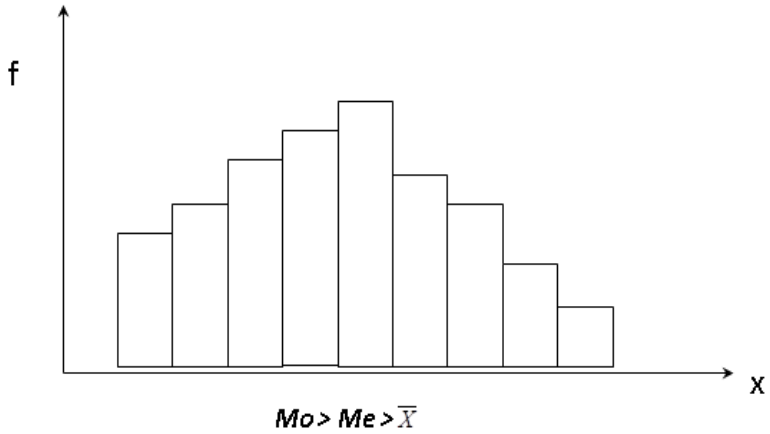
Характеристикою симетричності графіку розподілу виступає асиметрія. Перш за все величину асиметрії можна оцінити за рахунок середнього значення, моди та медіани розподілу.

Напрямок та міру асиметрії характеризують коефіцієнти асиметрії, які обчислюються за формулами:

$$A = \frac{\bar{X} - Mo}{\sigma} \text{ або } A = \frac{\bar{X} - Me}{\sigma}$$







Напрямок та міру асиметрії характеризують коефіцієнти асиметрії, які обчислюються за формулами:

$$A = \frac{\bar{X} - Mo}{\sigma} \text{ або } A = \frac{\bar{X} - Me}{\sigma}$$

При правосторонній асиметрії A > 0, при лівосторонній A < 0, при симетричному розподілі A = 0. Вважається, що при |A| < 0,25 асиметрія слабка, при 0,25<|A| <0,5 – середня, при |A| > 0,5 - сильна.

Наприклад, якщо $\sigma = 24{,}6$; $X = 180{,}5$; $\text{Mo} = 164{,}4$; $\text{Me} = 170{,}1$ коефіцієнти асиметрії становлять:

$$A = \frac{180{,}5 - 164{,}4}{24{,}6} = 0{,}86 \quad A = \frac{180{,}5 - 170{,}1}{24{,}6} = 0{,}43$$

Наявна правостороння асиметрія.

Коефіцієнт асиметрії можна також визначити за формулою:

$A = \frac{m_3}{\sigma^3}$ де $m_3 = i^3 \frac{\Sigma\left(\frac{x-\bar{x}}{i}\right)^3 f}{\Sigma f}$-центральний момент 3 порядку.

При дослідженні ступеня концентрації одиниць навколо середнього рівня (плосковершинності) визначають коефіцієнт ексцесу: $E = \frac{m_4}{\sigma^4} - 3$ де $m_4 = i^4 \frac{\sum\left(\frac{x-\bar{x}}{i}\right)^4 f}{\sum f}$-центральний момент 4 порядку.

E=0 розподіл нормальний

E>0 розподілтгостровершинний

E<0 розподіл плосковершинний

$$E = \frac{m_4}{\sigma^4} - 3 = \frac{878925}{24{,}6^4} - 3 = -0{,}6$$

Тут має місце плосковершинність розподілу.

Закономірність розподілу одиниць сукупності.

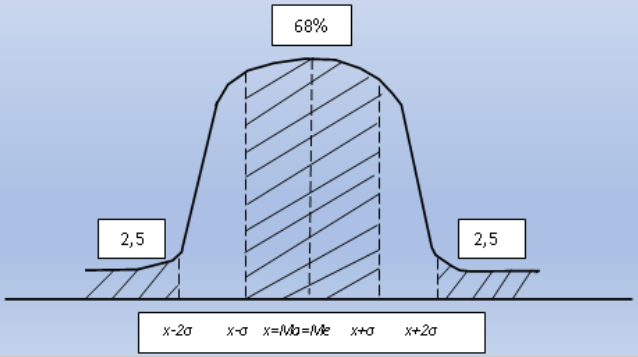
За значеннями варіючої ознаки можна описати певною функцією, яка має назву теоретичної кривої розподілу. Найбільш часто використовується крива нормального розподілу.

$$X = Mo = Me; \quad A = 0; \quad E = 0.$$

68,3% одиниць сукупності знаходяться в межах $X \pm \bar{\sigma}$;

95,5% — в межах $X \pm 2\bar{\sigma}$

99,7% — в межах $X \pm 3\bar{\sigma}$



Частоти, які розміщені на кривій нормального розподілу, називаються теоретичними частотами (ft).

$f_t = \frac{y_t \cdot i \cdot \sum f}{\sigma}$ і - величина інтервалу; σ - середнє квадратичне відхилення; y_t - ординати кривої нормального розподілу (знаходяться за спеціальними таблицями):

$y_t = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$ де $t = \frac{|x-\bar{x}|}{\sigma}$

Для об'єктивного судження про відповідність емпіричного розподілу кривій нормального розподілу використовуються спеціальні критерії відповідності (Пірсона, Колмогорова, Ястремського та ін.).

Критерій Пірсона χ^2 визначається за формулою:

$$\chi^2 = \sum \frac{(f - f_t)^2}{f_t}$$

де f — емпіричні частоти; f_t — теоретичні частоти.

Значення $\chi 2$ знаходяться з таблиці

Приклад розрахунку критерію Пірсона $\chi 2$: при $X = 72{,}6$ $\sigma = 20{,}52$

Інтервали	f	x	|x-x|	$t = \frac{\|x-\bar{x}\|}{\sigma}$	y_t	f_t	$(f-f_t)^2$	$\frac{(f-f_t)^2}{f_t}$
10-30	2	20	52,6	2,56	0,0151	1,47	0,281	0,191
30-50	8	40	32,6	1,59	0,1127	10,98	8,880	0,809
50-70	35	60	12,6	0,61	0,3312	32,28	7,398	0,225
70-90	41	80	7,4	0,36	0,3739	36,44	20,794	0,571
90-110	9	100	27,4	1,34	0,1626	15,85	46,923	2,960
110-130	4	120	47,4	2,31	0,0277	2,60	1,690	0,626
130 і більше	1	140	67,4	3,28	0,0040	0,38	0,372	0,954
РАЗОМ	**100**	**x**	**x**	**x**	**x**	**x**	**100**	**6,336**

Отже, $\chi 2 = 6{,}336$. При числі ступенів волі $k = 7 - 3 = 4$ та рівні ймовірності $0{,}95$ $\chi 2$таб $= 9{,}5$. Оскільки $\chi 2 < \chi 2$таб, розподіл можна вважати наближено нормальним.

2.8. Статистичні методи вивчення взаємозв'язків.

Види взаємозв'язків між явищами та процесами

Може існує багато видів взаємозв'язків між вибраними ознаками. Умовно \x можна поділити на дві нрупи – детерміновані та стохастичні.

Класифікація взаємозв'язків:

За видом

- **Функціональні (детерміновані)** зв'язки характеризуються тим, що одному значенню факторної ознаки (X) відповідає одне строго визначене (детерміноване) значення результативної ознаки (Y). Ці зв'язки завжди є повними, тобто значення результативної ознаки на 100% залежить від факторної. Наприклад, тарифний денний заробіток (Y) при фіксованій годинній тарифній ставці залежить від кількості відпрацьованих годин (X).
- **Стохастичні**. Одному значенню факторної ознаки (X) може відповідати декілька значень результативної ознаки (Y). Важливою особливістю цих зв'язків є те, що вони мають риси статистичної закономірності та проявляються у масі спостережень, при достатньо великій чисельності сукупності. Названі зв'язки є неповними, тому що завжди існують неврахованні фактори, отже значення Y залежить від значень X менше, ніж на 100%.

За напрямком зміни

- **Прямі**. При прямому зв'язку обидва показники змінюються в одному напрямку
- **Обернені**. При оберненому зв'язку напрямок зміни показників протилежний, тобто при зростанні X зменшується Y.

За аналітичним виразом

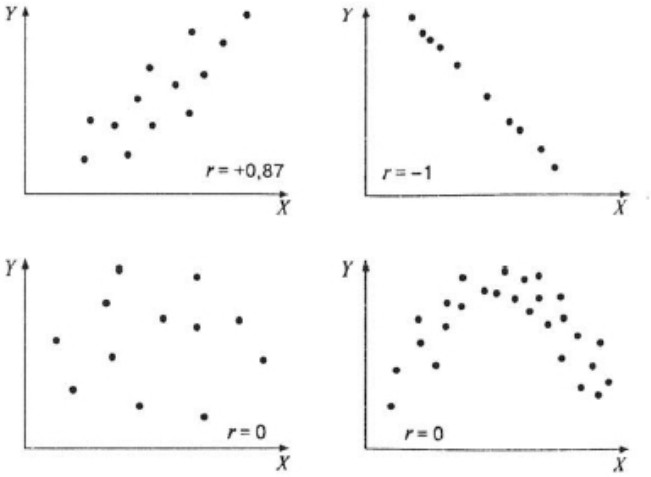
- лінійні
- нелінійні

Від числа факторних ознак

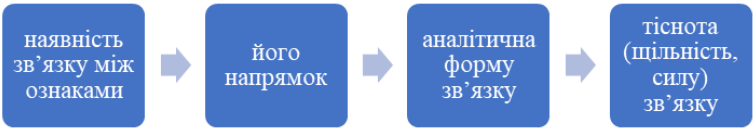
- однофакторні (парні)
- багатофакторні (множинні)

Для початку проводиться візуально-графічний аналіз. Тобто будується поле кореляції. Поле кореляції являє собою множину крапко, де по впливаюча ознака показана по осі абсцис, а результативна – ординат. Зазвичай, навіть на цьому етапі можна

зробити висновки про наявність і напрям зв’язку. Якщо крапки розміщені “кучно”, за якоюсь тенденцією, то зв’язок існує. Недоліки: суб’єктивність сприйняття малюнку та необхідність аналітичного знаходження тісноти зв’язку. Нижче приводяться зразки побудови поля кореляції.



Висновки після візуально-графічного аналізу:



Для вивчення стохастичних (кореляційних) зв'язків використовується метод порівняння паралельних рядів двох показників, один з яких є факторним (X), а другий – результативним (Y). При цьому основним завданням застосування цього методу є оцінка тісноти (сили) взаємозв'язку та визначення його напрямку на основі розрахунку спеціальних коефіцієнтів.

Найпростішим показником є **коефіцієнт Фехнера**.

Коефіцієнт Фехнера (Кф)

$$К_ф = \frac{С - Н}{С + Н}$$

де С – число співпадінь знаків відхилень від середньої;

Н – число не співпадінь знаків відхилень від середньої.

Якщо виконується нерівність $X \geq \overline{X}$ або $Y \geq \overline{Y}$, значенню присвоюється знак ” +”, в протилежному випадку – знак ”-”. В тому випадку, коли по обох показниках знаки однакові, має місце їх співпадіння, а коли вони різні – не співпадіння. Коефіцієнт Фехнера знаходиться в межах від -1 до +1. Якщо $|Кф| \to 0$, зв'язок між показниками слабкий, а при - зв'язок тісний.

Точнішим вважається **коефіцієнт кореляції рангів Спірмена**.

Коефіцієнт кореляції рангів Спірмена

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

де $d = r_x - r_y$ – різниця рангів факторного та результативного показників.

При цьому під рангом розуміють порядковий номер ранжованого показника. Варіація від -1 до +1. При $\rho > 0$ зв'язок між показниками прямий, а при $\rho < 0$ - обернений. Якщо $|\rho|$ наближається до 1, між показниками існує тісний (сильний) зв'язок, якщо $|\rho|$ <0,3 вважається, що взаємозв'язок відсутній.

Приклад

Ціна, грн. (X)	Обсяг продажу, шт. (Y)	Знаки відхилень		Ранги		d	d^2
		по X	по Y	по X	по Y		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
450	100	-	+	2	6	-4	16
560	84	+	-	5	2	3	9
730	56	+	-	8	1	7	49
480	91	-	-	3	4	-1	1
590	103	+	+	6	7	-1	1
620	85	+	-	7	3	4	16
360	120	-	+	1	8	-7	49
530	96	-	+	4	5	-1	1
4320	735	x	x	x	x	x	142

$$\overline{X}=\frac{\sum X}{n}=\frac{4320}{8}=540\ грн., \qquad \overline{Y}=\frac{\sum Y}{n}=\frac{735}{8}=92\ шт.$$

$$Кф=\frac{C-H}{C+H}=\frac{2-6}{2+6}=-0,5$$

$$\rho=1-\frac{6\sum d^2}{n(n^2-1)}=1-\frac{6*142}{8(64-1)}=-0,69$$

Очевидна наявність оберненого середнього зв'язку.

Метод аналітичного групування

Метод аналітичного групування полягає у тому, що сукупність розбивається на групи за факторною ознакою (X), далі по кожній групі та по сукупності визначаються середні значення X та Y. Порівняння середніх значень факторної та результативної ознак дозволяє зробити певні висновки про наявність та напрямок взаємозв'язку між ними.

Співвідношення між приростами середніх за формулою:

$$\frac{\Delta\overline{Y_i}}{\Delta\overline{X_i}} = \frac{\overline{Y_i} - \overline{Y}_{i-1}}{\overline{X_i} - \overline{X}_{i-1}}$$

де $\overline{X_i}, \overline{Y_i}$ - середні значення факторної та результативної ознаки по групах (групові середні). Якщо наведене співвідношення по групах приблизно стале, між показниками існує взаємозв'язок.

Для визначення тісноти зв'язку в стохастичних розподілах використовується емпіричне кореляційне відношення (η):

Емпіричне кореляційне відношення

$$\eta = \sqrt{\frac{\sigma_m^{\,2}}{\sigma^2}} = \sqrt{\frac{\sigma^2{}_m}{\sigma^2{}_m + \overline{\sigma^2{}_i}}},$$

де $\sigma^2{}_m$ - між групова дисперсія результативної ознаки;

σ^2 - загальна дисперсія результативної ознаки;

$\overline{\sigma^2{}_i}$ - середня із внутрішньо групових дисперсій ;

$$0 \le \eta \le 1$$

$\eta = 0$ зв'язок відсутній; $\eta = 1$ зв'язок функціональний;

Коефіцієнт детермінації (D) показує, на скільки відсотків варіація Y зумовлена варіацією X:

$$D = \eta^2 = \frac{\sigma^2{}_m}{\sigma^2} \quad .$$

Наведемо приклад розрахунку емпіричного кореляційного відношення та коефіцієнту детермінації за результатами аналітичного групування

Групи за ознакою	Кількість одиниць	$\overline{X_i}$	$\overline{Y_i}$	$\sigma^2{}_i$
1	15	150	34	14,9
2	25	270	39	15,3
3	40	340	45	13,8
4	12	410	49	18,9
5	8	500	56	25,4
Всього	100	х	х	х

$$\overline{Y_{заг}} = \frac{\sum \overline{Y_i} * f_i}{\sum f_i} = \frac{34 * 15 + 39 * 25 + 45 * 40 + 49 * 12 + 56 * 8}{100}$$

$$= \frac{4321}{100} = 43 од.$$

$$\sigma^2{}_m = \frac{\sum \left(\overline{Y_i} - \overline{Y_{заг}}\right)^2 f_i}{\sum f_i} =$$

$$= \frac{(34-43)^2 * 15 + (39-43)^2 * 25 + (45-43)^2 * 40 + (49-43)^2 * 12 + (56-43)^2 * 8}{100}$$

$$= 35{,}59$$

$$\overline{\sigma^2{}_i} = \frac{\sum \sigma^2{}_i * f_i}{\sum f_i}$$

$$= \frac{14{,}9 * 15 + 15{,}3 * 25 + 13{,}8 * 40 + 18{,}9 * 12 + 25{,}4 * 8}{100} = 15{,}88$$

$$\sigma^2 = \sigma^2{}_m + \overline{\sigma^2{}_i} = 35{,}59 + 15{,}88 = 51{,}47$$

$$\eta = \sqrt{\frac{\sigma_m{}^2}{\sigma^2}} = \sqrt{\frac{35{,}59}{51{,}47}} = 0{,}832$$

$$D = 0{,}8322 = 0{,}691 \text{ або } 69{,}1\%$$

Отже, між досліджуваними показниками існує прямий зв'язок, а варіація Y на 69,1% обумовлюється варіацією X.

Для перевірки суттєвості взаємозв'язку між X та Y використовують F-критерій за формулою :

$$F = \frac{\eta^2}{1-\eta^2} * \frac{K_2}{K_1},$$

де К2 = n-m, К1 = m-1 - число ступенів свободи при кількості одиниць n та кількості груп m. Критичні значення F-критерія занесені у спеціальні таблиці з яких визначається так зване табличне значення F-критерія (Fтабл). Якщо виконується умова F>Fтабл , зв'язок суттєвий (невипадковий).

У вищенаведеному прикладі $\eta^2 = 0{,}691$, К2 = 100-5 = 95, К1 = 5-1 = 4. $F = \frac{\eta^2}{1-\eta^2} * \frac{K_2}{K_1} = \frac{0{,}691}{1-0{,}691} * \frac{95}{4} = 53.1$ Оскільки F>Fтабл (53,1>2,5), зв'язок є суттєвим.

Парний кореляційно-регресійний аналіз

Стохастичні зв'язки, досліджуються за допомогою **кореляційно-регресійного аналізу**.

Найважливішою характеристикою кореляційного зв'язку є **лінія регресії**, котра пов'язує середні значення X та Y. Кореляційно-регресійна модель взаємозв'язку являє собою рівняння регресії, яке у загальному вигляді записується наступним чином:

$$y_x = f(X) + \xi,$$

де y_x – теоретичні значення Y ; $f(X)$- лінія регресії; ξ - залишкова компонента.

Переважно використовуються наступні функції (рівняння регресії):

лінійна $y_x = a_0 + a_1 X$

параболічна $y_x = a_0 + a_1 X + a_2 X^2$

степенева $y_x = a_0 * X^{a1}$

гіперболічна $y_x = a_0 + \frac{a_1}{X}$.

Розглянемо методику кореляційно-регресійного аналізу.

Етап 1. Застосовується візуально-графічний аналіз для з'язування загальної тенденції процесу.

Етап 2. Будується лінійне рівняння регресії.

Параметри лінійного рівняння регресії:

$$y_x = a_0 + a_1 X .$$

Для цього використовується метод найменших квадратів та розв'язується система рівнянь відносно a_0 і a_1 :

$$na_0 + a_1 \sum X = \sum Y$$
$$a_0 \sum X + a_1 \sum X^2 = \sum XY$$

З наведеної системи параметри рівняння регресії розраховуються різними способами, в тому числі за формулами:

$$a_0 = \frac{\sum Y \sum X^2 - \sum XY \sum X}{n \sum X^2 - \sum X \sum X} \quad або \quad a_0 = \overline{Y} - a_1 \overline{X}$$

$$a_1 = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - \sum X \sum X} \quad або \quad a_1 = \frac{\frac{1}{n} \sum XY - \overline{X}\,\overline{Y}}{\sigma^2 X}$$

Параметри a_1 називається коефіцієнтом регресії, що показує, на скільки одиниць змінюється Y при збільшенні X на одну. $a_1 > 0$ – зв'язок прямий, інакше – обернений.

Етап 3. Розраховуються теоретичні значення результативної ознаки Y_x:

середня квадратична (стандартна) помилка:

$$S = \sqrt{\frac{\sum (Y - Y_X)^2}{n}}$$

коефіцієнт апроксимації:

$$V = \frac{S}{Y} x100 .$$

Чим меншими є значення S та V, тим краще рівняння регресії описує (апроксимує) взаємозв'язок між X та Y .

Етап 4. Через коефіцієнт кореляції оцінюється тіснота зв'язку між X та У.

Лінійний коефіцієнт кореляції (r):

$$r = \frac{\overline{XY} - \overline{X} * \overline{Y}}{\sigma_X \sigma_Y} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\sum X^2 - \sum X \sum X\right)\left(n\sum Y^2 - \sum Y \sum Y\right)}} .$$

Значення коефіцієнта кореляції r в межах від -1 до +1. При $r > 0$ зв'язок між показниками прямий, а при $r < 0$ – обернений. Якщо : $|r| < 0{,}3$ вважається, що зв'язок відсутній; $0{,}.3 < |r| < 0{,}5$- зв'язок слабкий; $0{,}.5 < |r| < 0{,}7$- зв'язок середній; $0{,}7 < |r| < 0{,}9$- зв'язок сильний; $0{,}.9 < |r| < 1$- зв'язок дуже сильний.

Коефіцієнт детермінації $D = r^2$ показує, на скільки відсотків варіація Y обумовлюється варіацією X.

Часто також визначається коефіцієнт еластичності (Е) за формулою: $E = a_1 \frac{\bar{X}}{\bar{Y}}$. Цей коефіцієнт показує відсоток збільшення Y при збільшенні X на 1%.

На п'ятому етапі здійснюється перевірка суттєвості (невипадковості) взаємозв'язку між показниками за допомогою F-критерія Фішера:

$$F=\frac{r^2}{1-r^2}x\frac{K_2}{K_1},$$

де $К_1 = m-1;\ K_2 = n-m;\ n$ – кількість одиниць у сукупності; m - кількість параметрів у рівнянні регресії.

Таблиця 3

Ціна,грн. (X)	Обсяг продажу, шт. (Y)	XY	X^2	Y^2	Y_x	$(Y-Y_x)^2$
2,0	120	240,0	4,00	14400	120	0
2,4	88	212,2	5,76	7744	82	36
2,3	95	218,5	5,29	9025	92	9
2,5	60	150,0	6,25	3600	73	169
2,6	61	158,6	6,76	3721	64	9
2,5	62	155,0	6,25	3844	73	121
2,8	46	128,8	7,84	2116	45	1
2,9	40	116,0	8,41	1600	36	16
2,2	108	237,6	4,84	11664	101	49
2,8	50	140,0	7,84	2500	45	25
25,0	730	1755,7	63,24	60214	731	420

Визначаємо параметри лінійного рівняння регресії:

$$a_0=\frac{\sum Y\sum X^2-\sum XY\sum X}{n\sum X^2-\sum X\sum X}=\frac{730*63{,}24-1755{,}7*25}{10*63{,}24-25*25}=\frac{2272{,}7}{7{,}4}=307{,}1$$

$$a_1=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-\sum X\sum X}=\frac{10*1755{,}7-25*730}{10*63{,}24-25*25}=\frac{-693}{7{,}4}=-93{,}6$$

Отже, лінійне рівняння регресії має вигляд:

$$Yx = 307{,}1 - 93{,}6X.$$

Таким чином, при збільшенні ціни на 1 грн. Обсяг продажу в середньому зменшується на 93,6 од.

Теоретичні значення Y_x визначаються шляхом підстановки у рівняння регресії значень X:

$$Y_{x1} = 307{,}1 - 93{,}6 * 2{,}0 = 120$$

$$Y_{x2} = 307{,}1 - 93{,}6 * 2{,}4 = 82$$

Теоретичні значення Y занесені у 6 графу таблиці 3.

Визначаємо середню квадратичну (стандартну) помилку та коефіцієнт апроксимації:

$$S = \sqrt{\frac{\sum (Y - Y_X)^2}{n}} = \sqrt{\frac{420}{10}} = 6{,}48 \qquad \overline{Y} = \frac{\sum Y}{n} = \frac{730}{10} = 73$$

$$V = \frac{S}{\overline{Y}} x100 = \frac{6{,}48}{73} x100 = 8{,}9\%$$

.

Для оцінки тісноти взаємозв'язку між ознаками підрахуємо лінійний коефіцієнт кореляції:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - \sum X \sum X)(n\sum Y^2 - \sum Y \sum Y)}}$$

$$= \frac{10 * 1755{,}7 - 25 * 730}{\sqrt{(10 * 63{,}24 - 25 * 25)(10 * 60214 - 730 * 730)}} =$$

$$= \frac{-693}{715{,}8} = -0{,}968$$

Тісний кореляційний зв'язок. Коефіцієнт детермінації (D=r2=-0,9682=0,937) показує, що варіація Y на 93,7% зумовлена варіацією X. Коефіцієнт еластичності

$$E = a_1 \frac{\bar{X}}{\bar{Y}} = -93{,}6 \frac{2{,}5}{73} = -3{,}2\%$$

Отже, при збільшенні ціни на 1% обсяг продажу зменшується на 3,2%.

Далі здійснюється перевірка суттєвості зв'язку за допомогою F-критерія Фішера:

$F=\frac{r^2}{1-r^2}\cdot\frac{K_2}{K_1}=\frac{0{,}937}{1-0{,}937}\cdot\frac{8}{1}=119{,}9.$ Таким чином, F>Fтабл (119,9>5,3), а зв'язок між ознаками невипадковий (суттєвий).

2.9. Показники динаміки

Аналітичні показники динаміки

Ряди статистичних величин, які характеризують зміну явищ у часі, мають назву рядів динаміки. Вони складаються з двох елементів – показника часу (t) та рівнів ряду динаміки (y). Рівні ряду динаміки – це числові значення показника, котрі розташовані у хронологічній послідовності та відносяться до відповідного моменту або періоду часу.

Методи обробки рядів динаміки

Метод прямого перерахунку полягає у тому, що нові рівні ряду динаміки розраховуються повторно з врахуванням тих змін, які відбулися. Метод зімкнення ряду динаміки передбачає, що нові значення рівнів ряду динаміки визначаються на основі перехідного коефіцієнту. Цей коефіцієнт розраховуються як відношення значення показника в нових умовах до значення того ж показника у старих умовах, які обчислені за однаковий період або момент часу.

Місяці	1	2	3	4	5	6	7
Спожита електроенергія до зміни цін, тис. грн	100	112	112	120			
Спожита електроенергія після зміни цін, тис. грн				180	189	190	180

Визначаємо перехідний коефіцієнт (Кп):

$$Кп = \frac{180}{120} = 1{,}5$$

Підрахуємо скоректовані рівні ряду динаміки до зміни цін:

у1=100*1,5=150; у2=112*1,5=168; у3=112*1,5=168.

Одержали співставний ряд динаміки у нових цінах.

Класифікація рядів динаміки.

Маємо динамічний ряд шлюбів:

На дату	01.01.2012	01.01.2013	01.01.2014	01.01.2015	01.01.2016	01.01.2017	01.01.2018
Шлюбів на рік	567	674	578	823	784	563	724

Середня чисельність становить:

$$\bar{y} = \frac{\frac{1}{2}y_1 + y_2 + \ldots + \frac{1}{2}y_n}{n-1} = \frac{\frac{1}{2}567 + 674 + 578 + 823 + 784 + 563 + \frac{1}{2}724}{7-1} = \frac{4067{,}5}{7-1} = 678$$

Потрібно звернути увагу на те, що минулий приклад приводився для моментного ряду, тобто дані приводились на певний момент часу. В інтервальних рядах використовується інший підхід. Середній рівень ряду визначається за формулою середньої арифметичної простої або зваженої:

$$\overline{y} = \frac{\sum y}{n} \quad або \quad \overline{y} = \frac{\sum y * n}{\sum n}$$

Наприклад, маємо ряд динаміки виробництва продукції:

Квартал	I	II	III	IУ
Обсяг виробництва, т	200	212	195	220

$$\overline{y} = \frac{\sum y}{n} = \frac{200 + 212 + 195 + 220}{4} = \frac{827}{4} = 207\ m$$

Абсолютний розмір змін у часі показує абсолютний приріст. **Ланцюговий абсолютний приріст** ($Δл$) характеризує зміну показника за одиницю часу в абсолютному виразі:

$Δл = y_i - y_{i-1}$.

Базисний абсолютний приріст ($Δб$) показує зростання або зменшення показника в абсолютному виразі порівняно з рівнем, прийнятим за базу, тобто за певний період часу:

$Δб = y_i - y_0$.

$$Δб = \sum Δл; Δл = Δб_i - Δб_{i-1}$$

Середній абсолютний приріст показує середній розмір зміни показника за одиницю часу і розраховується за формулою:

$$\overline{\Delta} = \frac{\sum \Delta л}{n-1} = \frac{y_n - y_0}{n-1}.$$

Рік	Валовий збір,тон	*Δл*	*Δб*
2000	220,4	0	0
2001	219,3	-1,1	-1,1
2002	221,5	2,2	1,1
2003	225	3,5	4,6

$$\overline{\Delta} = \frac{-1{,}1 + 2{,}2 + 3{,}5}{3} = \frac{4{,}6}{3} = \frac{225{,}0 - 220{,}4}{4-1} = 1{,}53 тис.т$$

Для характеристики відносної зміни показника у часі

а) коефіцієнти росту:

ланцюгові $Кл = \frac{y_i}{y_{i-1}}$ базисні $Кб = \frac{y_i}{y_0}$

б) темпи росту:

ланцюгові $Тр = \frac{y_i}{y_{i-1}} x100$ базисні $Тб = \frac{y_i}{y_0} x100.$

Середній коефіцієнт або темп росту:

$$\overline{K} = \sqrt[n-1]{K_1 * K_2 * \ldots * K_{n-1}} = \sqrt[n-1]{\frac{y_n}{y_0}}$$

$$\overline{Т} = 100 * \sqrt[n-1]{K_1 * K_2 * \ldots * K_{n-1}} = 100 * \sqrt[n-1]{\frac{y_n}{y_0}}$$

Рік	Валовий збір,тон	Тл	Тб
2000	220,4	1	1
2001	219,3	0,9950	0,9950
2002	221,5	1,0100	1,0050
2003	225	1,0158	1,0209

$$\overline{T} = \sqrt[4-1]{\frac{225,0}{220,4}} = \sqrt[3]{1,021} = 1,007 або 100,7\%.$$

Темп приросту характеризує відносну швидкість зміни показника у часі, як правило, у процентному виразі. Темпи приросту визначаються наступним чином:

ланцюгові: $Тпр = \frac{Дл}{y_{i-1}} * 100 = Тл - 100$

базисні: $Тпр = \frac{Дб}{y_0} * 100 = Тб - 100.$

Середній темп приросту розраховується за формулою:

$$\overline{Тпр} = \overline{T} - 100 = 100,7 - 100 = 0,7\%.$$

Абсолютний вміст (значення) 1% приросту характеризує абсолютний розмір одного проценту зміни показника і визначається за формулами:

$$А = \frac{Дл}{Тпр ланц} = \frac{y_{i-1}}{100}.$$

Для порівняння швидкості зміни двох або більше взаємопов'язаних показників використовують метод приведення рядів динаміки до єдиної основи. Суть у тому, що вихідні ряди динаміки абсолютних значень показників замінюються базисними темпами зростання відносно моменту (періоду) часу.

Місяці	Витрати, тис. грн.	Прибуток, тис. грн.	Базисні темпи зростання, %	
			витрат	прибутку
1	10,7	3,1	100	100
2	12,4	3,2	115,9	103,2
3	15,1	3,5	141,1	112,9
4	14,9	3,7	139,3	119,4
5	16,2	3,6	151,4	116,1
6	15,8	6,4	147,7	206,5

Ми бачимо, що витрати зростають швидше ніж прибуток.

Для виявлення тенденції зміни показника у досить довгих рядах динаміки використовують метод збільшення інтервалів часу. При цьому вихідні рівні ряду замінюють сумарними або середніми значеннями показника за більші періоди часу.

обсяг продажу товару, шт.

Місяці	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
y	700	692	698	704	690	686	671	682	665	640	661	654

I квартал $\bar{y}_I = \frac{700+692+698}{3} = 697\,од.$

II квартал $\bar{y}_{II} = \frac{704+690+686}{3} = 693\,од.$

III квартал $\bar{y}_{III} = \frac{671+682+665}{3} = 673\,од.$

I***V*** квартал $\bar{y}_{IV} = \frac{640+661+654}{3} = 652\,од.$

Наявна чітко спадаюча тенденція.

Для обробки ряду динаміки з метою зменшення коливань його рівнів використовується метод рухомої середньої. Суть методу полягає у тому, що первинний ряд динаміки замінюється рядом середніх значень, підрахованих на основі рухомих сум. Рухома сума визначається шляхом додавання рівнів ряду, включених в інтервал вирівнювання (переважно це 3, 5, 7 рівнів).

Місяці	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
y	700	692	698	704	690	686	671	682	665	640
Рухомі суми	—	2090	2094	2092	2080	2047	2039	2018	1987	—
Рухомі середні	—	697	698	697	693	682	680	673	662	—

Найбільш детальне і таке, що набуло найширшого використання виявлення тенденції та закономірностей розподілу є аналітичне вирівнювання ряду.

Можемо сказати, що ми будуємо аналітичну модель за допомогою вибраної лінії тренду і потім перевіряємо її на адекватність. На практиці найчастіше використовують наступні рівняння тренду:

лінійне $y_t = a_0 + a_1 t$

параболічне $y_t = a_0 + a_1 t + a_2 t^2$

показникове $y_t = a_0 \cdot a_1^{\,t}$

степеневе $y_t = a_0 \cdot t^{a_1}$

гіперболічне $y_t = a_0 + \frac{a_1}{t}$

Коефіцієнти a_0 та a_1 знаходяться з системи нормальних рівнянь методом найменших квадратів. Для спрощення підрахунків параметр часу t задаються таким чином, щоб $\Sigma t = 0$.

$$a_0 = \frac{\Sigma y}{n}; \qquad a_1 = \frac{\Sigma yt}{\Sigma t^2}.$$

Місяці	y	t	yt	t^2	y_t
1	210	-2	-420	4	209
2	217	-1	-217	1	218
3	225	0	0	0	227
4	237	1	237	1	235
5	244	2	488	4	244
Разом	1133	0	88	10	1133

Для перевірки адекватності моделі (рівняння тренду) використовують наступні показники:

- середнє квадратичне (стандартне) відхилення

$$S = \sqrt{\frac{\Sigma(y-y_t)^2}{n}};$$

- коефіцієнт апроксимації

$$V = \frac{S}{\overline{y}} * 100.$$

Вважають, що рівняння тренду достатньою мірою апроксимує (описує) ряд динаміки, якщо $V < 10\%$.

$$a_0 = \frac{\Sigma y}{n} = \frac{1133}{5} = 226{,}6; a_1 = \frac{\Sigma yt}{\Sigma t^2} = \frac{88}{10} = 8{,}8.$$

$$y_t = 226{,}6 + 8{,}8t.$$

$$y_1 = 226{,}2 + 8{,}8x(-2) = 209{,}0;$$

$$y_2 = 226{,}2 + 8{,}8x(-1) = 218{,}0 im. \partial.$$

Місяці	1	2	3	4	5	Разом
y	210	217	225	237	244	1133
y_t	209	218	227	235	244	1133
$(y-y_t)^2$	1	1	4	4	0	10

$$S=\sqrt{\frac{10}{5}}=\sqrt{2}=1{,}4, \qquad V=\frac{1{,}4}{226{,}5}*100=0{,}6\%.$$

Коефіцієнт варіації показує, що рівняння тренду адекватно описує вихідний ряд динаміки.

Інтерполяція та екстраполяція

Інтерполяцією називають метод знаходження невідомого рівня в межах ряду враховуючи закономірність зміни показника у часі, що сформувалася у певному інтервалі. Інтерполяцію здійснюють на основі двох суміжних з невідомим рівнів ряду динаміки, або з використанням середніх показників динаміки.

Загальні формули для визначення невідомого рівня ряду (y_i) мають вигляд:

$$y_i=\frac{y_{i-1}+y_{i+1}}{2} \qquad y_i=y_{i-1}+\overline{\Delta} \qquad y_i=y_{i-1}*\overline{K}$$

де $\overline{\Delta}$ середній ланцюговий приріст, $\overline{K}$ середній ланцюговий темп росту

Дні	1	2	3	4	5	6	7	8	9	10
Обсяг добування газу, млн. м2	**70**	**72**	**71**	**73**	**...**	**75**	74	76	78	80

Визначимо невідомий п’ятий рівень ряду динаміки трьома способами:

1-й $y_5 = \frac{y_4 + y_6}{2} = \frac{73+75}{2} = 74\ (млн.м^3)$

2-й $\overline{\Delta} = \frac{y_n - y_0}{n-1} = \frac{80-70}{10-1} = 1{,}1, \quad y_5 = 73+1{,}1 = 74{,}1(млн.м^3)$

3-й $\overline{K} = \sqrt[n-1]{\frac{y_n}{y_0}} = \sqrt[9]{\frac{80}{70}} = 1{,}015 \quad y_5 = 73 \times 1{,}015 = 74{,}1(млн.м^3)$

Екстраполяція — це метод знаходження значення показника за межами відомого ряду. Цей метод передбачає поширення тенденції ряду у минуле або майбутнє, тому розрізняють ретроспективну та прогнозну екстраполяцію.

Екстраполяція знаходиться на основі середніх показників ряду динаміки або з допомогою рівняння тренду. Прогнозна екстраполяція першим способом виконується за формулами:

$$y_{t+l} = y_n + \overline{\Delta} \cdot l; \qquad y_{t+l} = y_n \cdot \overline{K}^l,$$

де l — період випередження (l = 1, 2, 3 ...), y_n – останній відомий рівень ряду динаміки.

При використання для прогнозу рівняння тренду в одержане рівняння підставляють наступні значення параметру часу $(t + l)$, наприклад, у лінійне рівняння тренду:

$$y_{t+l} = a_0 + a_1(t + l).$$

Розглянемо методику використання названих методів прогнозної екстраполяції:

Дні	Обсяг реалізації, т	t	t^2	y ·t	y_t
1	300	-3	9	-900	302
2	310	-2	4	-610	306
3	312	-1	1	-312	311
4	315	0	0	0	315
5	319	1	1	319	320
6	324	2	4	648	324
7	326	3	9	978	328
Разом	2206	0	28	123	2206

Визначимо середні показники ряду динаміки:

$$\overline{\Delta} = \frac{y_n - y_0}{n-1} = \frac{326-300}{7-1} = 4{,}3; \quad \overline{K} = \sqrt[n-1]{\frac{y_n}{y_0}} = \sqrt[7-1]{\frac{326}{300}} = 1{,}014$$

Підрахуємо прогнозні значення показника на наступні три дні двома способами:

	Перший спосіб	Другий спосіб
l = 1	$y_{t+1} = 326 + 4{,}3 \times 1 = 330{,}3;$	$y_{t+1} = 326 \times 1{,}014^1 = 330{,}6;$
l = 2	$y_{t+2} = 326 + 4{,}3 \times 2 = 334{,}6;$	$y_{t+2} = 326 \times 1{,}014^2 = 335{,}2;$
l = 3	$y_{t+3} = 326 + 4{,}3 \times 3 = 338{,}9;$	$y_{t+3} = 326 \times 1{,}014^3 = 339{,}9;$

$$a_0 = \frac{\Sigma y}{n} = \frac{2206}{7} = 315{,}14; \quad a_1 = \frac{\Sigma y \cdot t}{\Sigma t^2} = \frac{123}{28} = 4{,}39$$

$$y_t = 315{,}14 + 4{,}39t$$

$$l = 1 \qquad y_{t+1} = 315{,}14 + 4{,}39(3+1) = 332{,}7$$

$$l = 2 \qquad y_{t+2} = 315{,}14 + 4{,}39(3+2) = 337{,}1$$

$$l = 3 \qquad y_{t+3} = 315{,}14 + 4{,}39(3+3) = 341{,}5$$

При ретроспективній екстраполяції використовують ті ж показники, що й для прогнозування, але формули обчислень змінюються:

$$y_{t-l} = y_0 - \overline{\Delta} \cdot l; \qquad y_{t-l} = y_0 \div \overline{K}^l,$$

де l — період ретроспективи.

Наприклад, для наведеного ряду динаміки ретроспективні значення становлять:

l = 1 $\quad y_{t-1} = 300 - 4{,}3 \times 1 = 295{,}7; \quad y_{t-1} = 300 \div 1{,}014^1 = 295{,}9$

l = 2 $\quad y_{t-2} = 300 - 4{,}3 \times 2 = 291{,}4; \quad y_{t-2} = 300 \div 1{,}014^2 = 291{,}8$

На основі рівняння тренду ряду динаміки одержимо наступні ретроспективні значення:

l = 1 $\quad y_{t-1} = a_0 + a_1(t-1) = 315{,}14 + 4{,}39(-3-1) = 297{,}6$

l = 2 $\quad y_{t-2} = a_0 + a_1(t-2) = 315{,}14 + 4{,}39(-3-2) = 293{,}2$

Закономірності динаміки показників формуються під впливом систематичних та випадкових факторів, тому поряд із наявністю основної тенденції (тренду) їм притаманні відхилення від нього, сезонні коливання, структурні зрушення тощо.

Для вимірювання варіації в ряді динаміки використовують абсолютні та відносні показники: розмах варіації, середнє лінійне та середнє квадратичне відхилення, коефіцієнт варіації. Методика визначення цих показників аналогічна тій, що застосовується для індивідуальних даних:

$$R_t = y_{\max} - y_{\min}; \qquad \bar{l}_t = \frac{\Sigma|y - \bar{y}|}{n}; \quad \sigma_t = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}; \quad V_t = \frac{\sigma_t}{\bar{y}} \times 100.$$

Розрахункові значення yt одержують на основі середніх показників або рівняння тренду. Коливання фактичних значень ряду динаміки навколо тренду, що проявляється у відхиленні їх від розрахункових, обумовлюється впливом випадкових факторів і вимірюється абсолютними та відносними показниками:

стандартне відхилення: $\quad S = \sqrt{\frac{\Sigma(y-y_t)^2}{n}};$

коефіцієнт апроксимації: $V_a = \frac{S}{\bar{y}} \times 100.$

Показники S та V_a використовують також для підбору оптимального рівняння тренду, для якого $S(V_a) \to min.$

Наведемо приклад розрахунку названих показників, використовуючи раніш знайдене рівняння тренду:

Дні	y	$(y-\bar{y})^2$	y_t	$y-y_t$	$(y-\bar{y})^2$
1	300	225	302	-2	4
2	310	25	306	4	16
3	312	9	311	1	1
4	315	0	315	0	0
5	319	16	319	0	0
6	324	81	324	0	0
7	326	121	328	2	4
Разом	**2006**	**477**	**2205**	**x**	**25**

Розмах варіації:

R = ymax – ymin =326 – 300 = 26 (од.)

Середній рівень ряду:

$$\bar{y} = \frac{\Sigma y}{n} = \frac{2206}{7} = 315(од.)$$

Середнє квадратичне відхилення:

$$\sigma = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \frac{477}{7} = 8{,}25(од.)$$

Коефіцієнт варіації:

$$V = \frac{\sigma}{\bar{y}} \times 100 = \frac{8{,}25}{315} \times 100 = 2{,}6\%$$

Рівняння тренду:

$$y_t = 315{,}14 + 4{,}39t.$$

Стандартне відхилення:

$$\sigma = \sqrt{\frac{\Sigma(y - y_t)^2}{n}} = \frac{25}{7} = 1{,}89\,(од.)$$

Коефіцієнт апроксимації:

$$V_a = \frac{S}{\bar{y}} \times 100 = \frac{1{,}89}{315} \times 100 = 0{,}6\%.$$

Отже, ряд динаміки має слабку варіацію, а лінійне рівняння тренду добре описує (апроксимує) вихідні дані. Вплив випадкових факторів дуже слабкий, оскільки V_a = 0,6%.

Якщо в ряді динаміки спостерігаються стійкі відхилення від тренду, то можна припустити наявність у ньому одного чи кількох коливальних процесів. Особливо це помітно, коли явища, що вивчаються мають сезонний характер, тобто зростання чи спадання рівнів повторюється регулярно з певним інтервалом, за місяцями року, споживання палива та електроенергії для побутових потреб, сезонний розпродаж товарів і т.д.).

Індекси сезонності показують, у скільки разів фактичний рівень ряду динаміки в конкретний момент або період часу відрізняється від середнього рівня або рівня, що визначається за рівнянням тренду.

Рівень сезонності явища, що вивчається, оцінюється кількісно з допомогою індексів сезонності чи шляхом гармонійного аналізу. У тому випадку, коли коливання показника не мають чіткої тенденції, індекси сезонності визначаються за формулою:

$$I_{сез} = \frac{\bar{y}_i}{\bar{y}_{заг}} \times 100,$$

де$\bar{y}_i$— середнє значення показника за і – й період року:

$\bar{y}_{заг}$— загальне середнє значення за всі роки.

Індекси сезонності прийнято розраховувати не менше, ніж за три роки. Приклад обчислення індексів сезонності:

Квартали	**Товарооборот, тис.грн.**			$\bar{y}_i$	**$I_{сез}$ (%)**
	2001	2002	2003		
I	285	264	279	276	97,2
II	293	278	286	286	100,7
III	295	281	302	293	103,2
IV	290	265	287	281	98,9

Середні значення за квартали:

$$\bar{y}_I = \frac{285+264+279}{3} = 276\,(тис.грн.) \qquad \bar{y}_{II} = \frac{293+278+286}{3} = 286\,(тис.грн.)$$

$$\bar{y}_{III} = \frac{295+281+302}{3} = 293\,(тис.грн.) \qquad \bar{y}_{IV} = \frac{290+265+287}{3} = 281\,(тис.грн.)$$

Загальна середня:

$$\bar{y}_{заг} = \frac{276+286+293+281}{4} = 284\,(тис.грн)$$

Визначимо індекси сезонності:

$$I_{сез.I} = \frac{276}{284}\times 100 = 97{,}2\%; \qquad I_{сез.II} = \frac{286}{284}\times 100 = 100{,}7\%;$$

$$I_{сез.III} = \frac{293}{284}\times 100 = 103{,}2\%; \qquad I_{сез.IV} = \frac{281}{284}\times 100 = 98{,}9\%;$$

Отже, можна зробити висновки про те, що товарообороті має слабкі сезонні коливання.

Наведемо приклад побудови графіка сезонної хвилі на основі раніше обчислених індексів сезонності товарообороту.

Якщо спостерігається стабільний тренд за декілька років аналіз сезонності проводиться на основі аналітичного групування. В такому випадку будують лінію тренду, а потім на її основі знаходять теоретичні значення ряду. Після цього індекси сезонності розглядаються як співвідношення між фактичними та теоретичними рівнями.

$$I_{сез} = (\Sigma \frac{y}{y_t} \times 100) / n$$

де y — фактичні рівні ряду; y_t — теоретичні рівні ряду, обчислені за рівнянням тренду;

n — число років, за які є дані.

Наведемо приклад розрахунку індексів сезонності:

Квартали	**Роки**			**Теоретичні значення**			
	2001	*2002*	*2003*	*2001*	*2002*	*2003*	
I	190	198	220	203	207	211	
II	224	231	238	204	208	212	
III	211	214	209	205	209	213	
IV	180	182	193	206	210	214	

Виконаємо вирівнювання ряду динаміки по прямій: $y_t = a_0 + a_1 t$.

y	190	224	211	180	198	231	214	182	220	238	209	193
t	-5,5	-4,5	-3,5	-2,5	-1,5	-0,5	0,5	1,5	2,5	3,5	4,5	5,5
t^2	30,25	20,25	12,25	6,25	2,25	0,025	0,03	2,3	6,25	12,3	20,3	30,3
yt	-1045	-1008	-739	-450	-297	-116	107	273	550	833	941	1061
y_t	203	204	205	206	207	208	209	210	211	212	213	214
y/y_t	93,6	109,8	102,9	87,4	95,7	111,1	102,4	86,7	104,3	112,3	98,1	90,2

$\Sigma y = 2490;\ \Sigma t^2 = 143{,}0;\quad \Sigma yt = 111.$ Звідси:

$$a_0 = \frac{\Sigma y}{n} = \frac{2490}{12} = 207{,}5; \quad a_1 = \frac{\Sigma yt}{\Sigma t^2} = \frac{111}{143} = 0{,}8$$

Рівняння тренду: yt = 207,5+ 0,8 t

Визначимо теоретичні рівні ряду динаміки та занесемо їх у таблицю.

Далі знайдемо відношення y / yt в процентах та індекси сезонності:

$$I_{сезI} = \frac{93{,}6 + 95{,}7 + 104{,}3}{3} = 97{,}9\%;$$

$$I_{сезII} = \frac{109{,}8 + 111{,}1 + 112{,}3}{3} = 111{,}1\%;$$

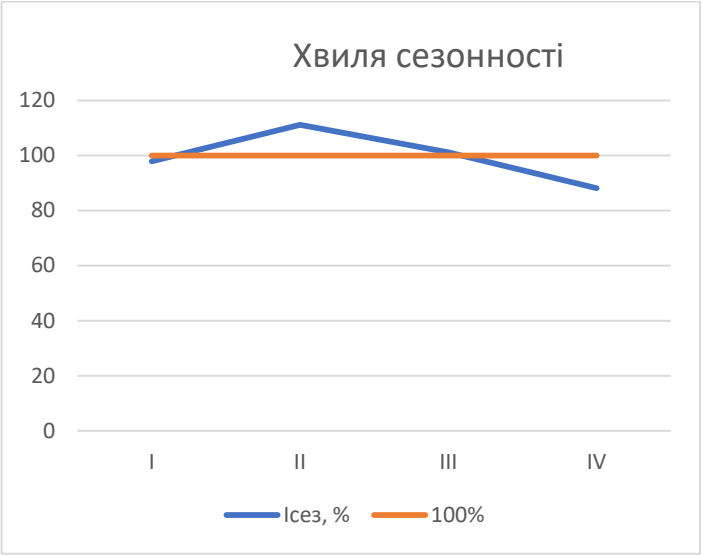
$$I_{сезIII} = \frac{102{,}9 + 102{,}4 + 98{,}1}{3} = 101{,}1\%;$$

$$I_{сезIV} = \frac{87{,}4 + 86{,}7 + 90{,}2}{3} = 88{,}1\%.$$

Занесемо результати підрахунків у таблицю:

Квартали	**(y/y_t) * 100**			**$I_{сез}$, %**
	2001	*2002*	*2003*	
I	93,6	95,7	104,3	97,9
II	109,8	111,1	112,3	111,1
III	102,9	102,4	98,1	101,1
IV	87,4	86,7	90,2	88,1

Побудуємо графік сезонної хвилі:

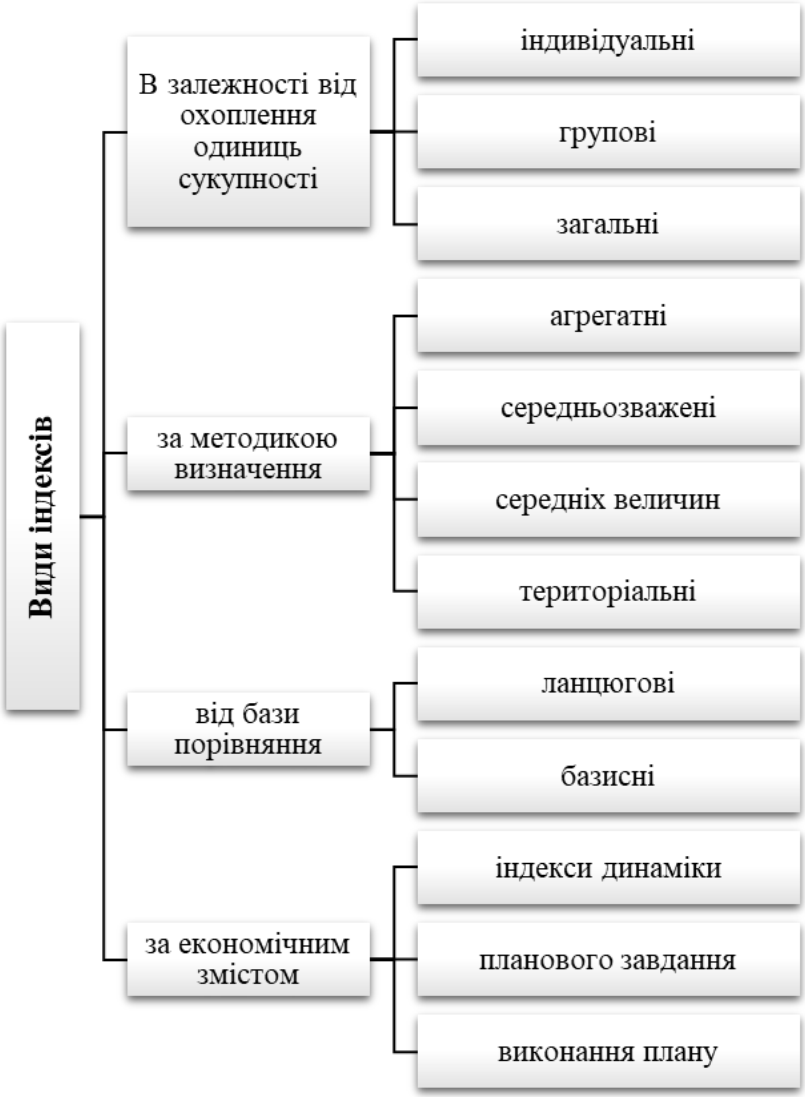


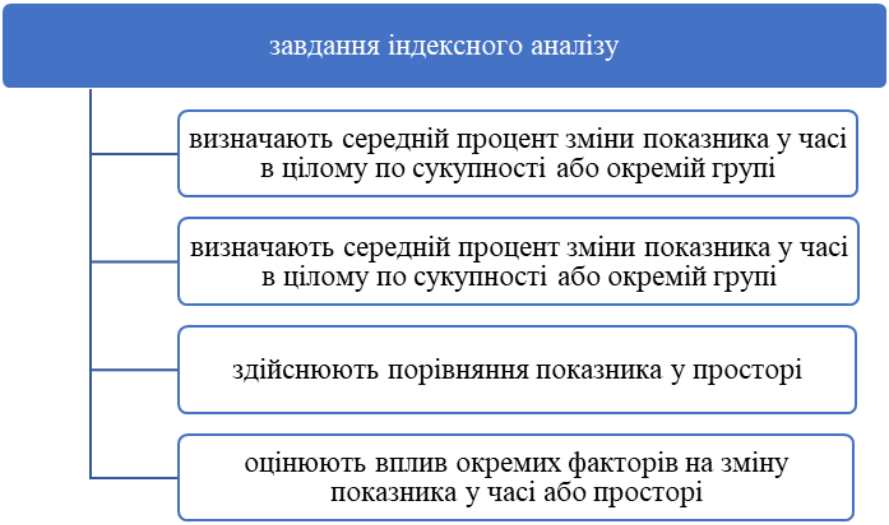
2.10. Індекси

У статистиці під індексом розуміють специфічну відносну величину, яка характеризує зміну показника у часі та просторі. Індекси можуть визначатися у неоднорідній сукупності, що складається з елементів, які безпосередньо не можна сумувати.

Індекс - це відносна величина, яка показує, у скільки разів рівень досліджуваного явища в конкретних умовах відрізняється від рівня того ж явища за інших умов. При цьому умови можуть різнитися в часу, тоді виходить динамічний індекс, у просторі, тоді виходить територіальний індекс, а також стосовно

планованого рівнем, тоді виходять індекси планового завдання та виконання плану.





У статистичному аналізі індекси використовуються не тільки для співставлення рівнів, але й оцінки значимості чинників, пояснюючих абсолютну відмінність рівнів результативного показника.

Залежно від складності розрізняють три типи індексів: індивідуальні індекси, загальні індекси та індекси середніх величин.

Існують формалізовані методи запису індексів. Якісні показники, які входять в співвідношення мають такий запис:

P – ціна одиниці товару;

Z – собівартість одиниці продукції;

f – зарплата одного працівника;

Q – кількість одиниць сукупності.

Відповідно кількісні показники:

q – кількість (фізичний обсяг) товару або продукції;

T – чисельність працівників;

S – посівні площі тощо.

На основі одного якісного та одного кількісного показника будуються об'ємні показники:

$pq = p * q$ – товарообороті або вартість продукції;

$zq = z * q$ – витрати на виробництво продукції;

$fT = f * T$ – фонд заробітної плати працівників;

$yS = y * S$ – валовий збір культур тощо.

Для позначення часу використовують, як правило, два періоди:

0 – попередній або базисний період,

1 – наступний або звітний період.

Якщо дані аналізуються більш ніж за два періоди, використовують порядкові номери 1,2,3, і т.д.

Індивідуальні індекси застосовуються, коли немає значення структура явища, що вивчається.

Індивідуальні індекси кількісних показників мають вигляд:

$$i_q = \frac{q_1}{q_0}; \quad i_T = \frac{T_1}{T_0}; \quad i_S = \frac{S_1}{S_0}.$$

Індивідуальні індекси об'ємних показників можна записати наступним чином:

$$i_{pq} = \frac{p_1 q_1}{p_0 q_0} = i_p * i_q; i_{zq} = \frac{z_1 q_1}{z_0 q_0} = i_z * i_q;$$

$$i_{fT} = \frac{f_1 T_1}{f_0 T_0} = i_f * i_T;$$

Очевидно, що маючи два індивідуальні індекси ми завжди можемо знайти третій, як їх співвідношення.

Індивідуальні індекси якісних показників визначаються за формулами:

$$i_p = \frac{p_1}{p_0}; i_z = \frac{z_1}{z_0}; i_f = \frac{f_1}{f_0}; i_y = \frac{y_1}{y_0}.$$

i_{fT} – індекс фонд заробітної плати працівників $= 1{,}756$;

$i_{\text{Т}}$ – індекс чисельності працівників $= 0{,}8$.

Індекс заробітної плати:

$$i_f = \frac{i_{fT}}{i_T} = \frac{1{,}756}{0{,}8} = 2{,}197.$$

Індивідуальні індекси:

Продукція	Собівартість, грн..		Кількість, шт.		$z0\ q0$	$z1\ q1$
	I кв, $z0$	II кв, $z1$	I кв, $g0$	II кв, $g1$		
А	80	90	100	120	8000	10800
Б	60	60	70	80	4200	4800
В	70	65	140	100	9800	6500

Визначимо індивідуальні індекси собівартості:

$$i_z^A = \frac{z_1}{z_0} = \frac{90}{80} = 1{,}125 \quad i_z^Б = \frac{60}{60} = 1 \quad i_z^B = \frac{65}{70} = 0{,}929.$$

Отже, по продукції А собівартість зросла на 12,5%, по продукції Б залишилася без змін, а по продукції В – зменшилася на 7,1%.

Індивідуальні індекси кількості (фізичного обсягу) продукції:

$$i_g^A = \frac{q_1}{q_0} = \frac{120}{100} = 1{,}2 \quad i_q^Б = \frac{80}{70} = 1{,}143 \quad i_q^B = \frac{100}{140} = 0{,}714$$

Таким чином, по продукції А виробництво зросло на 20%, по продукції Б – на 14,3%, а по продукції В – зменшилася на 28,6%.

Індивідуальні індекси витрат на виробництво продукції:

$$i_{zq}^{A} = \frac{z_1 q_1}{z_0 q_0} = \frac{10800}{8000} = 1{,}350 \qquad i_{zq}^{B} = \frac{4800}{4200} = 1{,}143$$

$$i_{zq}^{B} = \frac{6500}{9800} = 0{,}663$$

Витрати на виробництво продукції А зросли на 35%, по продукції Б – на 14,3%, а по продукції В – скоротилися на 33,7%.

Взаємозв'язок індексів:

$$i_{zq} = i_z x i_q = 1{,}125x1{,}2 = 1{,}350$$

$$i_{zq} = 1x1{,}143 = 1{,}143$$

$$i_{zq} = 0{,}929x0{,}714 = 0{,}663$$

Агрегатні індекси відносяться до загальних індексів, які характеризують середню зміну індексованого показника у часі та просторі.

Якщо першим фактором вважається кількість товарів, а другим – ціни, то визначається загальний кількісний індекс Ласпейреса та загальний ціновий індекс Пааше.

Якщо першим фактором вважаються ціни, а другим – кількість товарів, то визначається загальний ціновий індекс Ласпейреса та загальний кількісний індекс Пааше. На практиці використовується їх співвідношення.

агрегатний індекс ціни $I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1}$,

агрегатний індекс собівартості $I_z = \frac{\sum z_1 q_1}{\sum z_0 q_1}$,

агрегатний індекс зарплати $I_f = \frac{\sum f_1 T_1}{\sum f_0 T_1}$,

агрегатний індекс урожайності $I_y = \frac{\sum y_1 S_1}{\sum y_0 S_1}$.

Індекс стає агрегатним, коли вивчається неоднорідне явище (наприклад, виторг від продажу не одного, а всіх або декількох товарів), що унеможливлює підсумовування об'ємного показника в натуральних одиницях.

В агрегатний індексах кількісних показників індексований показник у чисельнику береться за звітний період, а у знаменнику за базисний, а співмножник (якісний показник) у чисельнику і знаменнику фіксується, тобто береться однаковим, на рівні базисного періоду

індекс фізичного обсягу $I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0}$ або $I_q = \frac{\sum z_0 q_1}{\sum z_0 q_0}$,

індекс чисельності працівників $I_T = \frac{\sum f_0 T_1}{\sum f_0 T_0}$,

індекс посівних площ $I_S = \frac{\sum y_0 S_1}{\sum y_0 S_0}$.

У цьому випадку загальний виторг можна записати в агрегатному вигляді як суму творів кількості товарів та їх ціни. Порівнюючи рівні цього показника в умовах звітного та базисного періоду, отримуємо агрегатний індекс.

Між агрегатними індексами показників існує взаємозв'язок: агрегатний індекс об'ємного показника дорівнює добутку агрегатних індексів якісного та кількісного показників.

$$I_{pq} = I_p x I_q, I_{zq} = I_z x I_q, I_{fT} = I_f x I_T, I_{yS} = I_y x I_S.$$

Товар	грн		шт.		$p0q0$	$p0q1$	$p1q1$
	I кв. (p$\mathbf{0}$)	II кв. (p$\mathbf{1}$)	I кв. ($q\mathbf{0}$)	II кв. ($q\mathbf{1}$)			
А	230	200	50	90	11500	20700	18000
Б	280	300	100	100	28000	28000	30000
В	160	150	150	130	24000	20800	19500
					63500	69500	67500

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{67500}{69500} = 0{,}971$$

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{69500}{63500} = 1{,}094$$

$$I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{67500}{63500} = 1{,}063$$

Отже, загалом ціни в середньому впали на 2,9%, кількість проданих одиниць (фізичний обсяг) збільшився в середньому на 9,4%, а товарообіг зріс на 6,3%.

Перерахуємо ті ж дані в абсолютному вимірі.

$$\Delta = \sum p_1 q_1 - \sum p_0 q_0 = 67500 - 63500 = 4000 грн.$$

$$\Delta_p = \sum p_1 q_1 - \sum p_0 q_1 = 67500 - 69500 = -2000 \text{грн}.$$

$$\Delta_q = \sum p_0 q_1 - \sum p_0 q_0 = 69500 - 63500 = 6000 грн.$$

Агрегатні індекси можна визначати як ланцюгові та базисні. В ланцюгових індексах індексований показник береться за суміжні періоди часу (наступний і попередній), а у базисних – у знаменнику беруться значення індексованого показника за базисний період. Таким чином, перші індекси характеризують середню зміну індексованого показника за одиницю часу.

Показники розраховуються за формулами:

Ланцюгові

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \; I_p = \frac{\sum p_2 q_2}{\sum p_1 q_2} \; I_p = \frac{\sum p_3 q_3}{\sum p_2 q_3} \; I_p = \frac{\sum p_4 q_4}{\sum p_3 q_4}$$

Базисні

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \; I_p = \frac{\sum p_2 q_2}{\sum p_0 q_2} \; I_p = \frac{\sum p_3 q_3}{\sum p_0 q_3} \; I_p = \frac{\sum p_4 q_4}{\sum p_0 q_4}$$

Визначимо ланцюгові та базисні індекси урожайності по двох регіонах за наведеними даними:

Роки	**Урожайність кукурудзи, ц/га**		**Посівні площі, га**	
	регіон 1	**регіон 2**	**регіон 1**	**регіон 2**
2000	35	46	1500	1100
2001	46	45	1600	1800
2002	49	38	1300	2000
2003	41	50	2200	1500

Ланцюгові індекси урожайності:

2001 р. $I_y = \frac{\sum y_1 S_1}{\sum y_0 S_1} = \frac{46*1600+45*1800}{35*1600+40*1800} = \frac{154600}{128000} = 1{,}028$

2002 р. $I_y = \frac{\sum y_2 S_2}{\sum y_1 S_2} = \frac{49*1300+38*2000}{46*1300+45*2000} = \frac{139700}{149800} = 0{,}933$

2003 р. $I_y = \frac{\sum y_3 S_3}{\sum y_2 S_3} = \frac{41*2200+50*1500}{49*2200+38*1500} = \frac{165200}{164800} = 1{,}002$

Отже, по двох регіонах урожайність у 2001 р. порівняно з 2000 р. в середньому зросла на 2,8%, у 2002р. порівняно з 2001 р. – зменшилася на 6,7%, а у 2003 р. порівняно з 2002 р. – зросла лише на 0,2%.

Базисні індекси урожайності:

2001 р. $I_y = \frac{\sum y_1 S_1}{\sum y_0 S_1} = \frac{46*1600+45*1800}{35*1600+40*1800} = \frac{154600}{128000} = 1{,}028$

2002 р. $I_y = \frac{\sum y_2 S_2}{\sum y_0 S_2} = \frac{49*1300+38*2000}{35*1300+40*2000} = \frac{139700}{125500} = 1{,}113$

2003 р. $I_y = \frac{\sum y_3 S_3}{\sum y_0 S_3} = \frac{41*2200+50*1500}{35*2200+40*1500} = \frac{165200}{137000} = 1{,}206$

Отже, по двох регіонах урожайність у 2001 р. порівняно з 2000 р. в середньому зросла на 2,8%, у 2002р. порівняно з 2000 р. – зросла на 11,3%, а у 2003 р. порівняно з 2000 р. – зросла на 20,6%.

Якщо від чисельника відповідного індексу відняти його знаменник, визначимо приріст валового збору за рахунок зміни урожайності або порівняно з попереднім роком, або з базисним:

$\Delta_y = \sum y_1 S_1 - \sum y_0 S_1 = 154600 - 128000 = 26600 ц$

$\Delta_y = \sum y_2 S_2 - \sum y_1 S_2 = 139700 - 149800 = -10100 ц$ і т. д.

Крім запису загальних індексів в агрегатному вигляді практично часто використовують формули їх розрахунку як середніх величин з відповідних індивідуальні індекси.

Наприклад, загальний індекс виручки може бути записаний як середня арифметична виважена з індивідуальних індексів виторгу. Для цього помножимо і розділимо його чисельник на допомогу базисного періоду.

Загальний індекс виручки також може бути записаний як середня гармонійна зважена з індивідуальних індексів виторгу.

Для цього помножимо і розділимо його знаменник на виручку звітного періоду.

Агрегатні індекси кількісних показників можна перетворити у середньоарифметичні індекси наступним чином:

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{\sum i_q * p_0 q_0}{\sum p_0 q_0}, \quad i_q = \frac{q_1}{q_0}, \quad q_1 = i_q * q_0$$

$$I_T = \frac{\sum f_0 T_1}{\sum f_0 T_0} = \frac{\sum i_T * f_0 T_0}{\sum f_0 T_0}, \quad i_T = \frac{T_1}{T_0}, \quad T_1 = i_T * T_0$$

Динаміка посівних площ.

Культури	Валовий збір у базисному році, ц $(y0S0)$	Зміна посівних площ порівняно з попереднім роком, %	Індивідуальні індекси посівних площ (is)
Пшениця	127900	+5	1,050
Жито	34400	-10	0,900
Ячмінь	20500	+16	1,160
Всього	182800	x	x

$$I_S = \frac{\sum i_S * y_0 S_0}{\sum y_0 S_0}$$

$$= \frac{1{,}050 * 127900 + 0{,}900 * 34400 + 1{,}160 * 20500}{182800} = 1{,}034$$

Тобто ріст заробітної плати на 3,4%.

Аналогічно через індивідуальні індекси кількості товару та ціни можуть бути виражені загальні індекси Ласпейреса та Пааше.

Динаміка заробітної плати:

Цех	Фонд заробітної плати у звітному періоді, тис. грн. (f_1T_1)	Зміна рівня заробітної плати порівняно з базисним періодом, %	Індивідуальні індекси заробітної плати (i_f)
№1	25,4	+15	1,150
№2	17,3	+7	1,070
№3	19,6	+25	1,250
Всього	62,3	х	х

Середньогармонійний індекс заробітної плати:

$$I_f = \frac{\sum f_1 T_1}{\sum \frac{f_1 T_1}{i_f}} = \frac{62,3}{\frac{25,4}{1,150} + \frac{17,3}{1,070} + \frac{19,6}{1,250}} = 1,156$$

Тобто ріст на 15,6%.

Розглянуті вище загальні індекси можуть застосовуватись і при вивченні однорідних об'єктів (наприклад, підприємств, що реалізують один і той же товар). І тут динаміку загальної кількості товару можна показати безпосередньо, тобто. окремо від динаміки цін. Аналіз якісних показників з певним відривом від цін здійснюється за допомогою індексу змінного складу, постійного складу та структурних зрушень.

Між названими трьома індексами існує взаємозв'язок: індекс змінного складу дорівнює добутку індексу постійного складу та індексу структурних зрушень. Отже,

$I_{\bar{p}} = I_p \times I_{сз}; I_{\bar{z}} = I_z \times I_{сз}; I_{\bar{f}} = I_f \times I_{сз}.$

Індекс змінного складу характеризує зміну у процентах середнього значення якісного показника у звітному періоді порівняно з базисним під впливом двох чинників разом. Цей індекс складається з двох дробів, причому перший дріб містить значення якісного та кількісного показників у звітному періоді, а другий – у базисному, тобто індекс є відношенням звітного середнього значення показника до базисного. Наприклад,

індекс ціни змінного складу $I_{\overline{p}} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\overline{p_1}}{\overline{p_0}}$;

індекс собівартості змінного складу $I_{\overline{z}} = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{\overline{z_1}}{\overline{z_0}}$;

індекс зарплати змінного складу $I_{\overline{f}} = \frac{\sum f_1 T_1}{\sum T_1} : \frac{\sum f_0 T_0}{\sum T_0} = \frac{\overline{f_1}}{\overline{f_0}}$.

Індекс постійного складу показує зміну (в %) середнього значення показника під впливом одного фактору – динаміки його індивідуальних значень. У цьому індексі індексується (змінюється) якісний показник, а кількісний фіксується на рівні звітного періоду. Наприклад,

індекс ціни постійного складу: $I_p = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$;

індекс собівартості постійного складу: $I_z = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_1}{\sum q_1} = \frac{\sum z_1 q_1}{\sum z_0 q_1}$;

індекс зарплати постійного складу: $I_f = \frac{\sum f_1 T_1}{\sum T_1} : \frac{\sum f_0 T_1}{\sum T_1} = \frac{\sum f_1 T_1}{\sum f_0 T_1}$.

Як і інших системах індексів можемо визначити зміни в абсолютному вираженні.

(загальна): $\Delta = \frac{\sum p_1 q_1}{\sum q_1} - \frac{\sum p_0 q_0}{\sum q_0} = \bar{p}_1 - \bar{p}_0$;

за рахунок зміни складу: $\Delta p = \frac{\sum p_1 q_1}{\sum q_1} - \frac{\sum p_0 q_1}{\sum q_1} = \bar{p}_1 - \frac{\sum p_0 q_1}{\sum q_1}$;

за рахунок зрушень: $\Delta \text{сз} = \frac{\sum p_0 q_1}{\sum q_1} - \frac{\sum p_0 q_0}{\sum q_0} = \frac{\sum p_0 q_1}{\sum q_1} - \bar{p}_0$.

Індекс структурних зрушень показує, на скільки процентів змінилося середнє значення показника під впливом змін у структурі сукупності. У даному випадку індексується кількісний показник, а якісний фіксується на рівні базисного періоду.

індекс структурних зрушень ціни: $I_{сз} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}$;

індекс структурних зрушень собівартості: $I_{сз} = \frac{\sum z_0 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0}$;

індекс структурних зрушень зарплати: $I_{сз} = \frac{\sum f_0 T_1}{\sum T_1} : \frac{\sum f_0 T_0}{\sum T_0}$.

$$I_{\bar{z}} = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{15530}{350} \div \frac{13650}{300} = \frac{44{,}37}{45{,}5} = 0{,}975$$

$$I_z = \frac{\sum z_1 q_1}{\sum q_1} : \frac{\sum z_0 q_1}{\sum q_1} = \frac{15530}{350} \div \frac{15750}{350} = \frac{44{,}37}{45} = 0{,}986$$

$$I_{сз} = \frac{\sum z_0 q_1}{\sum q_1} : \frac{\sum z_0 q_0}{\sum q_0} = \frac{15750}{350} \div \frac{13650}{300} = \frac{45}{45{,}5} = 0{,}989$$

Продукція	Собівартість, грн.		Кількість, тис. шт.		$z0q0$	$z0q1$	$z1q1$
	$z0$	$z1$	$q0$	$q1$			
А	35	37	105	130	3675	4550	4810
Б	45	44	75	90	3375	4050	3960
В	55	52	120	130	6600	7150	6760
Разом	**х**	**х**	**300**	**350**	**13650**	**15750**	**15530**

середня собівартість знизилась на 2,5% ;

під впливом зміни індивідуальних значень — на 1,4%

а за рахунок структурних зрушень на 1,1%.

Приріст середньої собівартості:

загальний $\Delta = \bar{Z}_1 - \bar{Z}_0 = 44{,}37 - 45{,}5 = -1{,}13$грн. ;

за рахунок зміни собівартості: $\Delta z = 44{,}37 - 45 = -0{,}63$грн. ;

за рахунок зрушень: Δсз $= 41{,}11 - 41{,}52 = -0{,}41$грн.

Факторний індексний аналіз використовується для вивчення впливу окремих факторних показників на результативний показник з допомогою системи взаємозв'язаних індексів. При цьому результативний показник функціонально залежить від факторних показників та дорівнює їх добутку:

$$y = a \times b \times c \times d.$$

При побудові індексів використовується наступне правило: індексований показник у чисельнику за звітний період, у знаменнику — за базисний, показники, які знаходяться перед індексованим показником фіксуються на рівні звітного періоду, а ті, що розташовані після індексованого показника — на рівні базисного періоду.

$$I_y = \frac{\sum a_1 b_1 c_1 d_1}{\sum a_0 b_0 c_0 d_0}; I_a = \frac{\sum a_1 b_0 c_0 d_0}{\sum a_0 b_0 c_0 d_0}; I_b = \frac{\sum a_1 b_1 c_0 d_0}{\sum a_1 b_0 c_0 d_0};$$

$$I_c = \frac{\sum a_1 b_1 c_1 d_0}{\sum a_1 b_1 c_0 d_0}; I_d = \frac{\sum a_1 b_1 c_1 d_1}{\sum a_1 b_1 c_1 d_0}.$$

Існує взаємозв'язок: $Iy = Ia \times Ib \times Ic \times Id.$

На основі обчисленої системи індексів можна визначити загальний абсолютний приріст результативного показника та факторні прирости, зумовлені впливом кожного фактора зокрема. Ці прирости визначаються як різниця між чисельником та знаменником відповідного індексу:

$$\Delta y = \sum a_1 b_1 c_1 d_1 - \sum a_0 b_0 c_0 d_0;$$

$$\Delta a = \sum a_1 b_0 c_0 d_0 - \sum a_0 b_0 c_0 d_0;$$

$$\Delta b = \sum a_1 b_1 c_0 d_0 - \sum a_1 b_0 c_0 d_0;$$

$$\Delta c = \sum a_1 b_1 c_1 d_0 - \sum a_1 b_1 c_0 d_0;$$

$$\Delta d = \sum a_1 b_1 c_1 d_1 - \sum a_1 b_1 c_1 d_0;$$

$$\Delta y = \Delta a + \Delta b + \Delta c + \Delta d.$$

Приклад розрахунку та економічної інтерпретації індексів:

Продукція	Витрати сировини на одиницю продукції (питомі витрати), кг		Ціна **1** кг сировини, грн.		Кількість виробленої продукції, шт.	
	I кв. (m_0)	I*V* кв. (m_1)	I кв. (p_0)	I*V* кв. (p_1)	I кв. (q_0)	I*V* кв. (q_1)
А	20	22	100	160	40	50
Б	17	13	220	270	100	140
В	40	45	350	360	200	170

$$y = m \times p \times q.$$

Індекс витрат сировини на одиницю продукції

(питомих витрат):

$$I_m = \frac{\sum m_1 p_0 q_0}{\sum m_0 p_0 q_0} = \frac{22\times100\times40+13\times220\times100+45\times350\times200}{20\times100\times40+17\times220\times100+40\times350\times200} = \frac{3524000}{3254000} = 1{,}083.$$

Індекс цін на сировину:

$$I_p = \frac{\sum m_1 p_1 q_0}{\sum m_1 p_0 q_0} = \frac{22 \times 160 \times 40 + 13 \times 270 \times 100 + 45 \times 360 \times 200}{3524000} = \frac{3731800}{3254000} = 1{,}059$$

Індекс кількості (фізичного обсягу) продукції:

$$I_q = \frac{\sum m_1 p_1 q_1}{\sum m_1 p_1 q_0} = \frac{22\times100\times45+13\times220\times140+45\times360\times170}{3731800} = \frac{3421400}{3731800} = 0{,}917.$$

Індекс витрат на виробництво:

$$I_y = \frac{\sum m_1 p_1 q_1}{\sum m_0 p_0 q_0} = \frac{3421400}{3254000} = 1{,}051.$$

Отже, витрати на виробництво зросли на 5,1%,

в тому числі

за рахунок збільшення витрат сировини на одиницю продукції — на 8,3%,

за рахунок зростання цін — на 5,9%,

а під впливом зменшення кількості виробленої продукції витрати зменшилися на 8,3%.

Визначимо абсолютний приріст витрат на виробництво:

$\Delta y = \sum m_1 p_1 q_1 - \sum m_0 p_0 q_0 = 3421400 - 3254000 = 167400(грн.)$.

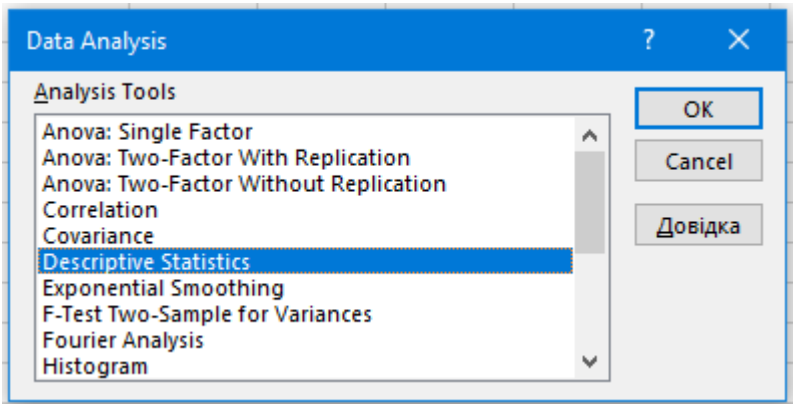
Прирости за рахунок зміни:
витрат сировини на одиницю продукції:
$\Delta m = \sum m_1 p_0 q_0 - \sum m_0 p_0 q_0 = 3524000 - 3254000 = 270000(грн.)$;
цін на сировину:
$\Delta y = \sum m_1 p_1 q_0 - \sum m_1 p_0 q_0 = 3731800 - 3524000 = 207800(грн.)$;
кількості (фізичного обсягу) продукції:
$\Delta q = \sum m_1 p_1 q_1 - \sum m_1 p_1 q_0 = 3421400 - 3731800 = -310400(грн.)$.
При цьому $\Delta y = 270000 + 207800 - 310400 = 167400(\text{грн.})$.

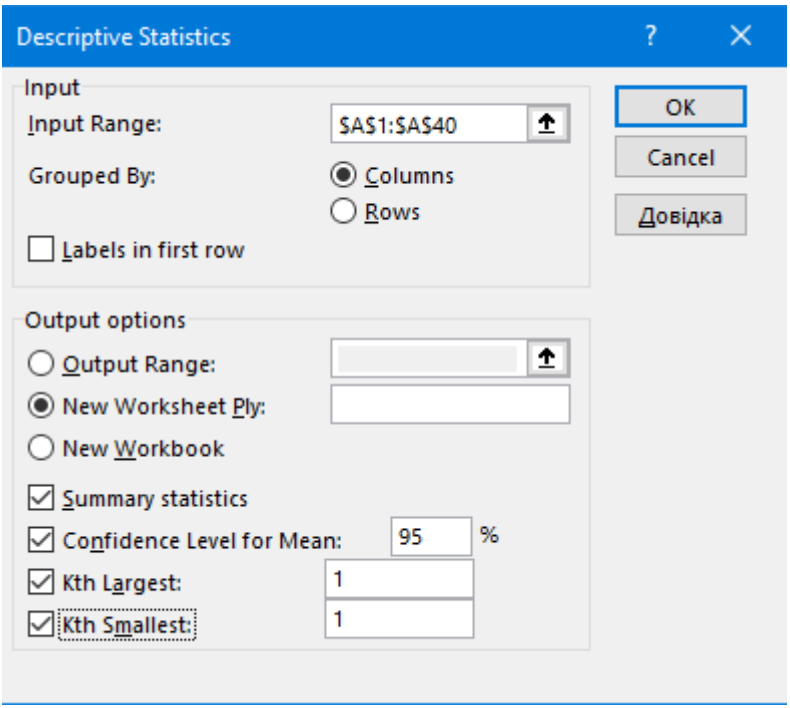
На практиці для статистичного аналізу даних використовують різні інструменти, як то табличні процесори, спеціалізовані пакети аналізу тощо.

Задано дискретний ряд.

1	19	43	87
2	20	45	87
2	20	53	88
3	22	55	89
3	23	62	91
14	31	65	96
15	33	67	96
18	35	72	97
19	36	72	102
19	39	82	103

Знайти його загальні статистичні характеристики за допомогою надбудови Excel «Аналіз даних».





Column1	
Mean	48,15
Standard Error	5,330073
Median	41
Mode	19
Standard Deviation	33,71034
Sample Variance	1136,387
Kurtosis	-1,40906
Skewness	0,217978
Range	102
Minimum	1
Maximum	103
Sum	1926
Count	40
Largest(1)	103
Smallest(1)	1
Confidence Level(95,0%)	10,78109

Для більш глибокого аналізу використовуються спеціалізовані продукти: Minitab, StatSoft (STATISTICA), COMSOL, SAS (програмне забезпечення для статистичного аналізу), MATLAB, SPSS (IBM), XL STAT.

Література до 2 частини.

1. Про державну статистику: Закон України // Голос України. – 1992. – 21 жовтня 1992.
2. Про заходи щодо розвитку державної статистики: Указ президента України від 22 листопада 1997 р. № 1299/97 // Статистика України. – 1998. - №1.

3. Програма реформування державної статистики на період до 2010 року: Постанова кабінету Міністрів України № 971 від 27.06.1998р.
4. Про концепцію побудови національної статистики України та державну програму переходу на міжнародну систему обліку і статистики. Затверджено постановою Кабінету Міністрів України від 4 травня 1993 р. № 326 // Зібрання постанов Уряду України. – 1994. - №2.
5. Большой экономический словарь / Под. ред. А.Н. Азрилияна. – 2-е изд., доп. И пере раб. – М.: Институт новой экономики, 1997.
6. Бараник З.П. Статистика: Навчальний посібник./ З.П. Бараник. – К.:2005. – 268с.
7. Бек В.Л. Теорія статистики: Навчальний посібник / В.К. Бек – К.: ЦУЛ, 2003. – 288 с.
8. Герасименко С.С. Статистика: Підручник / С.С. Герасименко, А.В. Головач, – К.: КНЕУ, 2000.
9. Гаркавий В.К. Статистика: Підручник / В.К. Гаркавий. – К.: Вища школа, 1995. – 415 с.
10. Двірник В.М., Корчажнікова Л.П. Теорія статистики: Навчальний посібник. / В.М. Двірник, Л.П. Корчажнікова. – Дніпропетровськ: ДАУБП, 2000, - 160с.

Навчальне видання

ТКАЛІЧЕНКО Сергій Володимирович
СУПРУН Анатолій Анатолійович

СТАТИСТИКА