МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

О. В. ПЕРЕГУДА
О. А. КАПУСТЯН
О. Б. КУРИЛКО

СТАТИСТИЧНА ОБРОБКА ДАНИХ

Навчальний посібник

КИЇВ 2022

Рецензенти:

доктор фізико-математичних наук О.А. Бурилко
доктор фізико-математичних наук К.В. Ральченко

Затверджено
вченою радою механіко-математичного факультету
(протокол №9 від 3 лютого 2022 року)

Перегуда О.В.
Статистична обробка даних: навч. посіб. / О.В. Перегуда,
О.А. Капустян, О.Б. Курилко.–Електронне видання, 2022.–103 с.

У навчальному посібнику, що відповідає загальноосвітнім стандартам, викладено основні методи статистичної обробки даних. Основною метою посібника є демонстрація використання методів статистичного дослідження, знаходження статистичних показників та інтерпретація їх значень при вирішенні економічних завдань, формування у студентів теоретичних знання і практичних навичок для аналізу досліджуваного об'єкта, розробки обґрунтованих рішень та рекомендацій.

Рекомендовано для студентів, аспірантів економічних спеціальностей, зокрема для економістів-міжнародників.

ЗМІСТ

ПЕРЕДМОВА

Розвиток сучасних наукових досліджень характеризуються їх математизацією, що виражається у використанні математичних методів і моделей. В прикладних науках використовуються різноманітні статистичні дані для побудови відповідних моделей об'єктів, явищ, закономірностей і процесів. Статистика стає основою досліджень даних і дає інструментарій, що дозволяє представити усю логічну послідовність обробки отриманої інформації. Статистика дає стислу й чітку характеристику досліджуваного масового явища і в той же час можливість виявити та дослідити тенденції його розвитку.

Змістом навчальної посібника "Статистична обробка даних" є найпоширеніші методи кількісного статистичного аналізу; методи моделювання та аналізу впливу певних факторів на становище об'єкта або явища. Навчальний посібник знайомить студентів з основними статистичними методами обробки даних, побудови та аналізу невідомих показників, моделей, що відображають можливості подальшого встановлення відповідних статистичних закономірностей функціонування систем різних типів.

В результаті вивчення навчальної дисципліни студенти повинні:

− знати основні кількісні та якісні методи і засоби аналізу закономірностей детермінованих і стохастичних систем прикладного характеру;

− вміти застосовувати імовірнісні методи при розв'язуванні прикладних задач, що враховують випадковий характер реальних явищ.

Основною метою навчального посібника «Статистична обробка даних» є формування у студентів знань статистичних методів моделювання та вмінь створення і використання математично-статистичних моделей у розв'язанні прикладних завдань фахового спрямування.

У першому та другому розділах посібника надано основні поняття з математичної статистики і стандартні прийоми

перевірки статистичних гіпотез, які необхідні для засвоєння навчальної дисципліни.

У третьому та четвертому розділах посібника розглянуто методи кореляційного і регресійного аналізів, що є основним інструментом обробки результатів прикладних досліджень.

П’ятий розділ присвячено питанням, які найчастіше виникають при аналізі результатів прикладних досліджень та дещо ускладнюють роботу дослідника.

У шостому розділі наведено приклад побудови та дослідження економіко-математичної моделі на основі статистичних даних.

Кожний розділ навчального посібника містить значну кількість прикладів застосування відповідних методів у розв’язанні прикладних завдань. Представлено реалізацію методів засобами Microsoft Excel.

РОЗДІЛ 1
ОСНОВНІ ПОНЯТТЯ

1.1. Поняття вибіркового методу в статистиці

Одним з розділів прикладної математики є математична статистика, предметом якого є розробка раціональних прийомів і методів отримання, опису та обробки експериментальних даних з метою вивчення закономірностей масових випадкових явищ.

Основними завданнями математичної статистики є:

– визначення за статистичними даними законів розподілу випадкових величин;

– визначення за статистичними даними параметрів розподілу випадкових величин;

– визначення за статистичними даними виду зв'язку між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкта);

– визначення сили (тісноти зв'язку) між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкта);

– перевірка вірогідності статистичних гіпотез;

– розробка рекомендацій щодо проведення експерименту та обробки його результатів.

У прикладних дослідженнях, зазвичай, необхідно вивчити сукупність однорідних об'єктів або спостережень за певною кількісною або якісною ознакою.

Сукупність об'єктів або спостережень, всі елементи якої підлягають вивченню при статистичному аналізі, називається *генеральною сукупністю***.**

Генеральна сукупність може бути скінченою або нескінченною. Так, при вивченні розподілу населення за родом занять розглядається велика, але скінчена генеральна сукупність об'єктів. При вивченні впливу яскравості освітлення робочого місця на продуктивність праці працівника генеральна сукупність спостережень теоретично нескінченна, оскільки яскравість освітлення може змінюватися неперервно у межах певного інтервалу.

Кількість об'єктів (спостережень) генеральної сукупності називається *об'ємом генеральної сукупності* і позначається *N*.

На практиці майже не реально досліджувати кожний елемент генеральної сукупності, оскільки це пов'язано з великими витратами засобів, коштів і часу. У деяких випадках дослідити всі об’єкти генеральної сукупності взагалі неможливо. Тому при статистичному аналізі, як правило, вивчається не вся

генеральна сукупність, а деяка її частина.

Частина об'єктів генеральної сукупності, що використовується для дослідження, називається *вибіркою*.

Кількість об'єктів (спостережень) вибірки називається її *об'ємом* і позначається *n*.

Наприклад, продукція у кількості *N* одиниць, вироблена підприємством протягом року, є генеральною сукупністю. Для дослідження якості продукції на практиці розглядається вибірка, що складається з *n* одиниць продукції.

Ознакою якості в даному дослідженні служить відповідність вибраної одиниці товару сертифікатним вимогам.

Мета вибіркового методу в статистиці полягає в тому, що висновки, зроблені на основі вивчення вибірки, можуть бути розповсюджені на всю генеральну сукупність.

Необхідно зазначити, що незалежно від способу утворення вибірки вона повинна правильно відображати кількісні та якісні співвідношення генеральної сукупності, тобто бути *репрезентативною*. Крім того, всі елементи генеральної сукупності повинні мати однакову ймовірність бути відібраними у вибірку, тобто вибірка має бути *випадковою*.

Існує кілька типів ймовірнісної вибірки, які відрізняються між собою характером використаних дослідником прийомів:

– проста ймовірнісна вибірка, яка проводиться шляхом випадкового відбору об’єктів у вибірку;

– стратифікована вибірка, що використовується тоді, коли цілі та завдання дослідження вимагають відбору об’єктів для вивчення за певними груповими критеріями;

– багатоступінчаста вибірка, для якої характерно декілька послідовних змін одиниць відбору.

Для результатів, що отримані при вибірковому дослідженні, необхідна перевірка на точність і статистичну значущість; спосіб формування вибірки та її об'єм повинні відповідати певному методу обробки даних.

1.2. Шкали вимірювань

Статистичній обробці підлягають тільки ті ознаки об'єктів або фактори, які можна виміряти за деякою шкалою.

Шкала − числова система, що відображає досліджувані властивості та ознаки об'єкта.

Існують такі шкали вимірювань: шкала найменувань (номінальна); шкала порядку; шкала інтервалів; шкала відношень.

Шкала найменувань (класифікації, номінальна).

Якщо дані вимірюються за шкалою найменувань, то над ними можливі тільки операції порівняння, такі як „рівні" або „нерівні". Дані номінальної шкали необхідні для ідентифікації певного об'єкта – місце розташування деякої організації, номер філіалу деякої фірми, номер методики навчання і т. ін.

Шкала порядку.

Якщо дані вимірюються за шкалою порядку, їх можна порівняти за величиною „більше", „менше" або „рівні". За такою шкалою вимірюються, наприклад, експертні оцінки.

Шкала інтервалів.

Якщо дані вимірюються за шкалою інтервалів, до них можна застосувати операції: порівняння − „більше", „менше", „рівні"; додавання і віднімання. Прикладом даних, які належать до цієї шкали, є результати вимірювання температури, тиску.

Шкала відношень.

Якщо дані вимірюються за шкалою відношень, їх можна порівняти за величиною та виконати всі арифметичні операції: додавання, віднімання, множення і ділення. Такою шкалою кодується вага, маса, ріст, довжина, дохід, обсяг виробництва.

В емпіричних дослідженнях найбільш використовуваною вважається шкала порядкового типу.

Від шкали вимірювання даних залежить можливість знаходження числових характеристик (табл. 1.1) і застосовування певного статистичного методу.

Таблиця 1.1 – Шкали вимірювань і числові характеристики

Назва шкали	**Числові характеристики**
Найменувань	Частоти
Порядку	Частоти, середнє, мода, медіана
Інтервалів	Частоти, середнє, мода, медіана, дисперсія
Відношень	Всі відомі

Для об'єктивного визначення типу шкали користуються способом її створення на основі шкали Лайкерта. Ця методика має ще назву «метод сумарних оцінок». На першому етапі група дослідників розробляє набір (навіть до сотні одиниць) роздумів відносно предмета вивчення. Наступним етапом є пілотажне дослідження на невеликій вибірці. Після отримання його результатів проводиться якісний аналіз набору даних і відсіюються малоефективні значення ознак.

1.3. Статистичні ряди

Припустимо, що необхідно вивчити деяку ознаку генеральної сукупності X, для чого було проведено n вимірювань цієї ознаки і складено вибірку її значень $\{x_1, x_2, ..., x_n\}$ об'єму n.

Елементи вибірки називаються *варіантами*. Число n_i, що показує, кількість появ варіанти x_i у вибірці, називається *частотою варіанти*. Число w_i, що дорівнює відношенню частоти варіанти n_i до об'єму вибірки n, називається *відносною частотою варіанти* x_i:

$$w_i = \frac{n_i}{n}.$$

Ряд варіант, розташованих в порядку зростання їх значень, називається *варіаційним рядом*. Послідовність, що складається з

варіант і відповідних їм частот (відносних частот), називається *статистичним рядом* або *рядом розподілу*. Групування кількісних результатів вимірювань у вигляді статистичних рядів є необхідним для застосування статистичних методів аналізу даних і побудови статистичних моделей.

Ознака X є випадковою величиною, а статистичний ряд – емпіричним (тобто отриманим у результаті експерименту або спостережень) законом її розподілу.

Статистичний ряд називається *дискретним*, якщо він є законом розподілу дискретної випадкової величини, та *інтервальним*, якщо він є законом розподілу неперервної випадкової величини.

Дискретний статистичний ряд у загальному вигляді можна зобразити у вигляді таблиці:

Варіанти x_i	x_1	x_2	$\cdots$	x_k
Частоти n_i (відносні частоти w_i)	$n_1(w_1)$	$n_2(w_i)$	$\cdots$	$n_k(w_k)$

де k – кількість варіант.

Інтервальний статистичний ряд у загальному вигляді можна представити у вигляді таблиці

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

де k – кількість інтервалів.

Для статистичних рядів виконуються рівності:

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

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

$$k = 1 + 1,4 \ln n .$$

Довжина кожного із інтервалів Δ розраховується за формулою

$$\Delta = \frac{x_{\max} - x_{\min}}{k},$$

де $x_{\max}, x_{\min}$ – максимальне і мінімальне значення у варіаційному ряді.

Підраховуючи кількість значень варіант, що потрапили в інтервал $\left[a_i; a_{i+1}\right)$, отримують відповідні частоти (відносні частоти).

Питання для самоконтролю

1. Які основні завдання вирішуються методами математичної статистики?
2. Що називається генеральною сукупністю і вибіркою?
3. У чому полягає вибірковий метод у статистиці?
4. Які вимоги висуваються до вибірки?
5. Що називається статистичним рядом?
6. Які види статистичних рядів?
7. Який алгоритм побудови статистичного ряду?

РОЗДІЛ 2
ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

2.1. Поняття про статистичні гіпотези

Дані вибіркових спостережень часто становлять основу для прийняття одного з кількох альтернативних рішень (продукція може бути бракованою або якісною, точність обробки виробу в межах норми або нижча від норми і т. д.). Із загальнометодологічного погляду тут йдеться про висунення деякої гіпотези, яку відхиляють або приймають після проведення деякого експерименту. Якщо цей експеримент має статистичний (стохастичний) характер, кажуть, що гіпотеза є *статистичною*.

Перевірка гіпотези про вид закону розподілу досліджуваної величини має велике значення для прикладних досліджень. Необхідність такої перевірки виникає при виборі критерію, оскільки для багатьох з них висувається вимога нормального розподілу статистичних даних

Означення. *Статистичною* називають гіпотезу про властивості генеральної сукупності, що перевіряється на основі вибірки.

Статистичними гіпотезами можуть бути, наприклад, такі твердження: розподіл імовірностей випадкової величини є нормальний; розподіл ймовірностей випадкової величини є пуассонівський; у нормальному розподілі випадкової величини параметри $a = 20$ і $\sigma = 1{,}5$; у показниковому розподілі випадкової величини параметр $\alpha = 5$; випадкові величини X і Y незалежні і т. д.

У математичній статистиці виділяють два основні типи статистичних гіпотез:

• гіпотези про закон розподілу ймовірностей випадкової величини (ознаки генеральної сукупності);

• гіпотези про значення параметрів розподілу випадкової величини (ознаки генеральної сукупності).

Статистичні гіпотези першого типу називають *непараметричними*, а другого типу – *параметричними*.

Означення. *Основною (нульовою*) називають висунуту гіпотезу і позначають H_0.

Альтернативною (конкуруючою) називають гіпотезу, яка повністю або частково логічно заперечує нульову гіпотезу, і позначають H_1 .

Наприклад, математичне сподівання нормально розподіленої випадкової величини $a = 10$ − основна гіпотеза; математичне сподівання нормально розподіленої випадкової величини $a \neq 10$ − альтернативна гіпотеза. Записують це так: H_0: $a = 10$; H_1: $a \neq 10$.

Задача про статистичну перевірку статистичних гіпотез формулюється так. Розглядають деяку гіпотезу про те, що розподіл ймовірностей деякої випадкової величини має той чи інший вигляд, або параметри розподілу мають ті чи інші значення.

Задача полягає у тому, щоб на основі вивчення статистичних даних (вибірки) підтвердити справедливість висунутої гіпотези чи спростувати її. При цьому вказується також ймовірність того, що прийняте рішення є правильним або помилковим. Проблема зменшення ймовірності того, що прийняте рішення є помилковим, є також однією із задач математичної статистики.

У результаті статистичної перевірки гіпотези може бути прийняте одне з двох правильних рішень: 1) гіпотеза приймається і вона істинна; 2) гіпотеза відхиляється і вона неістинна.

Поряд із тим у результаті статистичної перевірки статистичної гіпотези можуть бути допущені помилки (прийняті неправильні рішення) двох типів: 1) гіпотеза відхиляється, але вона істинна (помилка першого роду); 2) гіпотеза приймається, але вона неістинна (помилка другого роду).

Виявляється, що помилка першого роду має вагоміші наслідки, ніж помилка другого роду.

Щоб застерегти від помилки першого роду або принаймні звести до мінімуму ризик її допущення, вводиться спеціальне число α , яке виражає ймовірність відхилення правильної гіпотези.

Означення. Ймовірність допущення помилки першого роду називають *рівнем значущості* і позначають через α.

Число α задають наперед і найчастіше його вибирають рівним 0,1; 0,05; 0,01. Якщо $\alpha = 0,05$, то це означає, що ймовірність допустити помилку першого роду є мала, а саме – ми ризикуємо її допустити у 5-ти випадках зі 100.

2.2. Критерій статистичної перевірки гіпотези

Для перевірки нульової гіпотези вводять певну числову характеристику, яку обчислюють на основі вибірки і на підставі якої вирішують: прийняти основну гіпотезу чи альтернативну. Зрозуміло, що вибрана числова характеристика для різних вибірок матиме, загалом кажучи, різні значення, і тому вона є випадковою величиною.

Означення. *Статистичним критерієм* гіпотези (або просто *критерієм* гіпотези) називається випадкова величина *K*, за допомогою якої проводиться перевірка гіпотези.

Випадкову величину *K* вибирають такою, щоб її закон розподілу ймовірностей був відомий.

Означення. Значення випадкової величини *K*, обчислене на основі даних певної вибірки, називають *емпіричним* значенням критерію гіпотези і позначають $K_{емп.}$

Виявляється, що за одних значень $K_{емп}$ гіпотеза H_0 приймається, а за інших його значень – відхиляється.

Означення. Сукупність значень критерію *K*, за яких нульова гіпотеза H_0 відхиляється, називається *критичною областю*, а сукупність значень критерію *K*, за яких нульову гіпотезу H_0 приймають, називається *областю прийняття гіпотези.*

Отже, маємо таке *правило перевірки статистичних гіпотез:* якщо емпіричне значення критерію $K_{емп}$ належить критичній області, то нульову гіпотезу H_0 відхиляють; якщо емпіричне значення критерію K_0 належить області прийняття гіпотези, то нульову гіпотезу H_0 приймають.

Якщо випадкова величина *K* є одновимірна, то критична область, як правило, є множиною точок певних інтервалів на прямій, які відокремлені від області прийняття гіпотези так

званими критичними точками $k_{кр}$. Тобто для знаходження критичної області достатньо визначити критичні точки.

Залежно від конкуруючої гіпотези розглядають три типи критичних областей:

• *правостороння критична область* – це та область на числовій прямій, що визначається нерівністю $K > k_{кр}$;

• *лівостороння критична область* – це та область на числовій прямій, що визначається нерівністю $K < k_{кр}$;

• *двостороння критична область* – це та область на числовій прямій, що визначається сумою інтервалів $K < -k_{кр}$ і $K > k_{кр}$.

Для знаходження критичної області, задають рівнь значущості α і шукають критичні точки $k_{кр}$ з наступних співвідношень:

1) для правосторонньої критичної області:

$$P\{K > k_{kp}\} = \alpha \quad (k_{kp} > 0);$$

2) для лівосторонньої критичної області:

$$P\{K < k_{kp}\} = \alpha \quad (k_{kp} < 0);$$

3) для двосторонньої симетричної критичної області:

$$P\{K > k_{kp}\} = \frac{\alpha}{2} \quad (k_{kp} > 0).$$

Зрозуміло, що для певної гіпотези можна побудувати різні критерії її перевірки, і за кожним критерієм можемо одержувати різні результати щодо прийняття нульової гіпотези H_0 на основі тієї самої вибірки. Тому для визначення більш достовірного критерію вводиться характеристика, яка називається *потужністю* критерію.

Означення. *Потужністю критерію* називають ймовірність потрапляння критерію у критичну область за умови, що конкуруюча гіпотеза є істинною.

Тобто, потужність критерію визначається як імовірність не допустити помилку другого роду при вибраному критерії.

2.3. Перевірка гіпотези про вид закону розподілу досліджуваної величини. Критерій згоди Пірсона

Критерієм згоди називають статистичний критерій перевірки гіпотези про закон розподілу ймовірностей випадкової величини (ознаки генеральної сукупності). Є кілька критеріїв згоди: критерій Колмогорова, критерій Смірнова, критерій Пірсона та інші. Розглянемо критерій згоди Пірсона (критерій χ^2), який ґрунтується на порівнянні емпіричних і теоретичних частот.

Припустимо, що з деякої генеральної сукупності X, яка розглядається як випадкова величина, обрана вибірка $\{x_1, x_2, \ldots, x_n\}$.

Нехай висунуто гіпотезу H_0: *випадкова величина X має певний закон розподілу A*.

Здійснивши вибірку обсягу n, знаходять і записують у вигляді таблиці інтервальний статистичний розподіл частот:

$[z_{i-1}, z_i)$	$[z_0, z_1)$	$[z_1, z_2)$	…	$[z_{m-1}, z_m)$
n_i	n_1	n_2	…	n_m

або у вигляді таблиці дискретного статистичного розподілу частот

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

$$\sum_{i=1}^{m} n_i = n .$$

Оскільки перевіряється гіпотеза про те, що розподіл ознаки X генеральної сукупності описується певною (конкретною) функцією розподілу $F(x)$, або, що те ж саме, щільністю розподілу $f(x)$, то для кожного інтервалу $[z_{i-1}, z_i)$ можна визначити теоретичні ймовірності p_i попадання значень

випадкової величини X у цей інтервал, а отже, і теоретичні частоти $n' = np_i$.

Для обчислення ймовірностей p_i використовують формули:

$$p_i = P\{z_{i-1} \le X \le z_i\} = F(z_i) - F(z_{i-1}), \quad i = \overline{1,m}. \qquad (2.1)$$

Зазначимо, що для обчислення ймовірностей p_1 і p_m у формулах (2.1) покладають, відповідно, $z_0 = -\infty$ і $z_m = +\infty$. Тоді $\sum_{i=1}^{m} p_i = 1$.

Отримані результати обчислень зручно записати у формі таблиці:

$[z_{i-1}, z_i)$	$[z_0, z_1)$	$[z_1, z_2)$	…	$[z_{m-1}, z_m)$
n_i	n_1	n_2	…	n_m
p_i	p_1	p_2		p_m
$n_i' = np_i$	n_1'	n_2'		n_m'

Згідно з критерієм Пірсона для перевірки гіпотези H_0 вводиться випадкова величина (статистика) K:

$$K = \chi^2 = \sum_{i=1}^{m} \frac{\left(n_i - n_i'\right)^2}{n_i'}, \qquad (2.2)$$

де m – кількість груп у статистичному розподілі вибірки; n_i – емпірична частота ознаки X в i-тій групі; $n_i' = np_i$ – теоретична частота; p_i – ймовірність того, що значення X належить i-тій групі.

З формули (2.2) маємо, що у випадку, коли відповідні теоретичні та емпіричні частоти співпадають, то значення $\chi^2 = 0$. Тобто, чим ближче значення χ^2 до нуля, тим краще узгоджуються вибіркові дані та обраний теоретичний закон розподілу.

Відомо, що при $n \to \infty$ закон розподілу статистики K прямує до закону розподілу χ^2 з $k = m - r - 1$ ступенями вільності, де m – кількість груп у статистичному розподілі вибірки; r − кількість параметрів гіпотетичного розподілу A (наприклад, $r = 2$ для нормального розподілу, $r = 1$ для розподілу Пуассона, $r = 0$ для рівномірного розподілу).

Для критерію χ^2 будують правосторонню критичну область за правилом:

$$P\left\{\chi^2 > \chi^2_{кр}\right\} = \alpha . \qquad (2.3)$$

За заданим рівнем значущості α і кількістю ступенів вільності k по таблиці критичних точок розподілу χ^2 (в якій дано розв'язки рівняння (2.3)) знаходять критичну точку $k_{кр} = \chi^2_{кр}(\alpha, k)$.

На підставі даних вибірки, записаних у таблиці, обчислюють емпіричне значення критерію Пірсона:

$$K_{емп} = \sum_{i=1}^{m} \frac{\left(n_i - n_i'\right)^2}{n_i'} .$$

Порівнюємо значення $K_{емп}$ і $k_{кр}$: якщо , $K_{емп} \geq k_{кр}$ то гіпотезу H_0 відхиляють; якщо ж , $K_{емп} < k_{кр}$ то гіпотезу H_0 приймають.

Зауваження:

1) У деяких статистичних таблицях критичне значення χ^2 задається залежно від рівня вірогідності γ , а $\gamma = 1 - \alpha$.

2) Критичну точку $k_{кр} = \chi^2_{кр}(\alpha, k)$ можна визначити за допомогою вбудованої статистичної функції Excel ХИ2ОБР(α, l). Параметрами функції ХИ2ОБР є: α – рівень значущості; l – ступінь свободи, $l = k - r - 1$, де k – кількість груп емпіричного розподілу, r – кількість параметрів теоретичного розподілу.

Застосування критерію χ^2 вимагає дотримання наступних умов:

1) експериментальні дані мають бути незалежними, тобто вибірка має бути випадковою; 2) обсяг вибірки має бути достатньо великим (практично не меншим ніж 50 одиниць), а частота кожної групи – не меншою за 5. Якщо остання умова не виконується, то проводиться попереднє об'єднання нечисленних груп.

Критерій згоди Пірсона дає відповідь на питання, чи розбіжність між емпіричними і теоретичними частотами зумовлена випадковістю, чи вона є значущою. Як і будь-який інший критерій він *не доводить справедливості* гіпотези H_0 , а лише дозволяє встановити на прийнятному рівні значущості *узгодженість чи неузгодженість* гіпотези H_0 з даними спостережень.

Використовуючи статистичні методи обробки даних вибірки, досить часто висуваються вимоги до розподілу даних або до числових характеристик.

Приклад 2.1. За даним інтервальним статистичним рядом (табл. 2.2) знайти закон розподілу випадкової величини X.

Таблиця 2.2

$[z_{i-1}, z_i)$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5

Розв'язання. Для визначення вигляду закону розподілу побудуємо гістограму за даними табл. 2.2 (рис. 2.1). За виглядом гістограми висуваємо гіпотезу про нормальний закон розподілу даної випадкової величини:

H_0 – випадкова величина X розподілена за нормальним законом;

H_1 – випадкова величина X не розподілена за нормальним законом.

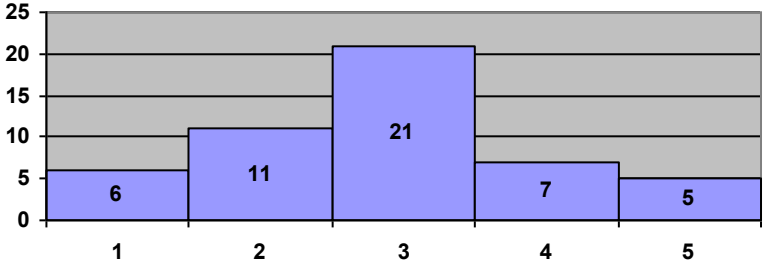


Рис. 2.1. Гістограма за даними табл. 2.2

Щільність розподілу випадкової величини, розподіленої за нормальним законом, має вигляд $f(x)=\frac{1}{2\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}$, де a і σ – параметри розподілу.

Знайдемо означені параметри, враховуючи їх точкові оцінки: $a=\bar{x}$, $\sigma^2=S^2$. Розрахунки оформимо у вигляді таблиці (табл. 2.3).

Таблиця 2.3

$[z_{i-1}, z_i)$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5
x_i	-1,6	0,8	0	0,8	1.6
$x_i n_i$	-9,6	8,8	0	5,6	8
$\left(x_i-\bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382

Обчислимо вибіркове середнє, вибіркову дисперсію і вибіркове середнє квадратичне відхилення відповідно:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i=\frac{1}{50}(-1,6\cdot 6-0,8\cdot 11+0\cdot 21+0,8\cdot 7+1,6\cdot 5)=-0,096;$$

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

;

$$= \frac{1}{50}(13,572 + 5,452 + 0,194 + 5,620 + 14,382) = 0,7844$$

$$S = \sqrt{S^2} \approx 0,886.$$

Отже, параметрами теоретичного закону розподілу є:
$a = \bar{x} = -0,096$, $\sigma = S = 0,886$.

Для знаходження значення критерію χ^2 розрахуємо теоретичні частоти n_i'. Теоретичні частоти можна знайти за формулою $n_i' = np_i$, де p_i – ймовірності попадання випадкової величини в певний інтервал. Для нормального закону розподілу означені ймовірності знаходяться за формулою

$$P\left(z_i < X < z_{i+1}\right) = p_i = \frac{1}{2}\left(\Phi\left(\frac{z_{i+1} - \bar{x}}{S}\right) - \Phi\left(\frac{z_i - \bar{x}}{S}\right)\right),$$

де $\Phi(x)$ – функція Лапласа, значення якої представлені у статистичних таблицях. Для зручності обчислень побудуємо таблицю (табл. 2.4).

За формулою (2.2) маємо: $K = \chi^2 = \sum_{i=1}^{m}\frac{\left(n_i - n_i'\right)^2}{n_i'} \approx 4,16$.

Знайдемо критичне значення $\chi^2_{\alpha,l}$, враховуючи, що $l = k - r - 1 = 5 - 2 - 1 = 2$. Рівень значущості α оберемо рівним 0,1. За допомогою Excel знаходимо ХИ2ОБР(0,1; 2) = 4,6.

Отже, оскільки $\chi^2 < \chi^2_{\alpha,l}$, то гіпотеза H_0 про нормальний розподіл приймається, гіпотеза H_1 відхиляється.

Таблиця 2.4

$[z_{i-1}, z_i)$	$[-2;-1,2)$	$[-1,2;-0,4)$	$[-0,4;0,4)$	$[0,4;1,2)$	$[1,2;2)$
n_i	6	11	21	7	5
x_i	-1,6	0,8	0	0,8	1.6
$x_i n_i$	-9,6	8,8	0	5,6	8
$\left(x_i - \bar{x}\right)^2 n_i$	13,572	5,452	0,194	5,620	14,382
p_i	0,0856	0,2595	0,3455	0,2155	0,063
$n_i^{'} = np_i$	4,325	12,975	17,275	10,775	3,15
$\frac{\left(n_i - n_i^{'}\right)^2}{n_i^{'}}$	0,649	0,301	0,803	1,323	1,087

2.4. Перевірка гіпотез про генеральні середні і дисперсії

В прикладних задачах часто виникає необхідність перевірки рівності середніх значень та дисперсій за даними двох або більше вибірок. Наприклад, коли визначається перевага однієї з технологій виготовлення певної продукції, або наявність підвищення продуктивності праці після внесення змін в процес виробництва, або при перевірці якості продукції. Здійснення означеної перевірки виконується за критеріями, що обираються залежно від виду розподілу вибіркових даних і мети дослідження. Для деяких критеріїв перевірки рівності середніх значень висувається додаткова вимога − про рівність генеральних дисперсій.

2.4.1. Перевірка гіпотези про рівність генеральних дисперсій. *F*- критерій (Фішера)

Перевірка гіпотези про рівність генеральних дисперсій здійснюється за *F*- критерієм (Фішера) тільки тоді, коли статистичні дані незалежні і розподілені за нормальним законом. Формулюються гіпотези наступним чином:

H_0 – дисперсії двох нормально розподілених генеральних сукупностей рівні, тобто $S_1^2 = S_2^2$;

H_1 − дисперсії двох нормально розподілених генеральних сукупностей не рівні, тобто $S_1^2 \neq S_2^2$.

F-критерій (Фішера) розраховується за формулою:

$$F = \frac{S_1^2}{S_2^2},\ S_1^2 > S_2^2. \qquad (2.4)$$

Гіпотеза H_0 приймається, якщо розраховане значення *F* менше критичного значення розподілу Фішера $F_{кр}$, взятого із рівнем значущості α і ступенями свободи l_1 та l_2 для чисельника і знаменника відповідно: $l_i = n_i - 1, i = 1,2$, де n_i – об'єми вибірок. $F_{кр}$ можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР (α ; l_1; l_2).

Зауваження. Дисперсія у чисельнику дробу в формулі (2.4) повинна бути більшою дисперсії у знаменнику, тобто значення *F*-критерія повинно бути більше одиниці.

Приклад 2.2. Відомо дані про продуктивність праці (одиниць продукції за зміну) двох груп працівників: група 1 складається з працівників, що пройшли тренінг з вдосконалення виробництва за новітніми технологіями; група 2 – із працівників, що не проходили відповідний тренінг (табл. 2.5). Враховуючи, що дані розподілені за нормальним законом, перевірити гіпотезу про рівність дисперсій.

Таблиця 2.5

	група 1					група 2				
Продуктивність праці	34	85	96	102	103	63	69	83	89	106
Кількість працівників	5	2	11	8	4	2	6	8	3	1

Розв'язання. Дані табл. 2.5 є двома вибірками. Перша – вибірка значень величини X_1 – продуктивність праці робітників,

що пройшли навчання, друга – вибірка величини X_2 – продуктивність праці робітників, що не пройшли тренінг.

Сформулюємо основну гіпотезу H_0 – дисперсії генеральних сукупностей, з яких зроблено вибірки, рівні, $S_1^2 = S_2^2$ та конкуруючу гіпотезу H_1 – дисперсії не рівні, $S_1^2 \neq S_2^2$. Перевіримо справедливість основної гіпотези H_0 за F-критерієм (Фішера).

Визначимо оцінку дисперсії вибірки X_1. Розрахунки подамо у вигляді таблиці.

Таблиця 2.6

x_i	34	85	96	102	103	Σ
n_i	5	2	11	8	4	30
$x_i n_i$	170	170	1056	816	412	2624
$\left(x_i - \overline{x}\right)^2 n_i$	14243,42	12,17	801,00	1689,74	965,14	17761,47

Отримаємо: $\overline{x_1} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{30} 2624 \approx 87,47$;

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i - \overline{x_1}\right)^2 n_i = \frac{1}{29}\cdot 17761,46 \approx 612,46 .$$

Аналогічно знайдемо оцінку дисперсії X_2 (табл. 2.7).

Таблиця 2.7

x_i	63	69	83	89	106	Σ
n_i	2	6	8	3	1	20
$x_i n_i$	126	414	664	267	106	1577
$\left(x_i - \overline{x}\right)^2 n_i$	502,45	582,13	137,78	309,07	737,12	2268,55

Отримаємо: $\overline{x_2} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{20}\cdot 1577 \approx 78,85$;

$$S_2^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(x_i - \overline{x_1}\right)^2 n_i = \frac{1}{19}\cdot 2268{,}55 \approx 119{,}40.$$

Знайдемо значення *F*-критерію за формулою (2.4). Оскільки $S_1^2 > S_2^2$, то $F = \frac{S_1^2}{S_2^2} = \frac{612{,}46}{119{,}40} \approx 5{,}13$. Знайдемо $F_{кр}$, враховуючи, що $l_1 = n_1 - 1 = 30 - 1 = 29$; $l_2 = n_2 - 1 = 20 - 1 = 19$. Рівень значущості оберемо $\alpha = 0{,}05$. Тоді, згідно табличних значень, отримаємо $F_{кр}$= FРАСПОБР (0,05; 29; 19) = 2,077.

Оскільки $F > F_{кр}$, то гіпотезу H_0 відхиляємо і приймаємо гіпотезу H_1. Отже, дисперсії вибірок не є рівними, вибірки здобуті з різних генеральних сукупностей.

Висновок: тренінг з вдосконалення виробництва за новітніми технологіями суттєво впливає на продуктивність праці робітників.

2.4.2. Перевірка гіпотези про рівність генеральних середніх. Критерій Стьюдента

Цей критерій був розроблений Вільямом Госсетом для оцінки якості пива в компанії Гіннес. У зв'язку із зобов'язаннями перед компанією з нерозголошення комерційної таємниці, стаття Госсета вийшла 1908 року у журналі «Біометрика» під псевдонімом "Student" (Студент).

Завдання порівняння середніх значень двох генеральних сукупностей постає тоді, коли практичне значення має саме величина досліджуваної ознаки. Наприклад, коли порівнюються терміни виробництва продукції за двома різними технологіями, або порівняння якості продукції, що вроблена при застосуванні тієї або іншої технології, при порівнянні ефективності нових технологій та інше. У цьому випадку можна використовувати *t*-критерій Стьюдента. Особливість використання *t*-критерію Стьюдента при обробці результатів дослідження полягає у необхідності виконання двох умов: нормального характеру розподілу даних, що вивчаються і рівності генеральних дисперсій груп, що порівнюються. В іншому випадку використання *t*-критерію Стьюдента некоректно. Цей критерій

застосовується для порівняння великих і середніх (обсягом понад 30 значень) нерівновеликих ($n_x \neq n_y$) вибірок, а також для малих (об'ємом менше 30 значень) рівновеликих ($n_x = n_y$) вибірок. Є можливість застосовувати *t*-критерій Стьюдента як у випадках порівняння незалежних вибірок, так і у випадку порівняння залежних вибірок.

Залежні вибірки - це вибірки, що являють собою параметри однієї і тієї ж сукупності до та після впливу деякого фактора. Найчастіше залежні вибірки - це результати вимірів однієї і тієї ж групи об'єктів у різні моменти часу (наприклад, до і після впливу певного фактора або факторів).

Вибірки є незалежними , якщо набір об'єктів дослідження в кожній групі здійснювався незалежно від того, які об'єкти дослідження включені до іншої групи. Допускається, щоб кількість об'єктів не залежних вибірках були різними.

Критерій Стьюдента застосовується для перевірки гіпотез про рівність генеральних середніх, якщо статистичні дані можуть бути розподілені за нормальним законом. Гіпотези формулюються наступним чином:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 – середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Перевірка виконується за даними двох вибірок об’ємом n_1 та n_2. При цьому можливі випадки.

Випадок 1. Генеральні дисперсії рівні ($S_1^2 = S_2^2$). Тоді *t*-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \qquad (2.5)$$

Розраховане значення *t*-критерія порівнюється з критичним значенням $t_{кр}$, де $t_{кр}$ – критичне значення розподілу Стьюдента з

параметрами $\frac{\alpha}{2}$ та ступенем свободи $l = n_1 + n_2 - 2$, яке надається в статистичних таблицях або може бути визначене, використовуючи функцію з Excel СТЬЮДРАСПОБР$\left(\frac{\alpha}{2}; l\right)$.

Випадок 2. Генеральні дисперсії не рівні ($S_1^2 \neq S_2^2$). Тоді *t*-критерій Стьюдента обчислюється за формулою:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \qquad (2.6)$$

Розраховане значення *t*-критерію також порівнюється з критичним значенням $t_{кр}$, але степінь свободи розраховується за формулою:

$$l = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}}. \qquad (2.7)$$

Випадок 3. Вибірки не є незалежними, оскільки на них впливає певний фактор і його вплив невідомий, або вибірки є даними, отриманими до і після проведення певного експерименту. Тоді формується парна вибірка і для кожної пари елементів знаходиться *d* – різниця їх значень. Подальша перевірка здійснюється над вибіркою різниць. Значення *t*-критерію Стьюдента обчислюється за формулою:

$$t = \frac{x_d \cdot \sqrt{n-1}}{S_d}. \qquad (2.8)$$

де x_d – вибіркове середнє для вибірки різниць, S_d – вибіркове середнє квадратичне відхилення для вибірки різниць, n – об'єм вибірки різниць.

Розраховане значення t-критерію порівнюється з критичним значенням розподілу Стьюдента з параметрами $\frac{\alpha}{2}$ та ступенем свободи $l = n - 1$.

У всіх випадках основна гіпотеза H_0 приймається, якщо розраховане значення t-критерію менше критичного значення $t_{кр}$ за абсолютною величиною $|t| < t_{кр}$.

Приклад 2.3. Проводилося дослідження продуктивності виробництва товарів двох типів товарів в залежності від обсягів сировини на підприємстві.

Вимірювався час виробництва товарів до повного використання ресурсу (табл. 2.8). Чи можна вважати, що виробництво двох типів товарі є однаково продуктивно?

Таблиця 2.8

Номер підприємства	1	2	3	4	5
Час виробництва першого типу продукції (дні)	60,2	62,3	61,3	60,7	63,4
Час виробництва другого типу продукції (дні)	59,4	58,3	62,1	63,4	60,8

Розв'язання. За умовами задачі було зроблено дві вибірки: перша вибірка об'єму n_1 = 5 значень величини X_1 – часу виробництва продукції першого типу ; друга вибірка об'єму $n_2 = 5$ значень величини X_2 – часу виробництва продукції другого типу. Потрібно перевірити гіпотезу про рівність генеральних середніх.

Сформулюємо гіпотези:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 – середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Оскільки вибірки не є незалежними, то необхідно сформувати парну вибірку (випадок 3). Потрібно знайти для кожної пари елементів різницю їх значень *d*, обчислити для парної вибірки $\overline{x_d}$ і S_d та розрахувати *t*-критерій Стьюдента за формулою (2.8). Розрахунки подамо у вигляді таблиці (табл. 2.9).

Таблиця 2.9

$\overline{x_{1i}}$	60,2	62,3	61,3	60,7	63,4	Σ
$\overline{x_{2i}}$	59,4	58,3	62,1	63,4	60,8	
$d_i = \overline{x_{1i}} - \overline{x_{2i}}$	0,8	4	-0,8	-2,7	2,6	3.9
$\left(d - \overline{x_d}\right)^2$	0,0004	10,3684	2,4964	12,1104	3,3124	28,288

Знайдемо $\overline{x_d}$ за формулою $\overline{x_d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{3,9}{5} = 0,78$.

Обчислимо S_d^2 наступним чином:

$$S_d^2 = \frac{1}{n}\sum_{i=1}^{n}\left(d_i - \overline{x_d}\right)^2 = \frac{28,288}{5} = 5,66 \ .$$

Отримаємо, що $S_d = \sqrt{5,66} \approx 2,38$.

Обчислимо значення *t*-критерія Стьюдента згідно формули формулою (2.9):

$$t = \frac{\overline{x_d}\sqrt{n-1}}{S_d} = \frac{0,78 \cdot 2}{2,38} \approx 0,66 .$$

Знайдемо критичне значення розподілу Стьюдента $t_{кр}$. Рівень значущості оберемо $\alpha = 0,05$, степенів вільності маємо $l = n - 1 = 4$, тоді $t_{кр} = 3,495$.

Отже маємо, що $|t| < t_{кр}$, тому основна гіпотеза H_0 про рівність генеральних середній має місце (приймається) на рівні значущості $\alpha = 0,05$.

Отримаємо висновок: продуктивність виробництва обох видів продукції можемо вважати однаковою.

2.4.3. Перевірка гіпотези про рівність генеральних середніх. Критерій Уілкоксона

Критерій Уілкоксона використовується для перевірки гіпотези про рівність генеральних середніх за статистичними даними двох вибірок тоді, коли дані не є розподілені за нормальном закон розподілу.

Формулювання гіпотез наступне:

H_0 – середні двох генеральних сукупностей рівні, тобто $\overline{x_1} = \overline{x_2}$;

H_1 – середні двох генеральних сукупностей не рівні, тобто $\overline{x_1} \neq \overline{x_2}$.

Для перевірки основної гіпотези необхідно, щоб об'єм першої вибірки n_1 був наприклад меншим за об'єм другої вибірки n_2. Якщо ця умова не виконується, то потрібно поміняти місцями вибірки, тобто першу вважати другою.

Перевірка гіпотези H_0 виконується наступним чином:

1) Формується об'єднана вибірка, об'ємом $n = n_1 + n_2$.

2) Елементам даної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, наступному найменшому – ранг 2 і так далі. При цьому, якщо деякі елементи вибірки співпадають, то їм присвоюються середні ранги. Для цього додаються ранги, які мали б ці елементи, якщо були б різні та розраховується їх середнє арифметичне значення ; кожному з однакових елементів присвоюється ранг, що дорівнює розрахованому середньому арифметичному значенню.

3) Розраховуються суми рангів елементів вихідних вибірок R_1 та R_2.

4) Обирається величина W: якщо вибірки рівні за об'ємом, то $W = R_1$ або $W = R_2$; якщо вибірки різні за об'ємом, то $W = R_1$, де R_1 – сума рангів меншої за об'ємом вибірки.

5) Розраховується значення W^* критерію Уілкоксона за формулою:

$$W^* = \frac{2W - n_1(n_1 + n_2 + 1) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot (n_1 + n_2 + 1)}}. \qquad (2.10)$$

6) Якщо у вибірці є дані з однаковими рангами , то критерій Уілкоксона W^* розраховується за формулою:

$$W^* = \frac{2W - n_1(n_1 + n_2 + 1) + 1}{\sqrt{\frac{n_1 n_2}{3} \cdot (n_1 + n_2 + 1) - \frac{\sum_{i=1}^{m} t_i(t_i^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}}},. \qquad (2.11)$$

де m – кількість груп однакових рангів, що містять дані обох вибірок, t_i – кількість елементів в i-тій групі.

7) Обираємо рівень значущості α.

8) Обчислюємо найменший рівень значущості $p_{гр}$ згідно формули:

$$p_{гр} = 2 \cdot (1 - \Phi\left(\left|W^*\right|\right), \qquad . \qquad (2.12)$$

де $\Phi(x)$ – значення функції нормального стандартного розподілу. Значення даної функції визначається за табличними данними або використовуючи вбудовану функцію Excel НОРМРАСП.

9) Порівнюємо рівень значущості α та отримане значення $p_{гр}$. Якщо $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається.

Приклад 2.4. При дослідженні попиту на два види товарів отримано статистичні дані реалізації цих видів товарів (табл. 2.18).

Враховуючи, що розподіл даних не відповідає розподілу за нормальним законом, перевірити гіпотезу про рівність середніх.

Таблиця 2.10

Вид товару	Кількість									
A	280	230	112	176	90	175	216	110	205	115
B	200	126	225	210	260	194	156	240	170	232

Розв'язання. Дані табл. 2.10 є двома вибірками. Перша – вибірка значень величини X_1 – кількість реалізації товару А; друга – вибірка величини X_2 – кількість реалізації товару В.

Гіпотези для дослідження сформулюємо наступним чином:

H_0 – середні генеральних сукупностей, з яких зроблено вибірки, рівні $\overline{x_1} = \overline{x_2}$;

H_1 – середні генеральних сукупностей, з яких зроблено вибірки, не рівні $\overline{x_1} \neq \overline{x_2}$.

Перевіримо справедливість гіпотези H_0 за критерієм Уілкоксона.

Утворимо об'єднану вибірку, надамо її елементам ранги і обчислимо їх суму. Результати розрахунків запишемо у вигляді таблиці (табл. 2.11). Для зручності виділимо елементи першої вибірки.

Оскільки вибірки рівні за об'ємом, то $W = R_1$ або $W = R_2$.

Розрахуємо значення W^* критерія Уілкоксона за формулою (2.10), якщо $W = R_1 = 89$, то маємо:

$$W^* = \frac{2 \cdot 89 - 10(10+10+1)}{\sqrt{\frac{10 \cdot 10}{3}(10+10+1)}} \approx -1,48.$$

Таблиця 2.11

Об'єднна вибірка	Впорядкована вибірка	Ранги елементів вибірки	Ранги елементів вибірки А	Ранги елементів вибірки В
280	90	1	1	
230	110	2	2	
112	112	3	3	
176	115	4	4	
90	126	5		5
175	156	6		6
216	170	7		7
110	175	8	8	
205	176	9	9	
115	194	10		10
200	200	11		11
126	205	12	12	
225	210	13		13
210	216	14	14	
260	225	15		15
194	230	16	16	
156	232	17		17
240	240	18		18
170	260	19		19
232	280	20		20
Сума			89	121

Оберемо рівень значущості $\alpha = 0,05$. Обчислимо найменший рівень значущості $p_{гр}$ згідно формули (2.12):

$$p_{гр} = 2 \cdot (1 - \Phi(|W^*|) = 2 \cdot (1 - 0,93) = 0,14.$$

Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості α = 0,05.

Розрахуємо значення W^* критерія Уілкоксона, якщо $W = R_2 = 121$:

$$W^* = \frac{2 \cdot 121 - 10(10+10+1)}{\sqrt{\frac{10 \cdot 10}{3}(10+10+1)}} \approx 1,25 .$$

Розрахуємо найменший рівень значущості $p_{гр}$:

якщо $W^* = 1,25$, то $p_{гр} = 2 \cdot (1 - \Phi(|W^*|) = 2 \cdot (1 - 0,89) = 0,22$.

Отже, оскільки $p_{гр} > \alpha$, то гіпотеза H_0 про рівність середніх приймається на рівні значущості $\alpha = 0,05$.

Питання для самоконтролю

1. Що називається статистичною гіпотезою?
2. Чим відрізняється рівень значущості від рівня довіри?
3. За якими етапами здійснюється перевірка статистичних гіпотез?
4. Що означає прийняття гіпотези?
5. Що називається теоретичним законом розподілу випадкової величини?

Емпіричним законом розподілу?

6. Що називається критерієм згоди?
7. Як перевірити гіпотезу про вид закону розподілу випадкової

величини?

8. Яка мета перевірки гіпотез про рівність середніх значень та дисперсій?
9. Які критерії перевірки гіпотез про рівність генеральних середніх?
10. Які умови застосування критеріїв про рівність генеральних середніх?
11. Які критерії перевірки гіпотез про рівність генеральних дисперсій ви

знаєте?

12. Які умови застосування критеріїв про рівність генеральних середніх?
13. Що називається рангом? Яка процедура ранжування?
14. Які засоби MS Excel призначені для перевірки статистичних гіпотез?

РОЗДІЛ 3
ОСНОВИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

Будь-який досліджуваний об'єкт або явище зазвичай характеризується декількома ознаками, тобто різними властивостями. Ці ознаки взаємозв'язані і впливають одна на одну. Крім того, може існувати зв'язок між ознаками різних об'єктів і явищ. Тому для оборобки та аналізу статистичних даних розроблений апарат для виявлення таких зв'язків і оцінки їх сили (тісноти). Цей математичний апарат називається кореляційним аналізом.

3.1. Кореляційний зв'язок між досліджуваними величинами

В багатьох прикладних задачах необхідно виявити залежність між двома властивостями (ознаками) X і Y одного і того ж економічного об'єкта або між певними ознаками різних об'єктів. Якщо вказані ознаки допускають кількісне вимірювання, і, з погляду економічної теорії, виходячи з економічної характеристики об'єкта, ознака Y залежить від ознаки X. Тоді X можна назвати незалежною змінною або *факторною ознакою*, а Y – залежною змінною або *результативною ознакою*.

Якщо кожному значенню факторної ознаки X відповідає одне і тільки одне значення результативної ознаки Y, то говорять, що між цими ознаками існує *функціональний зв'язок*: $Y = f(X)$.

Якщо кожному значенню факторної ознаки X відповідає безліч значень результативної ознаки Y, то говорять, що між цими ознаками існує *статистичний зв'язок*.

Наприклад, якщо X приймає l значень $X = \{x_1, \ldots, x_l\}$ і кожному її значенню x_i відповідає середнє множини значень Y, тобто:

– значенню x_1 відповідає множина $\{y_{11,}\ y_{12,}, \ldots, y_{1m1}\}$;

– значенню x_2 відповідає множина $\{y_{21,}\ y_{22,}, \ldots, y_{2m2}\}$;

…

– значенню x_l відповідає множина $\{y_{l1,}\ y_{l2,}\ldots,\ y_{lml}\}$,

то між *X* та *Y* існує статистичний зв’язок.

Вивчення статистичного зв'язку вважається досить складним процесом, у якому потрібно аналізувати багатовимірні таблиці даних. Тому, інколи вивчається не статистичний, а кореляційний зв'язок між *X* та *Y*.

Якщо кожному значенню факторної ознаки *X* відповідає певне середнє значення результативної ознаки *Y*, то говорять, що між цими ознаками існує *кореляційний зв’язок*. Тобто кореляційною є функціональна залежність між значеннями *X* і середніми значеннями Y : $\overline{Y} = f(X)$.

Наприклад, якщо X приймає l значень $X = \{x_1,\ldots, x_l\}$ і кожному її значенню x_i відповідає середнє множини значень *Y*, тобто:

– значенню x_1 відповідає $y_{x1} = \dfrac{y_{11} + y_{12} + \cdots + y_{1m1}}{m_1}$;

– значенню x_2 відповідає $y_{x2} = \dfrac{y_{21} + y_{22} + \cdots + y_{2m2}}{m_2}$;

…

– значенню x_l відповідає $y_{xl} = \dfrac{y_{l1} + y_{l2} + \cdots + y_{lml}}{m_l}$,

то між *X* та *Y* існує кореляційний зв’язок.

Основними задачами кореляційного аналізу є:

– вивчення сили зв’язку між двома і більше ознаками досліджуваного об’єкта;

– встановлення факторів, що найбільш суттєво впливають на результативну ознаку;

– виявлення невідомих причинно-наслідкових зв’язків між ознаками об’єкта.

3.2. Групування даних для кореляційного аналізу

Вибіркові дані для вивчення кореляційного зв'язку між ознаками X та Y мають вигляд пар їх значень: $(x_1; y_1), \ldots, (x_n, y_n)$, де x_i – значення величини X, y_i – значення величини Y, n – кількість пар значень, $i = \overline{1,n}$. Якщо кількість пар значень достатньо велика (принаймні n > 20), то для зручності розрахунків дані групуються.

Для групування даних необхідно:

1) Розбити множини значень X та Y на інтервали, використовуючи формулу Стерджеса (форм. 1.2) або іншу зручну. Кількість інтервалів для X та Y може бути різною (позначення: k – кількість інтервалів для X; m – кількість інтервалів для Y).

2) Зобразити дані графічно: побудувати на площині точки з координатами (x_i, y_j). В результаті отримується площина, розбита на прямокутники, в кожному з яких може бути множина точок (рис. 3.1). Вказане графічне зображення вибіркових даних називається *полем кореляції*.

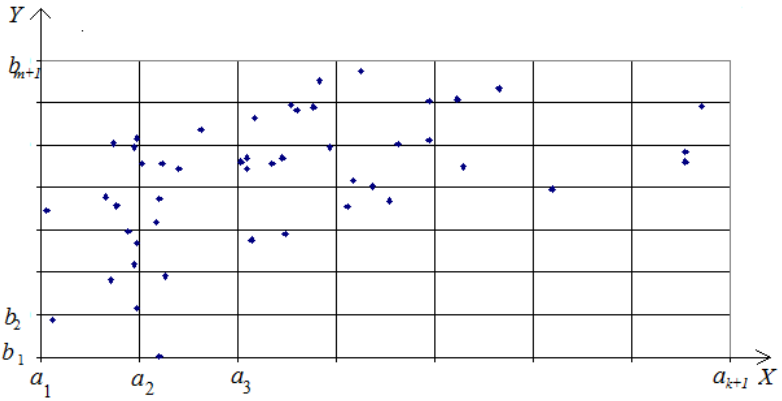


Рис. 3.1. Поле кореляції

3) Побудувати кореляційну таблицю (табл. 3.1). В першому рядку, розбитому на дві частини, записуються інтервали

$[a_i; a_{i+1})$ для X та їх середини x_i. У першому стовпці, розбитому на дві частини, записуються інтервали $[b_j; b_{j+1})$ для Y та їх середини y_j. В центральній частині таблиці записуються частоти n_{ij} – кількість точок, що потратили в прямокутник, обмежений по X інтервалом $[a_i; a_{i+1})$ і по Y – інтервалом $[b_j; b_{j+1})$. В останньому рядку таблиці записуються частоти n_i для X – кількості точок, що потратили в прямокутники, які відповідають інтервалу $[a_i; a_{i+1})$, тобто $n_i = \sum_{j=1}^{m} n_{ij}$ – сума частот n_{ij} в стовпці з номером i. В останньому стовпці таблиці записуються частоти n_j для Y – кількості точок, що потрапили в прямокутники, які відповідають інтервалу $[b_j; b_{j+1})$, тобто $n_j = \sum_{i=1}^{k} n_{ij}$ – сума частот n_{ij} в рядку з номером j.

Кореляційну таблицю можна розглядати як своєрідний подвійний статистичний ряд.

X та x_i		$[a_1; a_2)$	$[a_2; a_3)$	…	$[a_k; a_{k+1})$	$n_j = \sum_{i=1}^{k} n_{ij}$
Y та y_j		x_1	x_2	…	x_k	
$[b_1; b_2)$	y_1	n_{11}	n_{12}	…	n_{1k}	n_1
$[b_2; b_3)$	y_2	n_{21}	n_{22}	…	n_{2k}	n_2
…	…	…	…	…	…	
$[b_m; b_{m+1})$	y_m	n_{m1}	n_{m2}	…	n_{mk}	n_m
$n_i = \sum_{j=1}^{m} n_{ij}$		n_1	n_2	…	n_k	

4) За даними кореляційної таблиці будується ряд, що відображає залежність середнього значення Y від X (табл. 3.2). В першому рядку таблиці записуються середини інтервалів x_i. В

другому – відповідні середні значення $\overline{y_{x_i}}$, що знаходяться за формулами:

$$\overline{y_{x_1}} = \frac{y_1 n_{11} + \cdots + y_m n_{1m}}{n_1}, \ldots, \overline{y_{x_k}} = \frac{y_1 n_{k1} + \cdots + y_m n_{km}}{n_k}.$$

Таблиця 3.2

x_i	x_1	x_2	…	x_k
$\overline{y_{x_i}}$	$\overline{y_{x_1}}$	$\overline{y_{x_2}}$	…	$\overline{y_{x_k}}$
n_i	n_1	n_2	…	n_k

В результаті отримується статистичний ряд, що містить значення *X*, відповідні середні значення *Y* та частоти. За даними такого ряду проводиться кореляційний аналіз.

3.3. Коефіцієнт кореляції Пірсона

Для оцінки тісноти (або сили) зв'язку між *X* та *Y* існує коефіцієнт кореляції. У випадку, коли між *X* та *Y* існує лінійний зв'язок та вибіркові дані розподілені за нормальним законом, використовується *коефіцієнт кореляції Пірсона*, який ще називається параметричним коефіцієнтом кореляції.

Коефіцієнт кореляції Пірсона розраховується за формулою:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}, \qquad (3.1)$$

де $\bar{x}$ – вибіркове середнє величини *X*;

$\bar{y}$ – вибіркове середнє величини *Y*;

$\overline{xy}$ – вибіркове середнє величини *XY*;

S_x – вибіркове середнє квадратичне відхилення величини *X*;

S_y – вибіркове середнє квадратичне відхилення величини *Y*.

Враховуючи формули для знаходження вибіркових середніх і середніх квадратичних відхилень, а саме:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i\,,\quad \bar{y}=\frac{1}{n}\sum_{j=1}^{m}y_j n_j\,,\quad \overline{xy}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m}x_i y_j n_{ij}\,,$$

$$S_x=\sqrt{\frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i-\left(\frac{1}{n}\sum_{i=1}^{k}x_i n_i\right)^2}\,,$$

$$S_y=\sqrt{\frac{1}{n}\sum_{j=1}^{m}y_j^2 n_j-\left(\frac{1}{n}\sum_{j=1}^{m}y_j n_j\right)^2}$$

отримують більш зручну для розрахунків формулу:

$$r=\frac{n\sum_{i=1}^{k}\sum_{j=1}^{m}x_i y_j n_{ij}-\left(\sum_{i=1}^{k}x_i n_i\sum_{j=1}^{m}y_j n_j\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i-\left(\frac{1}{n}\sum_{i=1}^{k}x_i n_i\right)^2}\sqrt{\frac{1}{n}\sum_{j=1}^{m}y_j^2 n_j-\left(\frac{1}{n}\sum_{j=1}^{m}y_j n_j\right)^2}}\,. \quad (3.2)$$

У випадку незгрупованих даних розрахункова формула суттєво спрощується:

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2}}\,. \quad (3.3)$$

Властивості коефіцієнта кореляції Пірсона

1) Коефіцієнт кореляції Пірсона приймає значення на проміжку $[-1;1]$,тобто $-1\le r\le 1$.

2) Якщо $0 \le |r| \le 0,5$, то зв'язок вважається слабким; якщо $0,5 \le |r| \le 0,7$, то зв'язок вважається середнім; $0,7 \le |r| \le 1$, то зв'язок вважається сильним.

3) Якщо $r > 0$, то зв'язок називається додатнім, тобто зі збільшенням значень *X* значення *Y* також збільшуються. Якщо $r < 0$, то зв'язок називається від'ємним, тобто зі збільшенням значень *X* значення *Y* зменшуються.

Зауваження. Зауважемо, що коефіцієнт кореляції Пірсона показує силу лінійного зв'язку. Якщо між *X* та *Y* існує сильний нелінійний зв'язок, то коефіцієнт кореляції Пірсона може дорівнювати нулю.

Оскільки сила зв'язку між *X* та *Y* оцінюється за вибірковими даними, то необхідна перевірка її *статистичної значущості*, тобто оцінка можливості розповсюдити отримані результати на всю генеральну сукупність.

Перевірка статистичної значущості коефіцієнта кореляції Пірсона здійснюється за допомогою так званої *t*-статистики, яка розраховується за формулою:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (3.4)$$

Розраховане значення *t*-статистики порівнюється з критичним значенням $t_{кр}$. $t_{кр}$ – табличне значення розподілу Стьюдента, яке також можна знайти за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР (α; *l*), де α – обраний дослідником рівень значущості, *l* – степінь свободи, $l = n - 2$.

Якщо розраховане значення *t*-статистики більше критичного $|t| > t_{крит.}$, то коефіцієнт кореляції вважається значущим на обраному рівні α.

Приклад 3.1. За наявними даними про річний об'єм виробництва *Y* (тис. од. продукції) та основні фонди *X* (тис. у. од.) для 20 однотипних підприємств (табл. 3.5) оцінити тісноту

зв'язку між X і Y. Визначити можливість розповсюдження результатів розрахунків на всі підприємства такого типу.

Таблиця 3.3

Y \ X	12,5	17,5	22,5	27,5
20,5	1	-	-	-
21,5	-	2	-	-
22,5	-	1	2	-
23,5	-	-	3	3
24,5	-	-	-	8

Розв'язання. Згруповані вибіркові дані (табл. 3.3) запишемо у вигляді кореляційної таблиці (табл. 3.4).

Таблиця 3.4

Y_j \ X_i	12,5	17,5	22,5	27,5	n_j
20,5	1	0	0	0	1
21,5	0	2	0	0	2
22,5	0	1	2	0	3
23,5	0	0	3	3	6
24,5	-	-	-	8	8
n_i	1	3	5	11	

Для розрахунку коефіцієнта кореляції Пірсона скористаємося формулою (3.2). Розрахунки для зручності оформимо у вигляді таблиці (табл. 3.5).

Таблиця 3.5

x_i	n_i	y_j	n_j	$x_i n_i$	$y_j n_j$	x_i^2	$x_i^2 n_i$	y_j^2	$y_j^2 n_j$
12,5	1	20,5	1	12,5	20,5	156,25	156,25	420,25	420,25
17,5	3	21,5	2	52,5	43	306,25	918,75	462,25	924,5
22,5	5	22,5	3	112,5	67,5	506,25	2531,3	506,25	1518,8
27,5	11	23,5	6	302,5	141	756,25	8318,8	552,25	3313,5
		24,5	8		196			600,25	4802
Суми									

				480	468		11925		10979

Окремо розрахуємо $\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij}$:

$$\sum_{i=1}^{4}\sum_{j=1}^{5} x_i y_j n_{ij} = 20,5\cdot 12,5 + 21,5\cdot 17,5\cdot 2 + 22,5\cdot 17,5 +$$
$$+ 22,5\cdot 22,5 + 23,5\cdot 22,5\cdot 3 + 23,5\cdot 27,5\cdot 3 + 24,5\cdot 27,5\cdot 8 =$$
$$= 11330.$$

Підставимо знайдені суми у формулу (2.2) і отримаємо:

$$r = \frac{20\cdot 11330 - 480\cdot 468}{\sqrt{20\cdot 11925 - 480^2}\cdot\sqrt{20\cdot 10979 - 468^2}} = \frac{1960}{90\cdot 23,58} \approx 0,924.$$

За значенням коефіцієнта кореляції можна зробити висновок, що між X і Y існує сильний додатній зв'язок.

Перевіримо статистичну значущість знайденого коефіцієнта кореляції Пірсона. Розрахуємо t-статистику за формулою (2.4):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,924\sqrt{20-2}}{\sqrt{1-0,924^2}} = \frac{3,92}{\sqrt{0,146}} \approx 10,26.$$

Знайдемо $t_{кр}$, враховуючи, що $l = n - 2 = 20 - 2 = 18$. Оберемо рівень значущості $\alpha = 0,01$.

Тоді $t_{кр}$ = СТЬЮДРАСПОБР (0,01; 18) = 2,88.

Оскільки розраховане значення t-статистики більше критичного 10,26 > 2,88, то коефіцієнт кореляції можна вважати значущим на обраному рівні α = 0,01.

Висновок. Між річним об'ємом виробництва та основними фондами на підприємствах, що досліджувалися, існує сильний додатній зв'язок. Висновок дійсний для всіх підприємств заданого типу.

3.4. Коефіцієнт кореляції Спірмена

Для оцінки сили зв'язку між X та Y у випадку, коли між X та Y існує нелінійний зв'язок або вибіркові дані не розподілені за нормальним законом, доцільно використовувати коефіцієнт кореляції Спірмена.

Коефіцієнт кореляції Спірмена розраховується за формулою:

$$r_s(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_x + T_y}{n(n^2-1)}, \qquad (3.5)$$

де n – кількість пар вибіркових даних; d_i – різниці між рангами i-го значення X та відповідного значення Y; T_x, T_y – поправки, пов'язані з однаковими рангами, які розраховуються за формулами:

$$T_x = \frac{\sum_{i=1}^{L_x}\left(T_{xi}^3 - T_{xi}\right)}{12}; \quad T_y = \frac{\sum_{i=1}^{L_y}\left(T_{yi}^3 - T_{yi}\right)}{12}, \qquad (3.6)$$

де L_x, L_y – кількість зв'язок (груп однакових рангів); T_{xi}, T_{yi} – розміри i-тих зв'язок (кількість елементів в них).

Зауваження 3.1. Ранги присвоюються вибірковим даним звичайним способом. Даним об'єднаної вибірки присвоюються ранги (порядкові номери) за правилом: найменшому значенню присвоюється ранг 1, наступному найменшому – ранг 2 і т. д. При цьому, якщо деякі елементи вибірки співпадають, то їм присвоюються середні ранги. Для цього додаються ранги, які мали б ці елементи, якщо були б різні; розраховується їх середнє арифметичне; кожному із однакових елементів присвоюється ранг, що дорівнює розрахованому середньому арифметичному.

Зауваження 3.2. Статистична значущість коефіцієнта кореляції Спірмена перевіряється так, як і коефіцієнта кореляції Пірсона.

Приклад 3.2. Вивчається залежність між сумарним обсягом торгівельного бюджету двох країн X (тис. у.о.) та кількістю підприємств, що задіяні в торгівельних відносинах цих країн Y (сумарна кількість). Відповідні дані подано у табл. 3.6. Оцінити силу зв'язку між досліджуваними факторами за коефіцієнтом кореляції Спірмена. Перевірити його статистичну значущість.

Таблиця 3.6

X	52	37	32	26	53	31	36	32	54	64	47	35	34	28	36
Y	16	12	5	4	17	6	15	7	13	20	10	10	10	5	19

Розв'язання. Дані табл. 3.6 є вибірковими парами значень $(x_i; y_i)$, $i = \overline{1,n}$, n – кількість пар, n = 15. Знайдемо коефіцієнт кореляції Спірмена, необхідні розрахунки оформимо у вигляді таблиці (табл. 3.7), використовуючи позначення d_{xi} – ранг x_i, d_{yi} – ранг y_i.

Таблиця 3.7

x_i	y_i	d_{xi}	d_{yi}	d_i	d_i^2
52	16	12	12	0	0
37	12	10	9	–1	1
32	5	4,5	2,5	–2	4
26	4	1	1	0	0
53	17	13	13	0	0
31	6	3	4	1	1
36	15	8,5	11	2,5	6,25
32	7	4,5	5	0,5	0,25
54	13	14	10	–4	16
64	20	15	15	0	0
47	10	11	7	–4	16
35	10	7	7	0	0
34	10	6	7	1	1
28	5	2	2,5	0,5	0,25
36	19	8,5	14	5,5	30,25

Пояснимо, як заповнюється стовпець 3: знаходимо найменше зі значень x_i (це 26) та присвоюємо йому ранг 1; знаходимо наступне найменше (це 28) і присвоюємо йому ранг 2; наступним найменшим є 31, йому присвоюємо ранг 3; наступними найменшими є два значення 32, якщо б вони були різними, то їм би присвоїли ранги 4 і 5, але оскільки вони однакові, то присвоюємо їм середній ранг 4,5 і т. д.

Знаходимо суму квадратів різниць рангів: $\sum_{i=1}^{15} d_i^2 = 76$.

Знаходимо поправки, що пов'язані з однаковими рангами.

В рядку рангів d_{xi} є дві групи однакових рангів, в першій з них 2 елемента, в другій теж два. Отже, $L_x = 2$, $T_{x1} = 2$, $T_{x2} = 2$.

В рядку рангів d_{yi} є дві групи однакових рангів, в першій з них 2 елемента, в другій − три елемента. Отже, $L_y = 2$, $T_{y1} = 2$, $T_{y2} = 3$.

Підставимо отримані дані в формули (2.6) і знайдемо поправки:

$$T_x = 1,\ T_y = 2,5$$

Обчислимо коефіцієнт кореляції Спірмена за формулою (3.5):

$$r_s(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_x + T_y}{n(n^2-1)} = 1 - \frac{6\cdot 76 + 1 + 2,5}{15\left(15^2-1\right)} \approx 0,86.$$

Згідно значення коефіцієнта кореляції можна зробити висновок, що між *X* та *Y* існує сильний додатній зв'язок.

Перевіримо статистичну значущість знайденого коефіцієнта кореляції. Розрахуємо *t*-статистику за формулою (2.4)

$$t = \frac{0,86\sqrt{15-2}}{\sqrt{1-0,86^2}} \approx 6,17.$$

Знайдемо $t_{кр}$, враховуючи, що $l = n - 2 = 15 - 2 = 13$. Оберемо рівень значущості $\alpha = 0,001$.

Тоді $t_{крит}$ = СТЬЮДРАСПОБР (0,001; 13) = 4,22.

Оскільки розраховане значення *t*-статистики більше критичного 6,17 > 4,22, то коефіцієнт кореляції можна вважати значущим на обраному рівні α = 0,001.

Висновок. Між сумарним обсягом торгівельного бюджету двох країн та кількістю підприємств, що задіяні в торгівельних відносинах цих країн існує сильний додатній зв'язок. Висновок дійсний для всієї генеральної сукупності, з якої було зроблено вибірку.

3.5. Множинний та частинний коефіцієнти кореляції

У випадку, коли досліджуваний об'єкт або явище характеризується більш, ніж двома ознаками $X_1, X_2, \ldots, X_k$ необхідно вивчати множинні залежності. Для оцінки сили зв'язку між певною ознакою X_i та усіма іншими ознаками використовують *множинний коефіцієнт кореляції*, який позначається R_i.

Для розрахунку множинного коефіцієнта кореляції необхідно:

1) Побудувати матрицю парних коефіцієнтів кореляції $r_{ij}, i = \overline{1,k}$ між ознаками X_i та X_j:

$$A = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & r_{kk} \end{pmatrix}. \qquad (3.7)$$

2) Знайти визначник $|A|$ матриці A та алгебраїчне доповнення A_{ii} елемента r_{ii} цієї матриці.

3) Розрахувати множинний коефіцієнт кореляції за формулою:

$$R_i = \sqrt{1 - \frac{A_{ii}}{|A|}}. \qquad (3.8)$$

Перевірка статистичної значущості множинного коефіцієнта кореляції здійснюється за допомогою *t*-статистики, яка розраховується за формулою:

$$t = \frac{R^2 (n-k)}{\left(1-R^2\right)(k-1)}, \qquad (3.9)$$

де *n* – кількість взаємопов'язаних значень ознак X_i, $i = \overline{1,k}$.

Розраховане значення *t*-статистики порівнюється з критичним значенням $F_{кр}$. $F_{кр}$ – табличне значення розподілу Фішера, яке також можна знайти за допомогою вбудованої статистичної функції Excel FРАСПОБР $(\alpha; l_1; l_2)$, де α – обраний дослідником рівень значущості; l_1, l_2 – степені свободи: $l_1 = k-1$, $l_2 = n-k$.

Якщо розраховане значення *t*-статистики більше критичного $|t| > F_{кр}$, то множинний коефіцієнт кореляції вважається значущим на обраному рівні значущості α.

У випадку, коли необхідно дослідити кореляційний зв'язок між ознаками X_i та X_j, $i = \overline{1,k}$, $j = \overline{1,k}$ із множини ознак $X_1, X_2, \ldots, X_k$ досліджуваного об'єкта або явища, який не залежить від впливу інших ознак, розраховується *частинний коефіцієнт кореляції*, який позначається R_{ij}.

Для розрахунку частинного коефіцієнта кореляції необхідно:

1) Побудувати матрицю парних коефіцієнтів кореляції *A*.

2) Знайти алгебраїчні доповнення A_{ij} елементів r_{ij} відповідно.

3) Розрахувати частинний коефіцієнт кореляції за формулою:

$$R_{ij} = \frac{-A_{ij}}{\sqrt{A_{ii} A_{jj}}}. \qquad (3.10)$$

Перевірка статистичної значущості частинного коефіцієнта кореляції здійснюється за допомогою *t*-статистики, яка розраховується за формулою:

$$t = \frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}^2}}, \qquad (3.11)$$

де n – кількість взаємопов'язаних значень ознак X_i, $i = \overline{1,k}$.

Розраховане значення *t*-статистики порівнюється з критичним значенням $t_{кр}$. $t_{кр}$ – табличне значення розподілу Стьюдента, яке також можна знайти за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР (α; l), де α – обраний дослідником рівень значущості, l – степінь свободи, $l = n - k + 2$.

Якщо розраховане значення *t*-статистики більше критичного $|t| > t_{кр}$, то частинний коефіцієнт кореляції вважається значущим на обраному рівні значущості α.

Зауваження 3.3 Вважається, що для коректного використання множинного і частинного коефіцієнтів кореляції необхідно, щоб вибіркові дані мали сумісний нормальний розподіл, однак перевірка цієї умови на практиці зазвичай не виконується, оскільки пов'язана зі значними труднощами у розрахунках.

Зауваження 3.4 Замість парного коефіцієнта кореляції Пірсона можна використовувати також парний коефіцієнт кореляції Спірмена.

Зауваження 3.5 Кореляційна матриця завжди симетрична відносно головної діагоналі, оскільки $r_{ij} = r_{ji}$, $i = \overline{1,k}$, $j = \overline{1,k}$. Елементи головної діагоналі завжди дорівнюють 1, оскільки вони є коефіцієнтами кореляції X_i та X_j.

3.6. Кореляційний аналіз із використанням Microsoft Excel

Вбудовані сервісні функції Microsoft Excel дозволяють розраховувати парні коефіцієнти кореляції Пірсона. Для отримання матриці парних коефіцієнтів кореляції необхідно:

1) Вибрати *Сервис – Анализ данных*.

2) У діалоговому вікні для вибору інструмента аналізу вибрати інструмент *Корреляция*. З'явиться вікно для задання параметрів (рис. 3.2).

3) Задати параметри для розрахунку коефіцієнтів кореляції. У графі *Входной интервал* вказати масив даних; у графі *Группирование* вказати тип групування, наприклад *По столбцам*, у графі *Выходной интервал* вказати ту чарунку, починаючи з якої будуть представлятись вихідні дані – парні коефіцієнти кореляції. Натиснути *ОК*.

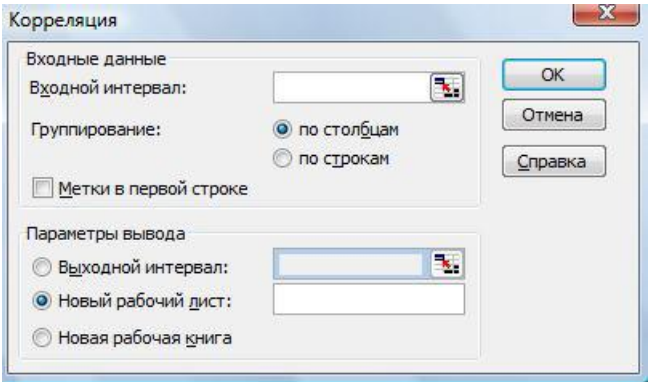
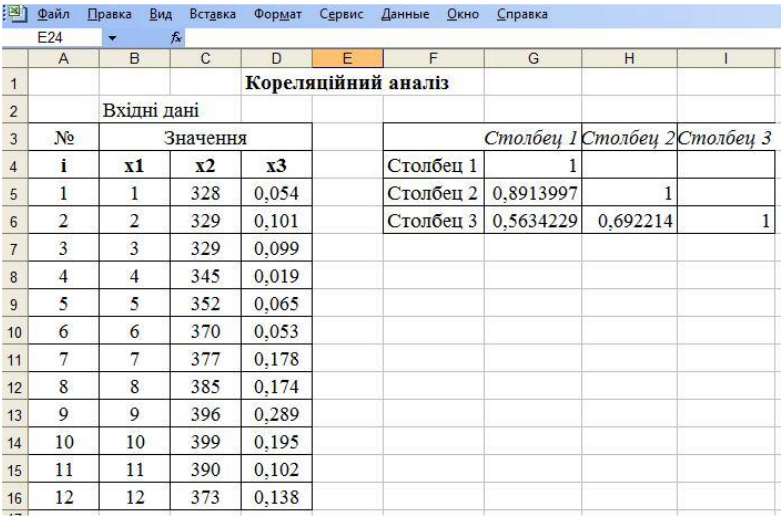


Рис. 3.2. Вікно надання параметрів кореляційного аналізу.

Приклад і результати розрахунків парних коефіцієнтів кореляції продемонстровано на рис. 3.3.



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

E24

	A	B	C	D	E	F	G	H	I
1				**Кореляційний аналіз**					
2		Вхідні дані							
3	№	Значення					*Столбец 1*	*Столбец 2*	*Столбец 3*
4	**i**	**x1**	**x2**	**x3**		Столбец 1	1		
5	1	1	328	0,054		Столбец 2	0,8913997	1	
6	2	2	329	0,101		Столбец 3	0,5634229	0,692214	1
7	3	3	329	0,099					
8	4	4	345	0,019					
9	5	5	352	0,065					
10	6	6	370	0,053					
11	7	7	377	0,178					
12	8	8	385	0,174					
13	9	9	396	0,289					
14	10	10	399	0,195					
15	11	11	390	0,102					
16	12	12	373	0,138					

Рис. 3.3. Результати розрахунку коефіцієнтів кореляції.

Зауваження.

1) В результаті роботи інструмента аналізу даних *Корреляция* розраховується матриця парних коефіцієнтів кореляції Пірсона навіть у випадку встановлення зв'язку між двома величинами.

2) Клітини матриці, що розташовані вище головної діагоналі, зазвичай залишаються незаповненими, оскільки матриця симетрична відносно головної діагоналі.

3) Засобами Microsoft Excel неможливо розрахувати парні або множинні коефіцієнти кореляції, однак можна значно спростити розрахунки, використовуючи вбудовану математичну функцію МОПРЕД, яка дозволяє обчислити визначник заданої матриці.

Приклад і результати обчислення визначника матриці показано на рис. 3.4.

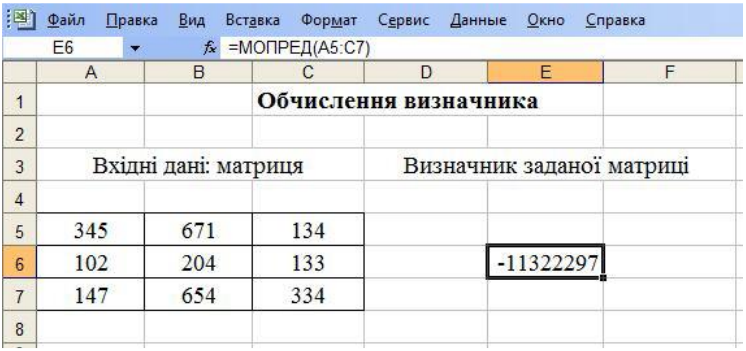


Рис. 3.4. Обчислення визначника заданої матриці.

Питання для самоконтролю

1. Що називається кореляційним аналізом? Яка мета кореляційного аналізу?
2. Що називається кореляційним зв’язком? Статистичним зв’язком?
3. Що називається коефіцієнтом кореляції? Як він використовується у статистичному моделюванні?
4. Як згрупувати вхідні дані при кореляційному аналізі?
5. Що називається полем кореляції? Кореляційною таблицею?
6. Як побудувати кореляційну таблицю?
7. Як за вибірковими даними визначити вид зв’язку?
8. Чим відрізняються коефіцієнти кореляції Пірсона та Спірмена? Які загальні риси мають коефіцієнти кореляції Пірсона і Спірмена?
9. Як мають задаватись вхідні дані для кореляційного аналізу у випадку лінійного зв'язку? У випадку нелінійного зв’язку?
10. Який зв’язок між факторами вважається сильним? Середнім? Слабким?
11. Чи показує коефіцієнт кореляції спрямованість зв’язку?
12. Який висновок робиться, якщо коефіцієнт кореляції є додатнім? Від’ємним?

13. Як присвоюються ранги, якщо вибіркові дані повторюються?

14. Що означає поняття «статистична значущість»?

15. Як перевірити зв’язок між декількома факторами?

16. Які коефіцієнти кореляції обчислюються при множинному кореляційному аналізі?

17. Що таке множинний кореляційний зв’язок?

18. Що таке чистий кореляційний зв’язок?

19. Для чого служить частинний коефіцієнт кореляції? Множинний коефіцієнт кореляції?

20. Що називається кореляційною матрицею? Для чого будується кореляційна матриця?

21. Які властивості кореляційної матриці?

22. Для чого перевіряється статистична значущість коефіцієнта кореляції?

23. Як побудувати кореляційну матрицю засобами MS Excel?

24. Чи можливо знайти множинний та частинний коефіцієнти кореляції засобами Microsoft Excel?

РОЗДІЛ 4

ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ

При вивченні тісноти зв'язку між різними ознаками економічного чи соціального об'єкта головною задачею є встановлення виду кореляційної залежності результативної ознаки (Y) від факторної (X), тобто виду функціональної залежності $\overline{Y} = f(X)$. В першу чергу це пов'язано з необхідністю прогнозування досліджуваних процесів. Математико-статистичний апарат, що дозволяє встановити вид кореляційної залежності називається *регресійним аналізом*, а функція, яка описує цю залежність, називається *рівнянням регресії*.

Регресійний аналіз проводиться за такими етапами:

1. Встановлення виду кореляційної залежності результативної ознаки Y від факторної ознаки X.
2. Побудова регресійної моделі.
3. Перевірка статистичної значущості побудованої моделі.

Перший етап регресійного аналізу є найважливішим, оскільки помилки у виборі виду залежності призводять до побудови регресійної моделі, що не відповідає емпіричним даним і не може використовуватися для прогнозування.

Вибіркові дані для вивчення кореляційного зв'язку між ознаками X та Y, зазвичай, мають вигляд пар їх значень: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, x_i – значення величини X, y_i – значення Y, n – кількість пар значень, $i = \overline{1,n}$.

Якщо їх кількість достатньо велика, то для зручності розрахунків дані групуються (див. п. 3.2) і будується статистичний ряд, що містить значення X, відповідні середні значення Y та частоти (табл. 4.1).

Таблиця 4.1

x_i	x_1	x_2	…	x_k
$\overline{y}_{x_i}$	$\overline{y}_{x_1}$	$\overline{y}_{x_2}$	…	$\overline{y}_{x_k}$
n_i	n_1	n_2		n_k

Згруповані дані (табл. 4.1) зображуються графічно, що часто дозволяє визначити вид залежності Y від X.

Ламана лінія, що сполучає точки з координатами $\left(x_i, \overline{y}_{x_i}\right)$, називається *емпіричною лінією регресії*.

Якщо емпірична лінія регресії значно наближається до прямої лінії, то висувається гіпотеза про наявність лінійного зв'язку між досліджуваними ознаками (рис. 4.1).

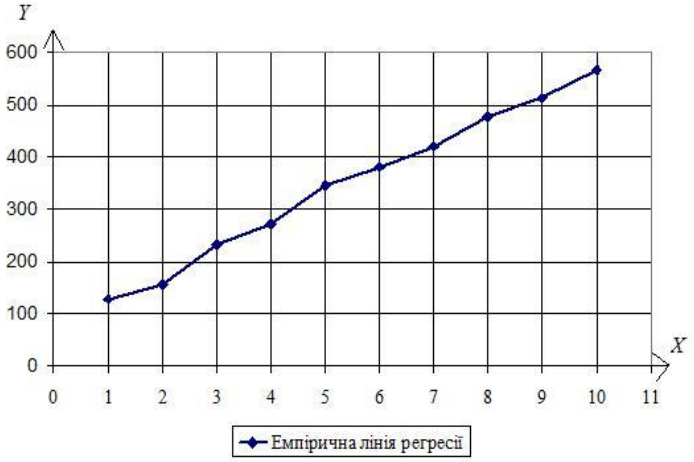


Рис. 4.1. Лінійний зв'язок між досліджуваними ознаками

В іншому випадку висувається гіпотеза про наявність нелінійного зв'язку (рис. 4.2).

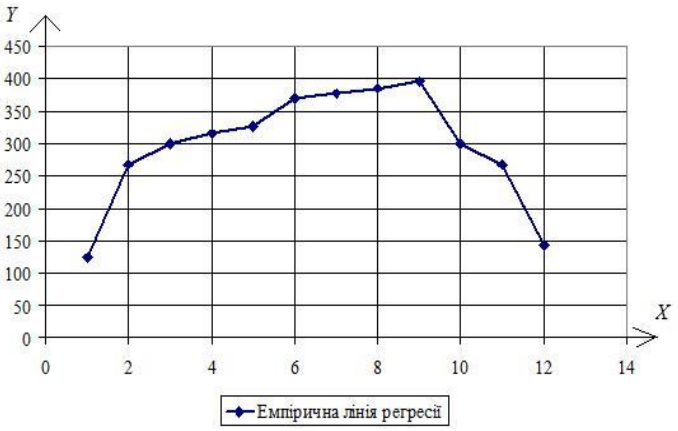


Рис. 4.2. Нелінійний зв'язок між досліджуваними ознаками

4.1 Побудова лінійної регресійної моделі

Якщо висунуто гіпотезу про наявність лінійної залежності результативної ознаки (Y) від факторної ознаки (X), то рівняння регресії має вигляд:

$$\overline{y}_x = ax + b, \qquad (4.1)$$

де a, b – параметри моделі.

Побудова лінійної регресійної моделі полягає в знаходженні параметрів рівняння (4.1). Параметри рівняння регресії знаходимо за *методом найменших квадратів*.

Метод найменших квадратів

Нехай при вивчення залежності Y від X було отримано вибіркові дані: $x_1, x_2, \ldots, x_n$ – значення величини X, $y_1, y_2, \ldots, y_n$ – відповідні значення Y. За вибірковими даними було побудовано рівняння регресії $y = ax + b$. Якщо в рівняння підставити замість x значення $x_1, x_2, \ldots, x_n$, то будуть отримані теоретичні значення Y: $y_{1,теор}, y_{2,теор}, \ldots, y_{n,теор}$, які

відрізняються від $y_1, y_2, \dots, y_n$. Різниця значень $y_{i,теор} - y_i$ називається похибкою регресійної моделі і позначається e_i. Якщо параметри рівняння підбирано так, щоб сума квадратів похибок була мінімальною, то говорять, що вони отримані методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії методом найменших квадратів знаходимо з системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum\limits_{i=1}^{k} x_i^2 n_i + b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \overline{y}_{x_i}, \\ a\sum\limits_{i=1}^{k} x_i n_i + b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \overline{y}_{x_i}. \end{cases} \tag{4.2}$$

Якщо вибіркові дані не згруповані, то система (4.2) значно спрощується:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i, \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i. \end{cases} \tag{4.3}$$

Перевірка адекватності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o, \tag{4.4}$$

де $Q = \sum\limits_{i=1}^{k} \left(\overline{y}_{x_i} - \overline{y} \right)^2 n_i$ – загальна варіація, тобто сума квадратів відхилень емпіричних значень Y від середнього $\overline{y} = \dfrac{\sum\limits_{i=1}^{k} \overline{y}_{x_i} n_i}{n}$;

$Q_p = \sum\limits_{i=1}^{k} \left(y_{i,теор} - \overline{y} \right)^2 n_i$ – варіація регресії, тобто сума квадратів

відхилень теоретичних значень Y від середнього значення, що обумовлена регресією; $Q_o = \sum_{i=1}^{k} \left(y_{i,теор} - \overline{y}_{x_i} \right)^2 n_i$ – варіація залишків, тобто сума квадратів відхилень теоретичних значень Y від емпіричних значень.

У випадку незгрупованих даних загальна варіація, варіація регресії і залишків знаходяться за формулами: $Q = \sum_{i=1}^{n} \left(\overline{y}_i - \overline{y} \right)^2$; $Q_p = \sum_{i=1}^{n} \left(y_{i,теор} - \overline{y} \right)^2$; $Q_o = \sum_{i=1}^{n} \left(y_{i,теор} - y_i \right)^2$; а середнє значення за формулою $\overline{y} = \frac{\sum_{i=1}^{n} \overline{y}_i}{n}$.

Для перевірки статистичної значущості рівняння регресії розраховується *F*-статистика за формулою:

$$F = \frac{Q_p \left(n - l \right)}{Q_o \left(l - 1 \right)}, \tag{4.5}$$

де n – кількість спостережень, l – кількість груп у кореляційній таблиці або кількість параметрів моделі у випадку незгрупованих даних. Розраховане значення *F*-статистики порівнюється з критичним значенням $F_{кр}$ розподілу Фішера, яке можна знайти за статистичними таблицями або за допомогою вбудованої функції в Excel $ФРАСПОБР\left(\alpha, k_1, k_2 \right)$, де $k_1 = l - 1$; $k_2 = n - l$ – степені свободи, α – рівень значущості.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2, що показує частину варіації значень результативної ознаки Y, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}. \tag{4.6}$$

Значення коефіцієнта детермінації знаходяться в інтервалі $[0;1]$, тобто $0 \le R^2 \le 1$. Чим ближче R^2 до 1, тим краще отримане рівняння регресії пояснює поведінку результативної ознаки. Наприклад, якщо $R^2 = 0,96$, то 96% варіації результативної ознаки Y пояснюється рівнянням регресії.

Приклад 4.1. Побудувати регресійну модель, що описує залежність виробничих затрат Y (тис. грн.) від об'ємів виробництва X (тис. од.). Відповідні статистичні дані подано у табл. 4.2.

Таблиця 4.2

X	41	44	52	57	59	64	68	70	73	75
Y	670	657	713	736	778	812	833	876	911	932

Розв'язання. В табл. 4.2 задано вибіркові дані: значення $x_i, i = \overline{1,n}$ величини X та відповідні значення $y_i, i = \overline{1,n}$; кількість пар $n = 10$ невелика, тому для проведення регресійного аналізу їх можна не групувати.

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.3).

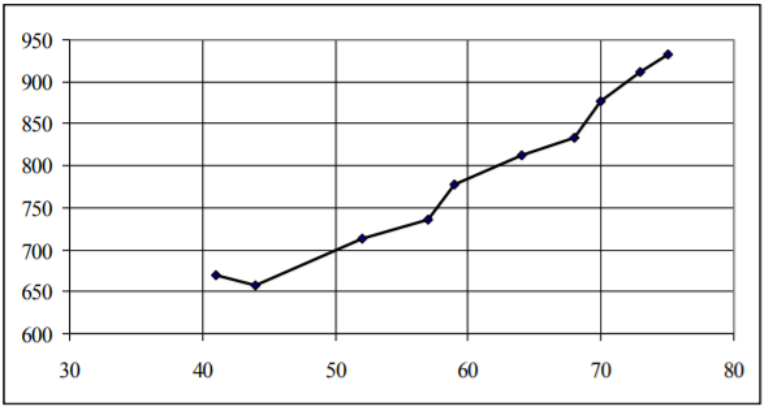


Рис. 4.3. Емпірична лінія регресії

Оскільки емпірична лінія регресії наближається до прямої лінії, то висуваємо гіпотезу про лінійну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $y = ax + b$.

Другий етап: знайдемо параметри a, b рівняння регресії, для чого складемо систему (4.3) для не згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.3).

Таблиця 4.3

	x_i	y_i	x_i^2	$x_i y_i$
Розрахункові дані	41	670	1681	27470
	44	657	1936	28908
	52	713	2704	37076
	57	736	3249	41952
	59	778	3481	45902
	64	812	4096	51968
	68	833	4624	56644
	70	876	4900	61320
	73	911	5329	66503
	75	932	5625	69900
Суми	603	7918	37625	487643

Отже, складемо систему для знаходження параметрів рівняння регресії та розв'яжемо її за правилом Крамера:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \end{cases} \Rightarrow \begin{cases} 37625a + 603b = 487643 \\ 603a + 10b = 7918 \end{cases}$$

Знайдемо визначник основної матриці системи, яка складена із коефіцієнтів перед невідомими:

$$\Delta = \begin{vmatrix} 37625 & 603 \\ 603 & 10 \end{vmatrix} = 37625 \times 10 - 603^2 = 12641.$$

Знайдемо допоміжні визначники, що отримуються із попереднього заміною відповідного стовпця коефіцієнтів на стовпець вільних членів:

$$\Delta_a = \begin{vmatrix} 487643 & 603 \\ 7918 & 10 \end{vmatrix} = 487643 \times 10 - 603 \times 7918 = 101876;$$

$$\Delta_b = \begin{vmatrix} 37625 & 487643 \\ 603 & 7918 \end{vmatrix} = 37625 \times 7918 - 603 \times 487643 = 38866021.$$

Знайдемо невідомі за формулами Крамера:

$$a = \frac{\Delta_a}{\Delta} = \frac{101876}{12641} \approx 8,06; \quad b = \frac{\Delta_b}{\Delta} = \frac{3866021}{12641} \approx 305,83.$$

Отже, знайдене рівняння регресії має вигляд

$$y = 8,06x + 305,83.$$

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для чого знайдемо загальну варіацію, варіацію регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.4).

Передусім знайдемо $\bar{y}$: $\bar{y} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} \approx 791,8$.

Таблиця 4.4

x_i	y_i	$y_{i,теор}$	$(\bar{y}_i - \bar{y})^2$	$(\bar{y}_{i,теор} - \bar{y})^2$	$(\bar{y}_{i,теор} - y_i)^2$
41	670	636,26	14835,24	24193,32	1138,52
44	657	660,44	18171,04	17256,64	11,80
52	713	724,91	6209,44	4474,42	141,82
57	736	765,20	3113,64	707,31	852,92
59	778	781,32	190,44	109,77	11,04
64	812	821,62	408,04	889,17	92,52

68	833	853,86	1697,44	3850,90	434,96
70	876	869,97	7089,64	6111,17	36,31
73	911	894,15	14208,64	10475,83	283,87
75	932	910,27	19656,04	14035,10	472,20
		Сума	85579,6	82103,63	3475,974

Отже, $Q = 85579,6$; $Q_p = 82103,63$; $Q_o = 3475,974$; тоді основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має вигляд: $85579,6 = 82103,63 + 3475,974$ і є тотожністю, тому рівняння регресії побудовано вірно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 10$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p\left(n-l\right)}{Q_o\left(l-1\right)} = \frac{82103,63\left(10-2\right)}{3475,974\left(2-1\right)} \approx 188,96.$$

Знайдемо $F_{кр}$: $F_{кр} = ФРАСПОБР\left(0,001;2-1,10-2\right) \approx 25,41$.

Розраховане значення *F*-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2:

$$R^2 = \frac{Q_p}{Q} = \frac{82103,63}{85579,6} \approx 0,96.$$

Значення коефіцієнта детермінації свідчить, що 96% варіації результативної ознаки Y пояснюється рівнянням регресії.

Висновок: Сумарні виробничі затрати Y (тис. грн.) лінійно залежать від об'єму виробництва X (тис. од.). Залежність описується рівнянням: $y = 8,06x + 305,83$, яке є статистично значущим на рівні значущості 0,001 та описує 96% вибіркових даних.

4.2 Побудова нелінійної регресійної моделі

Якщо висунуто гіпотезу про наявність нелінійної залежності результативної ознаки (Y) від факторної (X), то регресійний аналіз проводиться за тими ж етапами, як і у випадку лінійної залежності. Вид рівнянь регресії і системи для знаходження їх параметрів для нелінійних залежностей, що найчастіше зустрічаються на практиці , подано у табл. 4.5.

Таблиця 4.5

<table>
<tr><td colspan="2">Рівняння параболічної регресії:
$\overline{y_x} = ax^2 + bx + c$.</td></tr>
<tr><td colspan="2">Система для знаходження параметрів:</td></tr>
<tr><td>для згрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{k} x_i^4 n_i + b\sum_{i=1}^{k} x_i^3 n_i + c\sum_{i=1}^{k} x_i^2 n_i = \sum_{i=1}^{k} x_i^2 n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^3 n_i + b\sum_{i=1}^{k} x_i^2 n_i + c\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum_{i=1}^{k} x_i^2 n_i + b\sum_{i=1}^{k} x_i n_i + c\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}$</td><td>для незгрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$</td></tr>
<tr><td colspan="2">Рівняння гіперболічної регресії:
$\overline{y_x} = \frac{a}{x} + b$.</td></tr>
<tr><td colspan="2">Система для знаходження параметрів:</td></tr>
<tr><td>для згрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{k} \frac{1}{x_i^2} n_i + b\sum_{i=1}^{k} \frac{1}{x_i} n_i = \sum_{i=1}^{k} \frac{1}{x_i} \overline{y_{x_i}} n_i \\ a\sum_{i=1}^{k} \frac{1}{x_i} n_i + b\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \overline{y_{x_i}} n_i \end{cases}$</td><td>для незгрупованих вибіркових даних:
$\begin{cases} a\sum_{i=1}^{n} \frac{1}{x_i^2} + b\sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{n} \frac{1}{x_i} y_i \\ a\sum_{i=1}^{n} \frac{1}{x_i} + bn = \sum_{i=1}^{n} y_i \end{cases}$</td></tr>
<tr><td colspan="2">Рівняння показникової регресії:
$\overline{y_x} = ba^x$.</td></tr>
</table>

Система для знаходження параметрів:	
для згрупованих вибіркових даних:	для незгрупованих вибіркових даних:
$\begin{cases} \lg a\sum\limits_{i=1}^{k} x_i^2 n_i + \lg b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \lg \overline{y}_{x_i} \\ \lg a\sum\limits_{i=1}^{k} x_i n_i + \lg b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \lg \overline{y}_{x_i} \end{cases}$	$\begin{cases} \lg a\sum\limits_{i=1}^{n} x_i^2 + \lg b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i \lg y_i \\ \lg a\sum\limits_{i=1}^{n} x_i + n\lg b = \sum\limits_{i=1}^{n} \lg y_i \end{cases}$

Перевірка статистичної значущості нелінійної регресійної моделі також здійснюється за F-статистикою. При цьому для параболічної регресії кількість параметрів $l = 3$, для гіперболічної і показникової кількість параметрів $l = 2$.

Приклад 4.2. Дано розподіл однотипних підприємств за об'ємом виробництва X (тис. од.) і собівартістю одиниці продукції Y (грн.) (табл. 4.6). Знайти регресійну модель, що описує залежність собівартості продукції від об'єму виробництва.

Таблиця 4.6

X \ Y	10	15	20	25
25	–	–	1	2
50	–	2	2	–
75	–	5	3	1
100	1	3	–	–
125	3	1	1	–

Розв'язування. Для застуосування регресійного аналізу за даними табл. 4.6 побудуємо кореляційну таблицю (табл. 4.7).

Таблиця 4.7

y_j \ x_i	25	50	75	100	125	n_j
10	0	0	0	1	3	4
15	0	2	5	3	1	11

20	1	2	3	0	1	7
25	2	0	1	0	0	3
n_i	3	4	9	4	5	$n=25$

За даними кореляційної таблиці побудуємо ряд, що відображає залежність середнього значення Y від X, попередньо обчисливши середні значення $\overline{y}_{x_i}$ для кожного значення $x_i, i=\overline{1,5}$. Отримані результати подамо у вигляді таблиці 4.8:

$$\overline{y}_{x_1}=\frac{y_1n_{11}+y_2n_{12}+y_3n_{13}+y_4n_{14}}{n_1}=\frac{20\times1+25\times2}{3}\approx23,33;$$

$$\overline{y}_{x_2}=\frac{15\times2+20\times2}{4}=17,5;\ \ \overline{y}_{x_3}=\frac{15\times5+20\times3+25\times1}{9}=17,78;$$

$$\overline{y}_{x_4}=\frac{10\times1+15\times3}{4}=13,75;\ \overline{y}_{x_5}=\frac{10\times3+15\times1+20\times1}{5}=13.$$

Таблиця 4.8

x_i	25	50	75	100	125
$\overline{y}_{x_i}$	23,33	17,5	17,78	13,75	13
n_i	3	4	9	4	5

Перший етап аналізу: визначимо вид залежності Y від X. Побудуємо емпіричну лінію регресії (рис. 4.4).

Оскільки емпірична лінія регресії наближається до гіперболи, то висуваємо гіпотезу про гіперболічну залежність Y від X, тобто рівняння регресії будемо шукати у вигляді $\overline{y}_x=\frac{a}{x}+b$.

Другий етап: знайдемо параметри a,b рівняння регресії, для чого складемо систему для згрупованих даних. Необхідні розрахунки для зручності оформимо у вигляді таблиці (табл. 4.9), в останньому рядку якої знайдемо відповідні стовпцям суми.

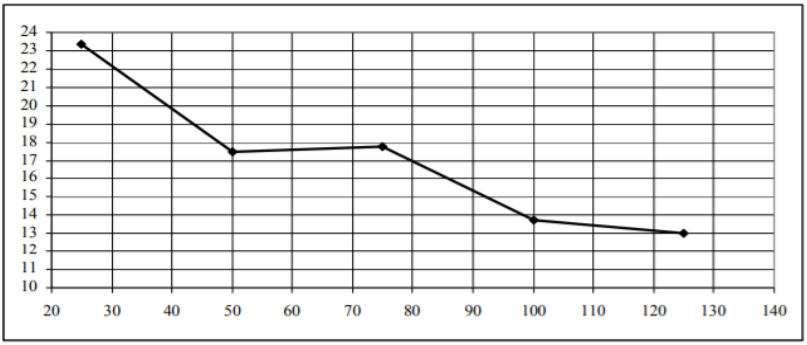


Рис. 4.4. Емпірична лінія регресії

Таблиця 4.9

x_i	$\overline{y}_{x_i}$	n_i	$\frac{1}{x_i}$	$\frac{1}{x_i}n_i$	$\frac{1}{x_i^2}n_i$	$\overline{y}_{x_i}n_i$	$\frac{1}{x_i}\overline{y}_{x_i}n_i$
25	23,33	3	0,04	0,12	0,0048	69,99	2,7996
50	17,5	4	0,02	0,08	0,0016	70	1,4
75	17,78	9	0,0133	0,12	0,0016	160,02	2,1336
100	13,75	4	0,01	0,04	0,0004	55	0,55
125	13	5	0,008	0,04	0,0003	65	0,52
			Суми	0,4	0,0087	420,01	7,4032

Отже, складемо систему для знаходження параметрів рівняння регресії та розв'яжемо її за правилом Крамера:

$$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y}_{x_i}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i} + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y}_{x_i}n_i \end{cases} \Rightarrow \begin{cases} 0,0087a + 0,4b = 7,4032 \\ 0,4a + 25b = 420,01 \end{cases}.$$

Основний визначник системи:

$$\Delta = \begin{vmatrix} 0,0087 & 0,4 \\ 0,4 & 25 \end{vmatrix} = 0,0087 \times 25 - 0,4^2 = 0,058 .$$

Допоміжні визначники:

$$\Delta_a = \begin{vmatrix} 7,4032 & 0,4 \\ 420 & 25 \end{vmatrix} = 7,4032 \times 25 - 420 \times 0,4 = 17,076;$$

$$\Delta_a = \begin{vmatrix} 0,0087 & 7,4032 \\ 0,4 & 420 \end{vmatrix} = 0,0087 \times 420 - 0,4 \times 7,4032 = 0,701207.$$

Тоді отримаємо:

$$a = \frac{\Delta_a}{\Delta} = \frac{17,076}{0,058} \approx 294,41;\ b = \frac{\Delta_b}{\Delta} = \frac{0,7012207}{0,058} \approx 12,09.$$

Отже, знайдене рівняння регресії має вигляд

$$\overline{y}_x = \frac{291,41}{x} + 12,09.$$

Третій етап: перевіримо правильність побудови моделі за рівнянням (4.4), її статистичну значущість за *F*-статистикою (4.5) і адекватність вибірковим даним за коефіцієнтом детермінації (4.6). Для цього знайдемо загальну варіацію, варіації регресії та залишків; необхідні розрахунки оформимо у вигляді таблиці (табл. 4.10).

Знайдемо $\overline{y}$: $\overline{y} = \dfrac{\sum_{i=1}^{n} \overline{y}_{x_i} n_i}{n} \approx 16,8.$

Таблиця 4.10

x_i	y_i	n_i	$y_{i,\text{теор}}$	$\left(y_i - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - \overline{y}\right)^2 n_i$	$\left(y_{i,\text{теор}} - y_i\right)^2 n_i$
25	23,33	3	23,866	127,907	149,782	0,863
50	17,5	4	17,978	1,958	5,547	0,914
75	17,78	9	16,015	8,637	5,547	28,028
100	13,75	4	15,034	37,220	12,482	6,594
125	13	5	14,445	72,215	27,737	10,441
			Суми	247,936	201,096	46,840

Отже, $Q = 247,936$; $Q_p = 201,096$; $Q_o = 46,840$. Основне варіаційне рівняння $Q = Q_p + Q_o$ для побудованої моделі має

вигляд: $247,936 = 201,096 + 46,840$ і є тотожністю, тому рівняння регресії побудовано вірно.

Для перевірки статистичної значущості рівняння регресії знайдемо *F*-статистику, враховуючи, що $n = 25$, $l = 2$ – оскільки шукали рівняння з двома параметрами:

$$F = \frac{Q_p\left(n-l\right)}{Q_o\left(l-1\right)} = \frac{201,096\left(25-2\right)}{46,840\left(2-1\right)} \approx 98,75.$$

Знайдемо $F_{кр}$: $F_{кр} = ФРАСПОБР\left(0,001; 2-1, 25-2\right) \approx 14,20$. Розраховане значення *F*-статистики більше критичного, тому регресійна модель є статистично значущою на рівні 0,001.

Знайдемо коефіцієнт детермінації R^2:

$$R^2 = \frac{Q_p}{Q} = \frac{201,096}{247,936} \approx 0,81.$$

Значення коефіцієнта детермінації свідчить, що 81% варіації результативної ознаки Y пояснюється рівнянням регресії.

Висновок: Залежність собівартості одиниці продукції Y (грн.) від об'єму виробництва X (тис. од.) описується рівнянням $\overline{y}_x = \frac{291,41}{x} + 12,09$, яке є статистично значущим на рівні значущості 0,001 та описує 81% вибіркових даних.

4.3 Побудова множинної лінійної регресійної моделі

У процесі аналізу діяльності економічного або соціального об'єкта часто виявляється, що на результативну ознаку цієї діяльності (наприклад, об'єм валової продукції, об'єм продаж, думку респондента відносно певного об'єкта та ін.) впливає декілька факторних ознак: час, вартість сировини і матеріалів, якість обладнання, продуктивність праці, соціальні установки, вплив зовнішніх і внутрішніх факторів та інше. Тоді як модель діяльності об'єкта використовують багатофакторну лінійну регресійну модель, на основі якої розробляються прогнози

діяльності, вивчається вплив на діяльність різноманітних показників і виявляються ті показники, покращення яких суттєво збільшує її кінцевий продукт.

Загальний вигляд багатофакторної лінійної регресійної моделі:

$$Y = f\left(X_1, X_2, ..., X_m\right) + \varepsilon, \qquad (4.7)$$

де Y – результативна ознака, $X_1, X_2, ..., X_m$ – факторні ознаки, m – кількість факторних ознак, ε – випадкова похибка моделі.

Зауваження 1. Задачі побудови багатофакторної регресійної моделі розв'язуються за умов, коли випадкова похибка ε має нормальний розподіл із нульовим математичним сподіванням, а випадкові похибки кожного вимірювання незалежні та мають однакові дисперсії. Кількість спостережень n повинна перевищувати величину $3(m+1)$.

Крім того, для забезпечення статистичної значущості моделі необхідно дотримуватися основного правила її побудови: *„Факторні ознаки, які включено у модель, повинні бути тісно пов'язані із результативною ознакою і слабо пов'язані (або не мати зв'язку) між собою"*.

Тіснота зв'язку між результативною і факторними ознаками та зв'язку факторних ознак між собою визначається за аналізом парних і частинних коефіцієнтів кореляції. В модель бажано включати тільки ті ознаки, що не мають статистично значущого зв'язку між собою, хоча й вважається, що сильний зв'язок між ними, зазвичай, не впливає на якість прогнозу за моделлю.

Якість моделі визначається за критерієм Фішера, тобто порівнянням статистики F моделі із критичним значенням $F_{кр}$, де $F_{кр}\left(\alpha, k_1, k_2\right)$ – табличне значення розподілу Фішера, що знаходиться за умов: $\alpha = 0,05$; $k_1 = m-1$; $k_2 = n-m$. Якщо $F > F_{кр}$, то модель є достовірною на рівні значущості $0,05$ (тобто 95% даних пояснюються побудованою моделлю, 5% – випадкові помилки моделі).

Відносну величину впливу факторних ознак на результативну можна оцінити за формулою:

$$r_{X_i}^2 = \frac{t_{X_i}^2 \times R^2}{\sum_{i=1}^{m} t_{X_i}^2} \ ; \qquad (4.8)$$

де $t_{X_i}^2$ розраховане значення розподілу Стьюдента для ознаки X_i, R^2 – загальний коефіцієнт детермінації моделі.

Обчислення, які необхідно провести для побудови багатофакторної регресійної моделі є певним чином досить об'ємними, тому досить часто застосовують засобу Excel *Регрессия* пакета *Анализ данных* значно полегшує цю роботу.

За допомогою засобу *Регрессия* отримують такі результати:

– параметри лінійної регресійної моделі виду $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_m X_m$, де $b_0, b_1, b_2, ..., b_m$ – параметри моделі.

– коефіцієнт детермінації;

– критеріальну статистику для перевірки статистичної значущості моделі;

– теоретичні значення результативної ознаки, отримані за побудованою моделлю та залишки – тобто різниці між теоретичними та емпіричними значеннями цієї ознаки.

Зауваження 2. Засобом *Регрессия* можна також користуватися при побудові моделі, нелінійної відносно певної факторної ознаки X_j, але лінійної відносно коефіцієнта цієї ознаки. Наприклад, якщо необхідно побудувати модель виду $Y = b_0 + b_1 X_1 + b_2 X_2^3$, то як вхідні дані вказують трійки $\left(y_j, X_{1j}, X_{2j}^3\right)$.

4.4. Лінійна регресія у Microsoft Excel

Пакет аналізу даних Microsoft Excel надає можливість будувати регресійні моделі, але тільки у випадку лінійної

залежності результативної ознаки *Y* від факторної ознаки *X* і тільки для незгрупованих вибіркових даних.

Для побудови лінійної регресійної моделі необхідно:

1) Активізувати *Сервис – Анализ данных – Регрессия – ОК*. З'явиться вікно для надання вхідних даних (рис. 4.5).

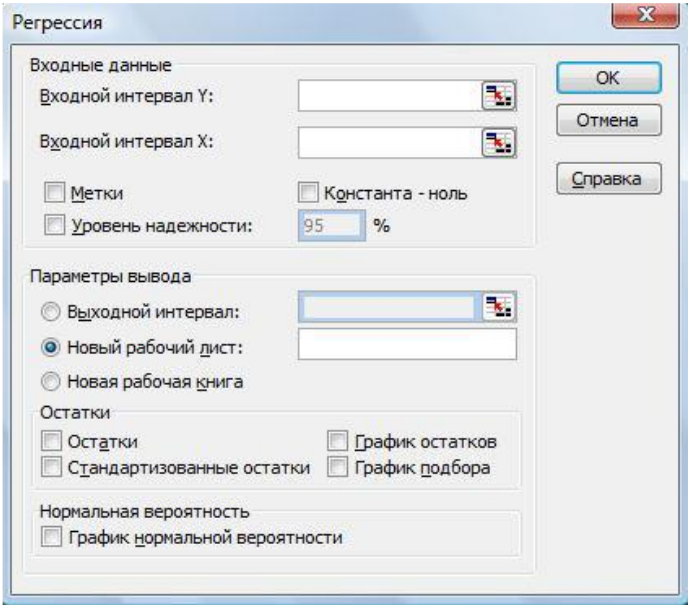
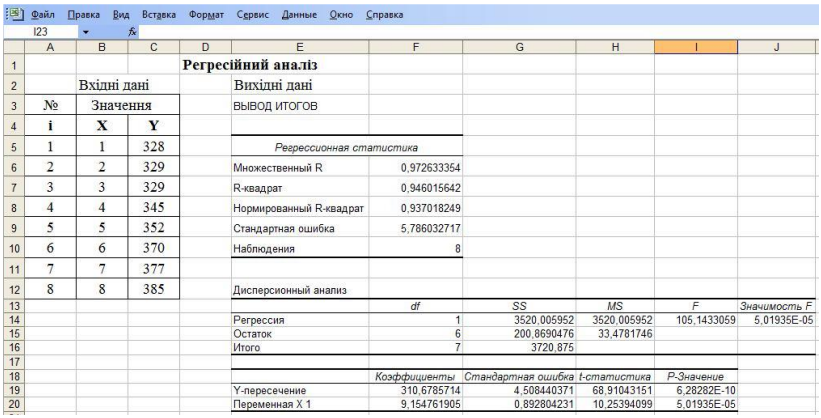


Рис. 4.5. Діалогове вікно функції *Регрессия*

2) У графі *Входной интервал Y* та *Входной интервал X* вказати відповідні стовпці даних; у графі *Выходной интервал* вказати ту комірку, починаючи з якої будуть надаватися вихідні дані – параметри рівняння регресії та результати її статистичного аналізу.

Приклад і результати роботи функції *Регрессия* представлено на рис. 4.6.



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

I23

	A	B	C	D	E	F	G	H	I	J
1					**Регресійний аналіз**					
2		Вхідні дані			Вихідні дані					
3	№	Значення			ВЫВОД ИТОГОВ					
4	**i**	**X**	**Y**							
5	1	1	328		*Регрессионная статистика*					
6	2	2	329		Множественный R	0,972633354				
7	3	3	329		R-квадрат	0,946015642				
8	4	4	345		Нормированный R-квадрат	0,937018249				
9	5	5	352		Стандартная ошибка	5,786032717				
10	6	6	370		Наблюдения	8				
11	7	7	377							
12	8	8	385		Дисперсионный анализ					
13						*df*	*SS*	*MS*	*F*	*Значимость F*
14					Регрессия	1	3520,005952	3520,005952	105,1433059	5,01935E-05
15					Остаток	6	200,8690476	33,4781746		
16					Итого	7	3720,875			
17										
18						*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	
19					Y-пересечение	310,6785714	4,508440371	68,91043151	6,28282E-10	
20					Переменная X 1	9,154761905	0,892804231	10,25394099	5,01935E-05	

Рис. 4.6. Результати регресійного аналізу

В таблиці (рис. 4.6) у графі *Коэффициенты* вказані значення параметрів моделі a та b: b − в графі *Y-пересечение*, a − в графі *Переменная X1*.

Отже, побудована лінійна регресійна модель має вигляд:

$$y = 69{,}15x + 310{,}68\,.$$

Для перевірки статистичної значущості моделі надається значення F-статистики у графі F: $F = 105{,}14$. Це значення обчислюється як відношення варіації регресії до варіації залишків (комірки H14 та H15). В стовпці *Значимость F* надано критеріальну статистику. Якщо це значення менше, ніж, наприклад, 0,05, то рівняння регресії є значущим на рівні 0,05. У даному завданні рівняння регресії є значущим на рівні 0,00005.

У графі *Множественный R-квадрат* надано значення множинного коефіцієнта кореляції, який показує силу залежності результативної ознаки від факторної (або декілька факторних ознак). У нашому випадку він дорівнює 0,97, що означає сильний зв'язок між Y та X.

Коефіцієнт детермінації моделі R_2 виводиться у графі *R-квадрат*, $R_2 = 0{,}97$, тобто 97% даних описується рівнянням регресії. Крім того, можна задати: графік підбору – порівняльна діаграма, що містить емпіричну і теоретичну лінії регресії;

таблиця залишків – різниць емпіричних і теоретичних значень Y (рис. 4.7).

	A	B	C	D	E	F	G
1					**Регресійний аналіз**		
2		Вхідні дані			Вихідні дані		
3	№	Значення			ВЫВОД ОСТАТКА		
4	**i**	**X**	**Y**				
5	1	1	328		*Наблюдение*	*Предсказанное Y*	*Остатки*
6	2	2	329		1	319,8333333	8,166666667
7	3	3	329		2	328,9880952	0,011904762
8	4	4	345		3	338,1428571	-9,142857143
9	5	5	352		4	347,297619	-2,297619048
10	6	6	370		5	356,452381	-4,452380952
11	7	7	377		6	365,6071429	4,392857143
12	8	8	385		7	374,7619048	2,238095238
13					8	383,9166667	1,083333333

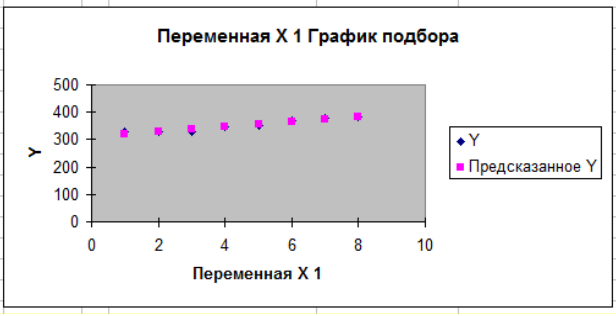


Рис. 4.7. Додаткові результати регресійного аналізу

Питання для самоконтролю

1. Що називається регресійним аналізом?
2. Що називається регресійною моделлю?
3. Що називається факторними ознаками? Результативною ознакою?
4. Які види регресійних моделей Ви знаєте?
5. Як повинні бути задані вхідні дані для регресійного аналізу?

6. Як сформулювати гіпотезу про вид регресійної моделі?

7. Що таке теоретичні та емпіричні значення результативної ознаки?

8. Що називається емпіричною лінією регресії? Теоретичною лінією регресії?

9. Як побудувати емпіричну лінію регресії? Теоретичну лінію регресії?

10. Як знайти параметри регресійної моделі?

11. Як будується розрахункова таблиця у регресійному аналізі?

12. Як перевірити правильність побудованої регресійної моделі?

13. Як перевірити адекватність побудованої регресійної моделі вхідним даним?

14. Що називається коефіцієнтом детермінації і як він використовується у статистичному моделюванні?

15. Як перевірити статистичну значущість побудованої регресійної моделі?

16. Що називається багатофакторною лінійною регресією?

17. Які етапи побудови багатофакторної лінійної регресійної моделі?

18. Як обґрунтовується вибір факторних ознак для побудови моделі?

19. Що називається кореляційними плеядами?

20. Як оцінюється вплив факторних ознак на результативну?

21. Як здійснюється прогноз за багатофакторною лінійною регресійною моделлю?

22. Як визначити вид нелінійної регресійної моделі?

23. Як перевірити статистичну значущість нелінійної регресійної моделі?

24. Як знайти критичне значення критерія Фішера?

25. Як здійснюється прогноз за нелінійною регресійною моделлю?

26. Для чого перевіряється статистична значущість регресійної моделі?

27. Чому при наявному нелінійному зв’язку будують декілька регресійних моделей?

28. Як з декількох регресійних моделей обрати найбільш адекватну?

29. Як побудувати порівняльну діаграму у регресійному аналізі?

30. Як побудувати регресійну модель засобами MS Excel?

РОЗДІЛ 5

ПРОБЛЕМНІ ПИТАННЯ ПРИКЛАДНИХ ДОСЛІДЖЕНЬ

5.1. Основні методи формування вибірки

Формування вибірки – це базовий етап будь-якого прикладного дослідження При формуванні вибірки досліднику необхідно визначити:

− що є елементом або *одиницею вибірки* виходячи від сутності дослідження;

− *контур вибірки*, тобто список усіх одиниць генеральної сукупності, з якої формується вибірка;

− *об'єм вибірки* − кількість елементів у ній.

Наприклад, якщо фірма – виробник мобільних телефонів бажає вивчити потенціальний ринок своєї продукції, то одиницями вибірки будуть особи, які приймають рішення про вибір комунікаційного обладнання: керівники організацій або голови родин. Контуром вибірки можуть бути списки: організацій, фірм.

Оскільки вибірка є лише частиною генеральної сукупності, то отримані на основі її вивчення результати не будуть точно відповідати результатам, які можна було б отримати при вивченні всієї генеральної сукупності. Різниця між результатами дослідження вибірки та генеральної сукупності називається *помилкою вибірки*. Помилки вибірки обумовлюються як методами її формування, так і її об'ємом.

5.1.1. Ймовірнісні методи формування вибірки

При формуванні вибірки часто використовують ймовірнісні методи, при яких всі елементи генеральної сукупності мають однакову ймовірність бути включеними в вибірку. Такими методами є простий випадковий відбір, систематичний відбір, кластерний відбір та стратифікований відбір.

При *простому випадковому відборі* передбачається, що для всіх елементів генеральної сукупності ймовірність бути обраним

в вибірку відома і однакова. Така ймовірність визначається як відношення об'єму вибірки до об'єму генеральної сукупності. Простий випадковий відбір може здійснюватись за допомогою таблиць випадкових чисел або генератора випадкових чисел MS Excel. При цьому: кожному елементу генеральної сукупності присвоюється порядковий номер; генерується необхідна кількість випадкових чисел; обираються ті елементи генеральної сукупності, номери яких співпадають з випадковими числами. Так, наприклад, при проведенні телефонного інтерв'ю комп'ютер випадково генерує телефонні номери.

Недоліком цього методу є необхідність визначення кожного елемента генеральної сукупності, що часто просто неможливо.

При *систематичному відборі* необхідно: визначити всі елементи генеральної сукупності; визначити інтервал відбору – відношення об'єму генеральної сукупності до об'єму вибірки; відібрати в вибірку елементи генеральної сукупності з урахуванням інтервалу.

Наприклад, якщо як генеральна сукупність використовується телефонний довідник, в якому 5500 номерів, а необхідний об'єм вибірки дорівнює 500, то інтервал відбору визначається як $5500/500=11$. Це означає, що кожний одинадцятий телефонний номер потрапляє у вибірку.

Кластерний відбір базується на розподілі елементів генеральної сукупності на групи (кластери), кожна з яких репрезентативна всій сукупності. У подальшому випадково обирається один із кластерів, елементи якого вважаються генеральною сукупністю. На ньому проводиться певне дослідження, результати якого розповсюджуються на всі інші кластери і на всю генеральну сукупність.

Наприклад, якщо планується дослідження думки населення певної області, то кластерами можуть вважатися райони цієї області. При цьому дослідник припускає, що населення кожного району ідентичне населенню області у цілому (що не завжди правильно).

У випадку, коли дослідження проводиться на певній території, можна розбити цю територію на кластери за

допомогою вибіркової решітки, яка накладається на карту території і визначає кластери.

Недоліком усіх описаних методів є необхідність припущення про симетричний розподіл характеристик генеральної сукупності, що практично зустрічається дуже рідко.

Якщо розподіл характеристик генеральної сукупності не є симетричним, використовується *стратифікований відбір*. В цьому випадку генеральна сукупність розподіляється на групи (страти) залежно від певної характеристики.

Наприклад, населення певного регіону розподіляється на групи залежно від рівня доходу. Кожна страта виконує роль генеральної сукупності, з якої формується вибірка. При цьому необхідно враховувати пропорційність стратифікування.

Якщо об'єм вибірки для певної страти пропорційний розміру страти у відношенні до всієї генеральної сукупності, то вибірка називається пропорційно стратифікованою. Якщо ця умова не виконується, то слід застосовувати вагові коефіцієнти, які урівноважать розміри страт.

5.1.2. Неймовірнісні методи формування вибірки

Використання ймовірнісних методів формування вибірки є найбільш правильним з точки зору статистики, але вони трудомісткі і, відповідно, матеріально невигідні. Тому в прикладних дослідженнях часто використовують неймовірнісні методи: відбір за критерієм, відбір згідно спостереження, відбір за квотами та відбір за принципом зручності.

Метод *відбору за принципом зручності* полягає в тому, що формування вибірки здійснюється самим зручним з боку дослідника способом за мінімум часу та грошових витрат. Наприклад, опитування покупців проводиться в певному магазині. При цьому вибірка часто не є репрезентативною, але існують способи оцінки її похибки.

При *формуванні вибірки за критеріями* виникає питання щодо її складу. Так часто формуються фокус-групи. Метод *відбору за спостереженням* використовується тоді, коли контури вибірки дуже обмежені. Наприклад, при проведенні

дослідження попиту на продукцію, що не має широкого кола споживачів.

За цим методом дослідник формує початкову вибірку, об'єм якої значно менший необхідного. У процесі дослідження вибірка розширюється за рахунок пропозицій і рекомендацій респондентів, які вже прийняли участь в опитуванні.

Метод *відбору за квотами* використовується за умови наявності чітко визначених характеристик респондентів, думку яких доцільно вивчити при дослідженні. У цьому випадку в вибірку відбирають ті елементи генеральної сукупності, що мають певні характеристики. Об'єм вибірки (квота) також визначається на початку дослідження. Відбір елементів закінчується при заповненні квоти.

Зауваження. При формуванні вибірки дослідник повинен знайти баланс між затратами на збір даних і об'ємом вибірки. Методи формування вибірки повинні відповідати цілям дослідження.

5.2. Визначення об'єму вибірки

В практичних дослідженнях використовується кілька методів розрахунку об'єму вибірки. Але незалежно від результатів розрахунків слід пам'ятати, що:

− об'єм вибірки залежить передусім від вартості дослідження;

− точність отриманих результатів залежить не стільки від об'єму вибірки, скільки від методу її формування.

Об'єм вибірки може бути просто процентом від об'єму генеральної сукупності. Наприклад, припускається, що для отримання досить точних результатів достатньо 5% елементів генеральної сукупності. Однак перевірити таке припущення неможливо, тому рівень точності результатів невідомий.

Якщо певне дослідження проводиться регулярно, то доцільно використовувати вибірки однакового об'єму. Крім того, якщо відомо, що дослідження певного типу вже проводилося, можна обрати той же об'єм вибірки. Але при

цьому не враховуються можливі зміни умов проведення дослідження.

При проведенні соціологічних опитувань вважається, що для пробних(пілотних) опитувань достатня вибірка об'ємом 100 – 250 осіб. При масових опитуваннях (генеральна сукупність становить 5000 осіб і більше) об'єм вибіркової сукупності повинен становити 10% генеральної сукупності, але не більше 2 – 2,5 тис. осіб. Це гарантує достатньо достовірні результати дослідження. Помилки вибірки, які інколи при цьому трапляються, бувають наслідком невірних вихідних статистичних даних про параметри контрольних ознак генеральної сукупності; недостатнього об'єму вибіркової сукупності, неправильного застосування способу відбору одиниць аналізу (наприклад, відбір із неправильно складеного списку, невдалий вибір місця, часу проведення опитування тощо).

Найбільш коректним є статистичний метод розрахунку об'єму вибірки, заснований на визначенні мінімально необхідного об'єму вибірки залежно від вимог до точності результатів дослідження.

У статистичному методі використовуються такі позначення:

n – об'єм вибірки;

N – об'єм генеральної сукупності;

z − нормативне відхилення оцінки. Обирається залежно від рівня значущості (табл. 5.1). Зазвичай потрібен рівень значущості $\alpha = 0{,}05$, тоді $z = 1{,}96$;

S^2 – вибіркова дисперсія ;

S – вибіркове середнє квадратичне відхилення;

p – варіація вибірки.

e – допустима помилка, яка обирається дослідником.

Таблиця 5.1

α	0,4	0,3	0,2	0,15	0,1	0,05	0,03	0,01	0,001
z	0,84	1,03	1,29	1,44	1,65	1,96	2,18	2,58	3,0

Зауваження. Для визначення варіації певної генеральної сукупності доцільно провести попереднє (пілотне) дослідження.

Крім того, потрібно враховувати, що максимальною є варіація $p = 50$. Тобто така варіація відповідає найгіршому випадку.

Наведемо розрахункові формули для визначення об'єму вибірки у випадку, коли вибірка становить менше 5% від генеральної сукупності і вважається великою:

$$n = \frac{z^2 pq}{e^2}. \tag{5.1}$$

Значення виразу (5.1) використовується, якщо потрібно розрахувати кількість елементів вибірки з генеральної сукупності за умов, що елементи можуть бути представлені у двох типах груп в залежності від певних ознак.

$$n = \frac{z^2 S^2}{e^2}. \tag{5.2}$$

Значення виразу (5.2) використовується, якщо з попередніх досліджень відома дисперсія або середнє квадратичне відхилення.

Якщо такі дослідження не проводилися і вибірка формується вперше, то використовується формула, в якій помилка e пов'язана із середнім квадратичним відхиленням:

$$n = \frac{z^2}{e_1^2}, \text{ де } e_1^2 = \frac{e}{S}. \tag{5.3}$$

Якщо об'єм вибірки більше 5% генеральної сукупності, то у вирази (5.1-5.3) вводиться уточнюючий коефіцієнт $k = \sqrt{\frac{N-n}{N-1}}$, тоді об'єм вибірки обчислюємо за формулою:

$$n_k = n \cdot k = n \cdot \sqrt{\frac{N-n}{N-1}}.$$

Приклад 5.1. Необхідно визначити кількість респондентів для опитування при дослідженні ринку продовольчих послуг.

Розв'язання. Якщо кількість споживачів невідома, приймаємо варіацію вибірки $p = 50$. Рівень значущості $\alpha = 0{,}05$, тоді $z = 1{,}96$, припустима помилка 4%. Тоді згідно формули (5.1) маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1{,}96^2 \cdot 50 \cdot 50}{4^2} = 600 \text{ осіб.}$$

Якщо потрібна менша помилка, наприклад 3%, то за тією ж формулою маємо:

$$n = \frac{z^2 pq}{e^2} = \frac{1{,}96^2 \cdot 50 \cdot 50}{3^2} = 1067 \text{ осіб.}$$

Якщо проводилося попереднє дослідження і відомо, що 70% респондентів є споживачами, то об'єм вибірки дорівнює:

$$n = \frac{z^2 pq}{e^2} = \frac{1{,}96^2 \cdot 70 \cdot 30}{3^2} = 504 \text{ особи.}$$

5.3. Обробка результатів експертного оцінювання

5.3.1. Коефіцієнт конкордації

Дуже важливим етапом у підведенні результатів дослідження є прогнозування. Часто прогнозування передбачає визначення значень економічних або соціологічних показників у майбутньому. Наприклад, прогнозування думки респондентів щодо рекламованого товару, ціни на товари, об'єм їх продаж.

Прогнозування є початковим етапом планування і включає в себе попередній і кінцевий (формальний) прогнози, для яких розробляється один або декілька сценаріїв майбутніх подій.

Методи експертних оцінок використовуються для прогнозування майбутніх подій, якщо відсутні статистичні дані або їх недостатньо. Вони також застосовуються для кількісного вимірювання таких подій, для яких не існує інших способів вимірювання.

Припускається, що експерт формує своє судження на аналізі групи факторів, оцінюючи ймовірності їх реалізації та впливу на результативну ознаку об'єкта вивчення. Але, при цьому отримані висновки та оцінки пов'язані з особистістю експерта, тому інший експерт, використовуючи ту саму інформацію, може дійти інших висновків. Тому вважається, що при розв'язанні проблем в умовах невизначеності, думка групи експертів дає більш надійні результати, ніж думка одного експерта.

Після отримання експертних оцінок проводиться їх обробка та оцінюється достовірність. Обробку результатів експертного оцінювання можна проводити за *коефіцієнтом конкордації*, який показує степінь згоди думок експертів. Найбільш достовірні оцінки отримуються за умов узгодженості думок експертів. Коефіцієнт конкордації W розраховується за формулою:

$$W = \frac{12}{m^2(n^3 - n)} \sum_{j=1}^{n} \sum_{i=1}^{m} (R_{ij} - \frac{n+1}{2})^2 , \qquad (5.4)$$

де n – кількість об'єктів оцінювання, m – кількість експертів, R_{ij} – ранг j-го об'єкта, представленого i-м експертом.

Статистична значущість коефіцієнта конкордації перевіряється порівнянням величини $n(m-1)W$ з табличним значенням розподілу χ^2 при рівні значущості $\alpha = 0{,}001$ з $n - 1$ степеням свободи.

Якщо коефіцієнт конкордації виявляється не значущим, то використовується методика виведення експерта, думка якого не узгоджується з думкою інших експертів. Для цього будується матриця коефіцієнтів кореляції Пірсона $(r(k,i))$ або рангових коефіцієнтів кореляції Спірмена $(r_s(k,i))$ та виявляється експерт, оцінка якого задовольняє умову:

$$r_j(k,i) = \min\{r(k,i)\},$$

що означає, що думка цього експерта найменше узгоджується з думкою інших експертів. Бали, подані таким експертом, у подальших розрахунках не враховуються.

Алгоритм повторюється, поки коефіцієнт конкордації не стає значущим.

5.3.2. Коефіцієнт компетенції

Використання коефіцієнта конкордації засновано на припущенні, що чим більш узгоджені думки експертів, тим достовірнішими є їх оцінки. Але практика показує, що це не завжди вірно, і експерт, який не згоден з думками більшості, може дати найточніші оцінки.

Якщо дослідник бажає врахувати думки всіх експертів, то обробку результатів експертного оцінювання слід виконувати за *коефіцієнтом компетентності* експерта.

Цей метод базується на використанні попередньої оцінки компетентності експертів, які приймають участь у дослідженні. Оцінка експертів проводиться за критеріями компетентності, серед яких можуть бути: рівень освіти; загальний стаж роботи; стаж роботи за проблемою дослідження; посада тощо.

Крім того, важливим критерієм є оцінка надійності експерта, яка розраховується як відношення його правильних оцінок до всіх проведених експертиз. Правильними вважаються ті оцінки, які з часом підтвердилися практикою.

При розрахунку коефіцієнтів компетентності експертів необхідно використовувати єдину для всіх критеріїв шкалу оцінювання. У противному випадку оцінки потрібно буде нормалізувати, тобто привести до однієї шкали.

Коефіцієнт компетентності розраховується за формулою:

$$KK_i = \frac{\sum_{j=1}^{m} k_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}},$$

де *n* – кількість експертів; *m* – кількість критеріїв оцінювання експертів; *kij* – бал, отриманий *i*-м експертом за *j*-м критерієм.

5.4. Елементи факторного аналізу

Факторний аналіз − сукупність моделей і методів, призначених для впорядкування інформації, яка міститься в кореляційній матриці. Він допомагає виявити неявні фактори, які пояснюють взаємозв'язки між спостережуваними ознаками досліджуваного об'єкта. Кількість ознак може бути великою і зв'язки між ними надзвичайно складними, однак, спостерігаючи за об'єктом, ми можемо виявити невелику кількість факторів, які впливають на досліджувані ознаки. Факторний аналіз передбачає класифікацію ознак, які мають подібний характер зміни при переході від одного об'єкта спостереження до іншого.

Обґрунтована заміна великої кількості ознак, описаних об'єктами спостережень, меншим числом комплексних характеристик (факторів) складають зміст факторного аналізу. Кожний *фактор* − це група взаємопов'язаних ознак, які визначають змістовну інтерпретацію даного фактора. При цьому в один фактор об'єднуються ознаки, які тісно корелюють між собою. Ознаки з різних факторів характеризуються слабким кореляційним зв'язком.

Основними етапами факторного аналізу є:

1) збір емпіричних даних і підготовка кореляційної (коваріаційної) матриці;

2) виділення початкових факторів і побудова факторної структури (обчислення факторних навантажень): проводиться вибір методу обчислення, визначається кількість факторів на основі змістовних або математичних міркувань;

3) обертання факторної структури: вибір критерію обертання;

4) змістовна інтерпретація результатів факторного аналізу;

5) обчислення факторних значень.

Припустимо, що досліджується деякий масив з *n* емпіричних ознак. За допомогою методів кореляційного аналізу можна встановити залежності між ними, обчисливши коефіцієнти кореляції. Тоді вся множина з *n* емпіричних ознак розіб'ється на

окремі групи за величиною коефіцієнтів кореляції. Наприклад, перша ознака тісно пов'язана з четвертою і шостою, а шоста з першою, перша з п’ятою і т. д., причому з іншими ознаками зв'язок виявляється значно слабкішим. Тоді ці взаємопов'язані ознаки утворюють загальну функціональну одиницю, яку ми і називаємо фактором. Наприклад, при аналізі успішності студентів деякий фактор має високу додатну кореляцію з оцінкою з вищої математики, інформатики, розділів фізики і високу від’ємну кореляцію з політології, філософії, історії − фактор характеризує точне мислення.

Створення математичної моделі факторного аналізу базується на припущенні про те, що усі зміни значень ознак обумовлені зміною деяких прихованих властивостей спостережуваних об'єктів. Ці приховані властивості називаються *загальними факторами* і їх кількість має бути меншою від числа ознак, за допомогою яких вони вимірюються. Кожний такий фактор має окреме значення значущості для різних досліджуваних ознак. Рівень значущості кожного фактора називається його *факторним навантаженням*. Він визначає степінь впливу загального фактора на зміну даної ознаки.

На зміну значень спостережуваної ознаки можуть впливати також деякі суб’єктивні, властиві тільки цій ознаці, зміни. Вони можуть бути викликані випадковими помилками, похибками вимірів і т.д. Причини усіх таких не взаємообумовлених змін об’єднуються в поняття *специфічного фактора***.** Отже, зміни значень спостережуваних ознак залежать від двох складових: загальних факторів і специфічних. Нехай X_i − *i*-та емпірична ознака. Позначимо через U_i − загальну частину цієї ознаки (частина змін, викликана впливом загальних факторів) і ε_i − специфічні фактори, зміни яких не пов'язані один з одним і не залежать від змін інших показників зміни ознаки X_i. Тоді зміна ознаки X_i розкладається на суму загальної частини усіх змін і змін, викликаних впливом специфічних факторів:

$$X_i = U_i + \varepsilon_i .$$

Подальший розвиток ідеї факторного аналізу ґрунтується на тому припущенні, що дані n змінних U_i є лінійними комбінаціями меншого числа інших змінних F_j, які називаються факторами, тобто

$$U_i = \omega_{1i}F_1 + \omega_{2i}F_2 + \ldots + \omega_{ki}F_k\,,\ i = \overline{1,n}$$

де ω_{ji} – факторні навантаження факторів F_j, які характеризують степінь впливу j -го загального фактора на i -у емпіричну ознаку. Об'єднуючи вище зазначені формули, отримаємо, що X_i виражається через загальні і специфічні фактори таким чином:

$$X_i = \omega_{1i}F_1 + \omega_{2i}F_2 + \ldots + \omega_{ki}F_k + \varepsilon_i\,,\ i = \overline{1,n}\,.$$

При цьому:

1. Загальні фактори F_j є або некорельованими випадковими величинами з дисперсією, що дорівнює 1, або невідомими невипадковими параметрами.

2. Специфічні фактори ε_i мають нормальний розподіл, не корелюють між собою і не залежать від загальних факторів.

Тому наступним кроком має бути оцінка значущості загальних і специфічних факторів для зміни значень ознаки X_i . Для цього розглянемо дисперсію ознаки X_i *:*

$$D(X_i) = \frac{1}{N-1}\sum_{p=1}^{N}(x_{ip} - \overline{x_i})^2\,,$$

де N – кількість об'єктів спостереження; x_{ip} – значення ознаки X_i для p–го об'єкта спостереження $p = \overline{1,n}$ *;* $\overline{x_i}$ – середнє значення i-ї ознаки. Так як U_i та ε_i не корелюють між собою, то:

$$D(X_i) = D(U_i) + D(\varepsilon_i)\,.$$

Математична модель факторного аналізу є розділом багатомірної статистики і потребує достатньо глибоких знань у

таких розділах вищої математики, як матрична алгебра, основи математичної статистики і математичного аналізу. Тому опишемо змістовну структуру факторного аналізу, опускаючи складні математичні і логічні обґрунтування.

Усі методи оцінки зв'язків ознак можна розділити на дві групи: прямі і непрямі. Прямі − це методи, які дозволяють визначити силу зв'язку до проведення факторного аналізу. До непрямих методів відносяться апостеріорні оцінки, коли до проведення факторного аналізу вони не відомі (беруться довільні оцінки), а в якості додаткової умови береться кількість факторів.

Найпоширеніший прямий метод оцінки зв'язку між ознаками базується на припущенні про те, що усі оцінки зв'язків дорівнюють 1. Тобто, вони дорівнюють діагональним елементам кореляційної матриці, а, отже, дисперсії специфічних факторів дорівнюють 0. У цьому полягає зміст однієї з найвідоміших моделей факторного аналізу − методу головних компонент.

Метод головних компонент. Для оцінки факторних навантажень, в якості критерію, використовується мінімум розбіжності між кореляційною матрицею початкових ознак і тією, яка виходить після оцінювання навантажень. Іншими словами, метод головних компонент здійснює перехід до нової системи координат, яка є системою ортонормованих лінійних комбінацій. Лінійні комбінації є власними векторами кореляційної матриці. Перша головна компонента − це лінійна комбінація, яка має найбільшу дисперсію. Друга компонента має найбільшу дисперсію серед усіх інших лінійних комбінацій, які не корелюють з першою головною компонентою.

Метод головних факторів. Це одна з найпоширеніших моделей факторного аналізу. Метод вимагає попередньої оцінки дисперсій і для нього критерієм оптимальної оцінки факторних навантажень є максимальна наближеність початкових кореляцій ознак до тих, які отримані в моделі після оцінювання навантажень. Для визначення кількості факторів використовуються різні статистичні критерії, за допомогою яких перевіряється гіпотеза про незначущість матриці кореляційних лишків.

Метод максимальної правдоподібності (Д. Лоулі) на відміну від попередніх моделей факторного аналізу ґрунтується не на попередній оцінці дисперсій, а на апріорному визначенні кількості загальних факторів. У разі великої вибірки дозволяє отримати статистичний критерій значущості отриманого факторного рішення.

Метод мінімальних залишків (Г. Харман) ґрунтується на мінімізації недіагональних елементів залишкової матриці кореляцій, і так як метод максимальної правдоподібності, вимагає попереднього вибору кількості факторів, що пояснюють спільні зміни ознак.

Перераховані методи відрізняються за способом пошуку розв'язання основного рівняння факторного аналізу. Вибір методу вимагає великого досвіду роботи. Проте деякі дослідники використовують відразу декілька методів, і виділені в усіх методах фактори вважають найбільш стійкими.

Моделі факторного аналізу називають іноді "прямими" або "початковими" в тому розумінні, що отримувані з їх допомогою оцінки навантажень напряму залежать від значень початкових ознак. Але на практиці може виникнути ситуація, коли навантаження при емпіричних ознаках утворюють таку комбінацію знаків і величин, яка важко інтерпретується. Ці випадки змусили дослідників шукати шляхи зміни початкових навантажень, щоб отримати результат, який краще інтерпретується. Один з методів вирішення цієї задачі − обертання факторів, тобто поворот відповідних факторам координатних осей, який проводиться не в просторі початкових ознак, а в просторі знайдених факторів.

Отже, наступним етапом факторного аналізу є обертання факторів, яке базується на принципах простої структури Терстоуна:

− в кожній стрічці факторної структури має бути хоча б один нуль;

− в кожному стовпчику − принаймні *к* нулів (*к* − кількість загальних факторів);

− для кожної пари стовпчиків можна знайти принаймні *к* параметрів (емпіричних ознак), для яких елементи факторної

структури дорівнюють нулю в одному з двох стовпчиків і не дорівнюють нулю − в іншому.

На основі цих принципів побудовано велику кількість аналітичних методів, які максимізують деякий критерій. Суть процесу обертання будь-якої пари векторів полягає в знаходженні такого кута між новим і старим напрямом факторів, який давав би найбільший приріст обраного критерію. Обертання факторів в просторі дає можливість кожній ознаці охарактеризуватися переважаючим впливом якогось одного фактора.

У сучасних пакетах статистичної обробки даних найчастіше використовуються методи обертання: варімакс, квартімакс і еквімакс. Обертання методом варімаксу спрощує значення стовпчиків факторної матриці, зводячи їх до 1 або 0. Обертання методом квартімаксу спрощує значення елементів стрічок факторної матриці. І, нарешті, еквімакс займає проміжне положення − при обертанні факторів за цим методом одночасно робиться спроба спростити значення елементів і стовпчиків, і стрічок.

Отримавши після процедури обертання факторне розв’язання (факторну матрицю), можна переходити до інтерпретації і найменування факторів. Цей етап роботи цілком і повністю залежить від інтуїції, рівня обізнаності і практичного досвіду дослідника. Щоб зрозуміти природу конкретного фактора, необхідно проаналізувати зміст ознак, які входять у даний фактор, і спробувати виявити спільні для них риси. Чим більша кількість ознак з великим значенням навантажень у цьому факторі, тим легше розкрити його природу. Для вибору назви фактора немає формалізованих прийомів. В якості попереднього варіанту можна використовувати ім'я ознаки, яка увійшла до фактора з найбільшим навантаженням.

Отже, процес інтерпретації факторних навантажень − це пошук таких загальних властивостей системи спостережень, які б могли бути описані в термінах одночасного збільшення (зменшення) однієї групи ознак на противагу зменшенню (збільшенню) іншої групи або просто збільшенню (зменшенню) якої-небудь частини ознак.

Якщо фактори знайдені і представлені, то на останньому кроці факторного аналізу, додатковим ознакам (додаткові запитання анкети) можна присвоїти значення цих факторів, так звані факторні значення. Тоді для кожного об'єкта спостереження значення великої кількості ознак можна перевести у значення невеликої кількості факторів − нових ознак.

Факторний аналіз є складною процедурою. Як правило, досконале факторне розв'язання (досить просте і таке, що змістовно інтерпретується) вдається отримати щонайменше після декількох циклів його проведення − від відбору ознак до спроби інтерпретації після обертання факторів. Для успішного проведення факторного аналізу необхідно дотримуватись основних вимог:

1) Змінні мають належати до шкали інтервалів (за класифікацією Стівенса). Передбачається, що порядкові змінні підлягають факторному аналізу, якщо їм надати числових значень. Не слід включати дихотомічні змінні в аналіз, якщо завдання не передбачають зменшення кількості ознак.

2) Відбираючи ознаки для факторного аналізу, слід враховувати, що на один фактор має припадати не менше трьох ознак.

3) Не варто включати у факторний аналіз ознаки, які мають дуже слабкі зв'язки з іншими ознаками. Велика ймовірність того, що вони не ввійдуть ні до одного фактора. Якщо в роботі не стоїть завдання сформувати шкалу нового опитування на основі факторного аналізу або якого-небудь аналогічного завдання, то не слід також включати усі ознаки, які мають між собою дуже тісні зв'язки. Швидше за все, вони утворюють один фактор. Чим більше таких ознак включається у факторний аналіз, тим більша ймовірність того, що вони утворюють перший фактор і до нього приєднається більшість інших ознак.

4) Стійкість виявленої факторної структури (її невипадковість) тим менша, чим більше складових її факторів. Вона також нестійка при малій кількості випробовуваних об'єктів.

Питання для самоконтролю

1. Що таке тренд?
2. Що таке сезонна варіація?
3. Як розраховується ковзке середнє?
4. Який алгоритм побудови адитивної моделі?
5. Який алгоритм побудови мультиплікативної моделі?
6. Який алгоритм побудови експоненційного згладжування?
7. Як здійснюється прогноз за методом експоненційного згладжування?
8. Яким чином константа згладжування впливає на прогноз?
9. Як розраховується помилка прогнозу?
10. Як обґрунтовується оптимальність моделі для прогнозування?
11. Що таке фактор?
12. Що таке факторне навантаження?
13. Чим відрізняються загальні та спеціальні фактори?
14. Які методи оцінки зв’язків факторного аналізу ви знаєте?

РОЗДІЛ 6

ПОБУДОВА ТА ДОСЛІДЖЕННЯ “ПАВУТИНОПОДІБНОЇ” МОДЕЛІ РИНКУ

“Павутиноподібна” модель ринку це приклад однієї з динамічних моделей, що описує процес формування попиту і пропозиції певного товару чи виду послуг на конкурентному ринку (за умов досконалої конкуренції). Дана модель адекватно описує динамічний процес коригування ціни і обсягу попиту(пропозиції) від одного стану локальної рівноваги до іншого.

При застосуванні “павутиноподібної” моделі має місце формалізація економічного закону попиту та пропозиції, який проголошує:

- кількість товару, яку можна реалізувати на ринку(попит), змінюється у напрямі, протилежному до зміни ціни товару;
- кількість товару, яку виробляють і доставляють на ринок(пропозиція), змінюється у тому ж напрямі, що й ціна товару;
- реальна ринкова ціна формується на рівні, за якого попит і пропозиція перебувають у стані локальної рівноваги.

Розглядають два випадки даної моделі: модель з запізненням пропозиції або модель з запізненням попиту. В розглянутій моделі цього розділу, вважаємо, що пропозиція із запізненням реагує на значення ціни в теперішній момент часу.

Нехай задано вибірку статистичних даних на основі значень попиту та пропозиції, що залежать від значення ціни.

	Ціна	**Попит**	**Пропозиція**
1	7,50	40,98	10,13
2	15,38	39,37	12,87
3	19,25	37,76	14,24
4	21,12	36,16	14,90
5	24,21	34,55	15,98
6	25,53	32,94	16,44
7	28,25	31,34	17,39

8	33,75	29,73	19,32
9	36,34	28,12	20,23
10	39,55	26,51	21,35
11	42,77	24,91	22,48
12	45,98	23,30	23,61
13	49,19	23,25	24,73
14	52,41	19,32	25,86
15	55,62	17,39	26,99
16	58,84	16,44	28,11
17	62,06	14,90	29,24
18	65,27	14,24	30,36
19	68,49	12,89	31,49
20	71,70	10,13	32,62

Побудуємо графіки емпіричних значень попиту та пропозиції з урахуванням вищезазначених статистичних даних (рис. 6.1):

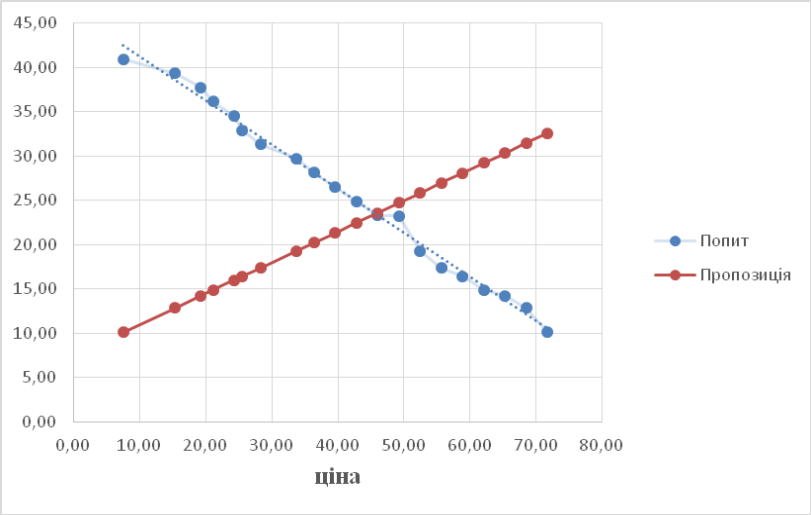


Рис. 6.1. Емпіричні значення попиту та пропозиції

Для знаходження теоретичних функціональних залежностей попиту та пропозиції очевидно необхідно застосувати алгоритм побудови лінійної регресійної моделі. Використовуючи функцію лінійної регресії, проведемо обчислення параметрів лінійної регресійної залежності попиту та пропозиції від ціни. Для цього розглянемо теоретичне рівняння лінійної регресії у вигляді $y = b_1 x + b_2$.

Значення параметрів b_1 та b_2 обчислимо використовуючи рівняння для знаходження цих змінних:

$$\begin{cases} b_1 \cdot \sum_{i=1}^{20} x_i^2 + b_2 \cdot \sum_{i=1}^{20} x_i = \sum_{i=1}^{20} x_i y_i, \\ b_1 \cdot \sum_{i=1}^{20} x_i + 20 b_2 = \sum_{i=1}^{20} y_i \end{cases}.$$

Отже отримаємо дані обчислень для теоретичної функції попиту та теоретичної функції пропозиції:

	b_1	b_2
D (попит)	-0,498238792	46,21897198
S (пропозиція)	0,350376303	7,494918313

На основі отриманих лінійних рівнянь теоретичної функції попиту та пропозиції, застосовуючи засоби Microsoft Excel, можемо провести регресійний аналіз коефіцієнтів цих рівнянь:

Регресійний аналіз попиту		
Регресійна статистика		
Множинний *R*	0,997	
R-квадрат	0,993	
Нормований *R*-квадрат	0,993	
Стандартна похибка	0,792	
Значення	20	
	Коефіцієнти	*Стандартна*

		похибка
Y-перетин	46,22	0,43
Змінна $x(D)$	–0,498	0,01

Регресійний аналіз пропозиції		
Регресійна статистика		
Множинний *R*	1	
R-квадрат	1	
Нормований *R*-квадрат	1	
Стандартна похибка	0,004708	
Значення	20	
	Коефіцієнт	*Стандартна похибка*
Y-перетин	7,494906	0,00256
Змінна $x(S)$	0,350387	5,67E-05

Визначимо теоретичну рівноважну P_E ціну для отриманих функції:

$$-0,49 \cdot P_E + 46,21 = 0,35 \cdot P_E + 7,49,$$
$$38,72 = 0,84 \cdot P_E,$$
$$P_E = 46,1.$$

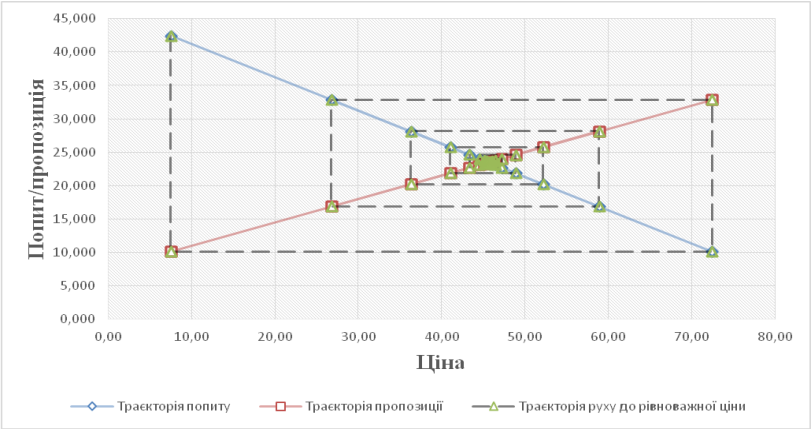
Дослідимо в аналітичній формі стан рівноваги на ринку в разі дискретної моделі (тобто з урахуванням запізнення реакції пропозиції на зміну попиту):

$$-0,49 \cdot P_t + 46,21 = 0,35 \cdot P_{t-1} + 7,49,$$
$$P_t = 77,72 - 0,703 \cdot P_{t-1}.$$

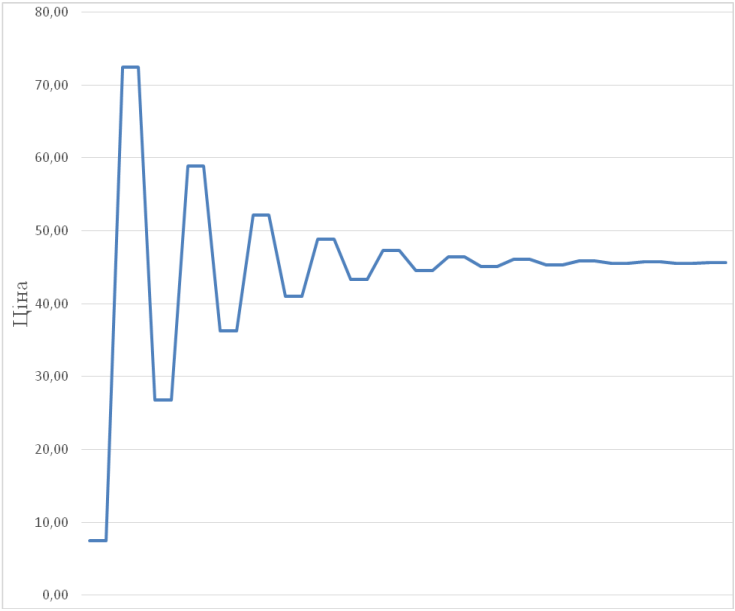
Із використанням функцій попиту та пропозиції й вищезазначеного рівняння рівноваги визначимо траєкторії зміни ціни, попиту та пропозиції. Для цього поведемо необхідні розрахунки:

Ціна	Траєкторія попиту	Траєкторія пропозиції	Траєкторія руху до рівноважної ціни
7,5	42,482	10,123	42,482
7,5	42,482	10,123	10,123
72,45	10,123	32,879	10,123
72,45	10,122	32,880	32,879
26,77	32,880	16,876	32,880
26,77	32,881	16,874	16,876
58,9	16,874	28,131	16,874
58,9	16,873	28,132	28,131
36,3	28,122	20,214	28,132
36,3	28,123	20,214	20,214
52,19	20,214	25,783	20,214
52,19	20,216	25,781	25,783
41,02	25,781	21,867	25,781
41,02	25,781	21,867	21,867
48,88	21,867	24,620	21,867
48,88	21,865	24,683	24,620
43,35	24,681	22,684	24,621
43,35	24,680	22,684	22,683
47,24	22,684	24,046	22,684
47,24	22,682	24,047	24,046
44,50	24,047	23,087	24,047
44,50	24,047	23,087	23,087
46,03	23,087	23,762	23,087
46,03	23,086	23,763	23,762
45,07	23,767	23,287	23,763
45,07	23,767	23,286	23,287
46,03	23,286	23,622	23,286
46,03	23,285	23,623	23,622

Побудуємо траєкторію руху рівноважної ціни згідно отриманих обчислень



та залежність зміни ціни з часом відносно рівноважної ціни:



З побудованих траєкторій отримаємо, що павутиноподібній спіралі притаманний збіжний цикл. Про це свідчать і затухання

коливань ціни та, відповідно, відновлення порушеної ринкової рівноваги.

Крім того, показник моделі $k = \left| \frac{b_1(s)}{b_1(d)} \right| = 0,607 < 1$, тобто ринкова ціна буде наближатися до рівноважної (46,1).

Отже маємо, що ринкові відносини знаходяться у стані стійкої рівноваги.

Водночас стійка ринкова рівновага, що спостерігається у досліджуваному кейсі, не повністю характеризує динаміку рівноважної ціни, тому така стійкість має відносний характер.

Загалом рух до ринкової рівноваги проходить низку циклів:

1. Надлишкова пропозиція призводить до зниження цін, унаслідок чого зрештою виникає надлишковий попит, що змушує ціни зростати.
2. Така ситуація згодом стає причиною нового надлишку пропозиції і т.д.

Необхідно відзначити, що "павутиноподібна" модель ринку може мати три стани ринкових відносин в залежності від отриманих характеристик рівнянь попиту та пропозиції:

- Ринок перебуває в стані стійкої рівноваги. Будь-які зміни, що відбуваються на ринку, не змінюють її. Відповідні зміни в обсягах попиту та пропозиції в кінцевому результаті приводять до стану ринкової стабільності.
- Ринок перебуває в стані циклічних коливань. Цей стан характеризується сталим відношенням (різницею) між обсягами попиту та пропозиції.
- Ринок перебуває в нестабільному стані. Будь-які намагання змінити та корегувати обсяги попиту та пропозиції призводять до додаткових і складніших змін, і тому система в цілому має нестабільний, нестійкий характер.

Список літератури

1. *Вітлінський В.В.* Економіко-математичне моделювання. Навч. посібник. / В.В. Вітлінський, С.І. Наконечний, О.Д. Шарапов та ін. – К.: КНЕУ, 2008.
2. *Голиков А.П.* Економіко-математичне моделювання світогосподарських процесів: навчальний посібник для студентів вищих навчальних закладів / А.П. Голиков. – Х.: ХНУ імені Каразіна, 2006.
3. *Леоненко М.М.* Теоретико-ймовірнісні та статистичні методи в економетриці тафінансовій математиці / М.М. Леоненко, Ю.С. Мішура, В.М. Пархоменко, М.Й. Ядренко. – К.: ІНФОРМТЕХНІКА, 1995. – 380 с.
4. *Пономаренко О.І.* Сучасний економічний аналіз. Мікроекономіка. Навч. посібник / О.І. Пономаренко, М.О. Перестюк, В.М. Бурим. – К.: Вища школа, 2004.
5. *Вашків П.Г.* Теорія статистики: Навчальний посібник / П.Г. Вашків, П.І. Пастер, В.П. Сторожук, Є.І. Ткач. – К.: Либідь, 2001. – 320 с.
6. *Бахрушин В.Є.* Методи аналізу даних: навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя: КПУ, 2011. – 268 с.
7. *Єріна А.М.* Статистичне моделювання та прогнозування / А.М. Єріна. – Київ: КНТЕУ, 2001. – 196 с.
8. *Клебанова Т.С.* Математичні методи і моделі ринкової економіки: навч. посібн. / Т.С. Клебанова, М.О. Кизим, О.І. Черняктаін. – Х. : ВД "ІНЖЕК", 2009. – 456 с.
9. *Пономаренко О.І.* Основи математичної економіки / О.І. Пономаренко, М.О. Перестюк, В.М. Бурим. – К.: ІНФОРМТЕХНІКА, 1995. – 319 с.
10. *Федоренко І.К.* Дослідження операцій в економіці: підручник / І. К. Федоренко, О. І. Черняк, Г. О. Чорноус,

О. О. Карагодова, О. В. Горбунов. – К.: Знання, КОО, 2007. – 558 с.

11. *Крючкова І.В.* Макроекономічне моделювання та короткострокове прогнозування. / За ред. І. В. Крючкової – Харків: Форт, 2000. – 336 с.

12. *Бутник О.М.* Економіко-математичне моделювання перехідних процесів у соціально-економічних системах: [монографія] / О.М. Бутник. – Х.:Видавничий Дім „ІНЖЕК"; СПД Лібуркіна Л.М., 2004. – 304 с.

13. *Медведєв М.Г.* Дослідження операцій: Навч. посіб. / М.Г. Медведєв, О.В. Колодінська. [2-ге вид., перер. і доп.]. – К.: Вид-во Європ. ун-ту, 2006. – 158 с.

14. *Карагодова О.О.* Дослідження операцій: навч. посіб. / О.О. Карагодова, В.Р. Кігель, В.Д. Рожок. – К.: Центр учбової літератури, 2007. – 256 с.

15. *Геєць В.М.* Моделі і методи соціально-економічного прогнозування: підручник / В.М. Геєць, Т.С. Клебанова, О.І. Черняк, А.В. Ставицький та інші. 2-е вид., виправ. – Х.: ВД «ІНЖЕК», 2008. – 396 с.

16. *Присенко Г.В.* Прогнозування соціально–економічних процесів: навч. посіб. / Г.В. Присенко, Є.І. Равікович. – Київ : КНЕУ, 2005. – 378 с.

17. *Макаренко Т.І.* Моделювання та прогнозування у маркетингу: навч.посіб. / Т.І. Макаренко. – К.: Центр навчальної літератури, 2005. – 160 с.

18. *Лук'яненко І.Г.* Сучасні економетричні методи у фінансах. Навчальний посібник / І.Г. Лук'яненко, Ю.О. Городніченко. – К.: Літера ЛТД, 2002. – 352 с.

19. *Грубер Й.В.* Економетрія: Вступ до множинної регресії та економетрії / Й.В. Грубер 2-х т. – К.: Нічлава, 1998. – Т.1. (384 с.), 1999. – Т.2 (308 с.).

20. *Хейс Д.М.* Причинный анализ в статистических исследованиях / Д.М. Хейс. – Финансы и статистика, 1981. – 225 с.

21. *Екимов С.В.* Нетрадиционные подходы в экономико-математическоммоделировании: [монография] / С.В. Екимов. – Днепропетровск: Наука и образование, 2004. – 240 с.

22. *Levine D.* Applied Statistics for Engineers and Scientists: Using Microsoft Excel & Minitab / D. Levine, P. Ramsey, R.Smidt. – Upper Saddle River, New Jersey: Prentice Hall Inc., 2001.

23. *Hill T.* STATISTICS: Methods and Applications / T. Hill, P.Lewicki. – Tulsa: StatSoft, OK, 2007. – 719 p. (StatSoft, Inc. (2011). Electronic Statistics Textbook. Tulsa, OK: StatSoft) (http://www.statsoft.com/textbook/)