7. Основні елементи телекомунікаційних мереж.
8. Призначення телекомунікацій.

ТЕМА 5. Статистичні методи аналізу маркетингової інформації

План

1. Обробка маркетингової інформації як шлях зменшення невизначеності маркетингового середовища.
2. Вибірковий аналіз.
3. Дисперсійний аналіз.
4. Кореляційний та регресійний аналіз.
5. Дискримінантний аналіз.
6. Кластерний аналіз.
7. Факторний аналіз.
8. Аналіз часових рядів.

Інформаційні джерела

1. Оксанич А. П. Іфнормаційні системи і технології маркетингу / А. П. Оксанич, В. Р. Петренко, О. П. Костенко : навч.-практ. посіб. – Київ : Видавничий дім «Професіонал», 2008. – 320 с.
2. Валентинов В. А. Эконометрика : практикум / В. А. Валентинов. – Москва : РДЛ, 2007. – 436 с.
3. Дослідження операцій : навч. посіб. / М. Г. Медведєв, О. В. Колодінська. – 2-ге вид., перероб. і допов. – Київ : Вид-во Європ. ун-ту, 2006. – 158 с.
4. Excel для экономистов и менеджеров / А. Г. Дубина, С. С. Орлова, И. Ю. Шубина, А. В. Хромов. – Санкт-Петербург : Питер, 2004. – 295 с.
5. Карагодова О. О. Дослідження операцій : навч. посіб. / О. О. Карагодова, В. Р. Кігель, В. Д. Рожок. – Київ : Центр учбової літератури, 2007. – 256 с.
6. Лапач С. Н. Статистика в науке и бизнесе / С. Н. Лапач, А. В. Чубенко, П. Н. Бабич. – Киев : МОРИОН, 2002. – 640 с.
7. Макаренко Т. І. Моделювання та прогнозування у маркетингу : навч. посіб. / Т. І. Макаренко. – Київ : Центр навч. л-ри, 2005. – 160 с.

8. Невежин В. П. Сборник задач по курсу «Экономико-математическое моделирование» / В. П. Невежин, С. И. Кружилов. – Москва : ОАО Издательский дом «Городец», 2005. – 320 с.
9. Просветов Г. И. Эконометрика: Задачи и решения : учебно-методическое пособие / Г. И. Просветов. – 4-е изд., доп. – Москва : Издательство РДЛ, 2007. – 192 с.
10. Системи оброблення економічної інформації : навч.-метод, посіб. для самост. вивч. дисц. / за заг. ред. В. Ф. Ситника. – Київ : КНЕУ, 2004. – 332 с.

1. Обробка маркетингової інформації як шлях зменшення невизначеності маркетингового середовища

Головна проблема маркетингу – це проблема інформаційного забезпечення. Залежно від того, якого типу ця інформація, якими є джерела інформації, які засоби збирання, передавання, дослідження та тлумачення результатів, вирішується питання використання її менеджерами підприємств. Саме за допомогою обґрунтованих маркетингових рішень підприємства пристосовують свою продукцію та послуги до потреб споживачів.

Кожен об'єкт (система) існує у реальному житті, перебуваючи у певній взаємодії із зовнішнім середовищем. Це об'єктивна реальність, незалежна від людини. Якщо цей об'єкт не впливає на наше життя і поведінку та взагалі ми нічого не знаємо про нього, то він для нас (суб'єктивно) не існує і не береться до уваги для прийняття рішень щодо нашого життя. Якщо ж, навпаки, існування об'єкта якимось чином причетне до нас, то, щоб визначити, яким є його вплив, необхідно дізнатись про нього «все» або принаймні ті його характеристики, що допоможуть нам скласти про нього уяву. Чим більше ми знатимемо про об'єкт, тим точніше наша уява про нього збігатиметься з реальним об'єктом. Наше уявлення про об'єкт є моделлю об'єкта, побудованою на тих даних (характеристиках), які нам відомі на той час. Наше уявлення про поведінку об'єкта насправді є поведінкою моделі з урахуванням відомих нам даних. І якщо вони не збігаються, то це означає, що ми мало знаємо про об'єкт, тобто існує деяка невизначеність щодо нього. Щоб позбутися цієї невизначеності або принаймні зменшити її, потрібні додаткові дані про об'єкт.

Одержати достовірну й достатню для прийняття рішення інформацію дасть змогу відповідний механізм роботи з інформацією, тобто визначена послідовність процедур аналізу інформації та її використання. Один із таких механізмів передбачає реалізацію таких процедур:

Перевірка інформації. Найбільш надійний і поширений метод перевірки інформації – порівняльний аналіз, тобто одержання однозначної відповіді на одне й те саме питання з різних джерел. Якщо інформація не підтвердилася одним або декількома джерелами чи отримано суперечливі відомості, необхідно поставити під сумнів усю отриману інформацію і підтвердити або спростувати припущення про її помилковість. Така неузгоджена інформація є неякісною і непридатною для прийняття рішення.

Обробка інформації. Після перевірки інформації на вірогідність відбувається аналіз її та формування (синтезування) виводу, тобто узагальнення, що пояснює всі встановлені факти. Після того, як сформульовано виводи, переходять до рекомендацій. Щоб підготувати вмотивовані виводи та рекомендації, необхідно дотримуватися двох обов'язкових умов: щоб з інформацією працював професіонал і щоб він був досить обережний, обґрунтовуючи рекомендації для керівництва фірми.

Відсіювання надлишкової інформації. Інформація не тільки полегшує і забезпечує швидке обґрунтоване прийняття рішень у будь-якій галузі діяльності з мінімальним ризиком, а й, якщо вона є надлишковою, паралізує аналіз інформації. Прийняття рішення в такій ситуації досить сумнівне, бо виявити ключову інформацію дуже складно, не говорячи вже про можливість отримання дезінформації.

Швидке змінювання інформації веде до перенапруження та прийняття помилкових рішень. Утримування штату кваліфікованих референтів-аналітиків для фірми не завжди можливе. Одним із способів вирішення цієї проблеми є використання «фільтра» на вході сторонньої інформації, яку не запитували.

Отримання маркетингової інформації у потрібному виді передбачає **використання** для обробки первинної інформації **різних математичних методів** (від простих вибіркових досліджень до використання штучних нейронних мереж).

Будь-який економічний об'єкт чи множина взаємодіючих об'єктів, об'єднаних у єдине ціле, розглядається як **система.** За системного підходу будь-який економічний об'єкт (наприклад ринок) чи множина взаємодіючих об'єктів, об'єднаних у єдине ціле, розглядається як система. Якою б не була система, її специфіка не вичерпується особливостями її складових, а ґрунтується на характері зв'язків і відношень між ними, що й визначає цілісність системи, її структуру та якісно нові властивості – системні властивості, що мають імовірнісно-статистичну природу й відображають статистичні закономірності функціонування і розвитку системи. Такі закономірності можна апроксимувати економіко-статистичними моделями. Ці моделі можуть класифікуватися за характером виявлених взаємозв'язків, за засобом відтворення їх, за характером використовуваної інформації, за засобом відображення структури впливів. Адекватність моделі реальному процесу залежить від методологічних принципів моделювання. Наприклад, за характером взаємозв'язків розрізняють моделі стохастичні та функціональні. Перші відображають стохастичний характер закономірностей функціонування системи, другі – зв'язок складових елементів розрахункових формул економічних показників. Найпростіший і найзручніший для аналізу варіант системи – сукупність великого обсягу однорідних елементів. *Однорідність* – це не точний збіг властивостей елементів, а наявність спорідненості в головному. Виділяють три форми зовнішнього прояву неоднорідності: у межах системи виділяються чітко розмежовані класи (типи) елементів; окремі елементи системи не можна однозначно зарахувати до якогось класу через відсутність чітких меж між типами (розмиті класи); окремі аномальні об'єкти, які мають своєрідні, нетипові для системи в цілому умови функціонування.

Для кожної з цих форм існує свій найбільш раціональний спосіб побудови моделей.

Склад незалежних змінних моделі називають ознаковою множиною, вони характеризують якісну особливість статистичних систем, специфіку зв'язку. Змінні включаються до моделі в результаті емпіричної перевірки їх впливу за допомогою статистичних критеріїв. Крім того, виконується диференційна оцінка їх значущості.

Процес вибору адекватної моделі має ітеративний характер і включає такі етапи:

1) із деякої множини допустимих моделей вибирається робоча модель;
2) вибрана модель застосовується до наявних даних;
3) визначається ступінь відповідності моделі реальним даним.

Після опрацювання всіх допустимих моделей вибирається модель із найвищим ступенем відповідності реальним даним і здійснюється осмислення отриманих результатів. Якщо вони тлумачаться як негативні, то необхідно перейти до пошуку іншої множини допустимих моделей.

Практично в усіх маркетингових дослідженнях використовуються статистичні методи аналізу інформації, які можна розбити на такі групи:

вибірковий аналіз – дозволяє встановити характер розподілу аналізуємого показника, визначити оцінки його математичного сподівання та дисперсії;

дисперсійний аналіз – використовується для виявлення впливу деякого фактора на певний економічний показник;

кореляційний аналіз – вивчає взаємодію та силу взаємозв'язку показників системи у процесі її функціонування;

регресійний аналіз – використовується для визначення залежності змінної від однієї чи декількох незалежних змінних;

факторний аналіз – використовується для дослідження взаємозв'язку між змінними з метою визначення найбільш впливових суттєвих факторів;

дискримінантний аналіз – використовується для визначення меж між заданими (існуючими) групами об'єктів за допомогою комбінації значень декількох незалежних змінних, що характеризують об'єкти, достатніх для розмежування груп та для віднесення будь-якого нового об'єкта до певної групи за його характеристиками ;

кластерний аналіз – використовується для об'єднання об'єктів у групи (кластери) так, щоб відмінності між об'єктами одного кластеру були меншими, ніж відмінності між об'єктами різних кластерів;

аналіз часових рядів **–** використовується для моделювання і прогнозування показників, дані про які представлені в вигляді часових рядів.

2. Вибірковий аналіз

Вибірковий аналіз – це аналіз, що ґрунтується на вивченні не всіх, а лише певної частини об'єктів, відібраних у випадковому порядку. Випадковість відбору гарантує незалежність вибірки від суб'єктивізму, упереджує умисність, тенденційність виконавців.

Основними поняттями під час проведення вибіркового аналізу є:

Генеральна сукупність – сукупність усіх уявних спостережень, які могли б бути виконані за даного реального комплексу умов. Поняття генеральної сукупності є поняття абстрактне і його не потрібно плутати з реальними сукупностями, що підлягають статистичному дослідженню. Генеральна сукупність називається скінченою або нескінченою, залежно від того, скінчена чи нескінчена сукупність усіх уявних спостережень.

Вибірка із даної генеральної сукупності – це результати обмеженого ряду спостережень $x_1, x_2, ... x_n$ випадкової величини x. Вибірку розглядають ще як деякий емпіричний аналог генеральної сукупності. Число n спостережень, утворюючих вибірку, називають ***об'ємом вибірки*.**

Одним із найважливіших питань, які вирішують під час формування вибірки, є забезпечення її репрезентативності, тобто повноти й адекватності репрезентації нею характеристик і властивостей усієї генеральної сукупності.

Вибірковий аналіз включає такі методи первинної обробки результатів спостережень, які забезпечують вирішення основних завдань:

– *перевірка однорідності вибірки та незалежності результатів спостережень* виконується з метою доказу, що всі результати спостережень належать до однієї генеральної сукупності. Для перевірки гіпотези про однорідність результатів спостережень використовують різні евристичні процедури, параметричні й непараметричні (рангові) критерії.

Приблизне уявлення про однорідність результатів спостережень визначається по величині вибіркового коефіцієнта варіації:

$$v(x) = s(x) / \bar{x}\,, \qquad (1)$$

де $s(x) = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$ – вибіркове середньоквадратичне відхилення;

$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ – вибіркове середнє значення;

n – об'єм вибірки.

Якщо виконується умова $v(x) < v_{таб}$ ($v_{таб}$ – табличне середнє значення варіації), то вважається, що ***вибірка – однорідна***. В іншому випадку – ***неоднорідна,*** з неї виключають крайні (екстремальні) значення, і для зменшеної таким чином вибірки знову виконується перевірка гіпотези про однорідність результатів спостережень.

Інша, більш строга процедура перевірки на однорідність [1] передбачає виключення аномалій результатів спостережень. Для цього із сукупності спостережень $x_i : i = \overline{1,n}$ обирається екстремальне (нехай, x) і обчислюється статистика (2):

$$\tau_p = \frac{|x - \bar{x}|}{s(x) \cdot \sqrt{(n-1)/n}}\,, \qquad (2)$$

яка залежить від рівня значимості α і кількості ступенів свободи $\nu = n - 1$. Отримане значення τ_p порівнюється з табличним $\tau_t(\alpha, \nu)$.

Якщо $\tau_p > \tau_t$, то x виключається із сукупності результатів спостережень і для зменшеної вибірки знову повторюється описана процедура перевірки на однорідність.

У разі впорядкування результатів спостереження за зростанням їх значення, використовують правило виключення k най-

більших членів варіаційного ряду (3.1):

$$L_K = \frac{\sum_{i=1}^{n-k}(x_i - \overline{x_k})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.1)$$

де $\overline{x_k}$ – середнє перших $n-k$ членів варіаційного ряду;

$\bar{x}$ – середнє за всією вибіркою.

За наявності аномальних спостережень статистика L_K буде меншою від табличного критичного значення, яке розміщено у спеціальних таблицях.

Якщо у вибірці можливі викиди і вліво, і вправо, тоді використовують таку формулу модифікацію статистики:

$$E_K = \frac{\sum_{i=1}^{n-k}(z_i - \overline{z_k})^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}, \qquad (3.2)$$

де $z = \left|x - \bar{x}\right|$ – елементи варіаційного ряду абсолютних відхилень спостережень від середнього значення.

У ході перевірки однорідності двох вибірок об'ємом n_1 і n_2, які мають нормальний розподіл, можна користуватись t-критерієм, з гідно з яким гіпотеза про однорідність приймається за умови виконання нерівності:

$$t_p = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{(n-1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_\tau(\alpha / 2, v), \qquad (4)$$

де $\bar{x}_i, \sigma_i^2$ – відповідно емпіричні середні і дисперсії вибірок;

α – рівень значимості;

$v = n_1 + n_2 - 2$ – кількість ступенів свободи;

$t_T(\alpha/2,\nu)$ – критичне значення, визначається за таблицями t-розподілу для заданих α і ν.

Якщо необхідно порівняти m вибірок на однорідність дисперсії (наприклад, під час оцінки однорідності випускаємої продукції), то рекомендується використовувати **критерій Барлетта**, згідно з яким гіпотеза про рівність дисперсій вибірок, що порівнюються, приймається за умови справедливості такої нерівності:

$$Q_p = \frac{2.3\cdot\left(\sum_{i=1}^{m} n_i \lg s^2 - \sum_{i=1}^{m} n_i \lg s_i^2\right)}{1+\frac{1}{3(m-1)}\left(\sum_{i=1}^{m}\frac{1}{n_i} - \frac{1}{\sum_{i=1}^{m} n_i}\right)} \le \chi^2(\alpha,\nu), \quad (5)$$

де $s = \frac{\sum_{i=1}^{m} n_i s_i^2}{\sum_{i=1}^{m} n_i}$, $s_i^2, i = \overline{1,m}$ – вибіркові дисперсії;

$n_i, i = \overline{1,m}$ – об'єми вибірок;

α – рівень значимості;

$\nu = m-1$ – кількість ступенів свободи;

$\chi^2(\alpha,\nu)$ – значення розподілу для заданих α і ν.

У випадку, коли $n_1 = n_2 = \ldots = n_m$, більш доцільніше для перевірки гіпотези про однорідність дисперсій вибірок використовувати зручний та більш точний **критерій Кохрана**. Гіпотеза про однорідність дисперсій за цим критерієм відхиляється за умови справедливості нерівності (6):

$$G_p = \frac{s_{\max}^2}{\sum_{i=1}^{m} s_i^2} \ge G(\alpha,m,\nu) \quad (6)$$

де $s_{\max}^2$ – максимальна вибіркова дисперсія;

α – рівень значимості;
$v = n - m$ – кількість ступенів свободи;
$G(\alpha, m, v)$ – критичне значення.

Щоб уникнути систематичної похибки результатів моделювання, необхідно перед статистичною обробкою вибіркової сукупності впевнитись в тім, що її можна розглядати як випадкову вибірку з незалежними даними. Така ситуація виникає, наприклад, під час побудови багатофакторної регресійної моделі, коли необхідно впевнитися, що включені до моделі вхідні показники є незалежними. Для перевірки незалежності спостережень використовується декілька критеріїв: критерій χ^2, ранговий критерій Спірмена, ранговий критерій Кендалла;

– *ідентифікація закону розподілу вибіркової сукупності.*

За відсутності апріорної інформації про закон розподілу вибіркової сукупності спочатку перевіряється її належність до симетричного розподілу, а потім до нормального або іншого закону розподілу, який є найбільш прийнятим за природою випадковості розглядаємого процесу.

Для перевірки *гіпотези про симетричність розподілу* необхідно побудувати інтервал довіри для невідомої ймовірності подій $x < \bar{x}$ за обчисленою частотою. Гіпотеза не відхиляється, якщо за ймовірності довіри $1 - \alpha = 0{,}95$ значення ймовірності появи дії в одиничному випробуванні $p = 0{,}5$ потрапляє в інтервал довіри:

$$\omega - 1{,}96\sqrt{\frac{\omega(1-\omega)}{n}} \le p \le \omega + 1{,}96\sqrt{\frac{\omega(1-\omega)}{n}}, \qquad (7)$$

де ω – вибіркове значення частоти.

Для ідентифікації *закону розподілу результатів спостережень* розроблена значна кількість статистичних критеріїв згоди, з яких найбільш часто використовувані ***критерій χ^2 Пірсона, критерій Колмогорова-Смирнова, критерій ω^2 Крамера-Мізеса-Смирнова.***

Критерії згоди призначені для перевірки гіпотези:

$$H_0 : F_\varepsilon(x) = F_{mod}(x;\theta_1,\theta_2,...,\theta_s) \tag{8}$$

і засновані на використанні різних мір відстані між аналізуємою емпіричною функцією розподілу $\hat{F}(x)$ (що визначається за вибіркою) і гіпотетичною модельною $F_{mod}(x;\theta_1,\theta_2,...,\theta_s)$.

На практиці спочатку необхідно визначити приблизне уявлення про відповідність емпіричного й модельного розподілу шляхом використання показників асиметрії $\hat{A}_s$ і ексцесу $\hat{E}_x$, які обчислюються за формулами:

$$\hat{A}_s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{nS^3(x)}; \tag{9}$$

$$\hat{E}_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{nS^4(x)} - 3, \tag{10}$$

де $\bar{x}$ – середнє значення вибірки;

$S(x)$ – середнє квадратичне відхилення.

Емпіричний розподіл узгоджується з теоретичним за умови, що вибірковий коефіцієнт асиметрії й ексцесу відрізняється за модулем від своїх математичних сподівань не більше, ніж на потроєні середні квадратичні відхилення, які визначаються за формулами (11) та (12):

$$S(\hat{A}_s) = \sqrt{6(n-2)/((n+1)(n+3))}, \tag{11}$$

$$S(\hat{E}_x) = \sqrt{24n(n-2)(n-3)/((n+1)^2(n+3)(n+5))}. \tag{12}$$

Тобто, для прийняття гіпотези H_0 необхідно, щоб одночасно виконувались такі нерівності (13):

$$\begin{aligned} \left|\hat{A}_s - E(A_s)\right| &\le 3S(\hat{A}_s) \\ \left|\hat{E}_x - E(E_x)\right| &\le 3S(\hat{E}_x) \end{aligned}, \tag{13}$$

де $E(\cdot)$ – оператор математичного сподівання.

Для нормального розподілу $E(A_s)=0$ і $E(E_x)=0$, тому (13) приймає вигляд (14):

$$\begin{aligned} \left|\hat{A}_s\right| &\le 3S(\hat{A}_s) \\ \left|\hat{E}_x\right| &\le 3S(\hat{E}_x) \end{aligned}. \qquad (14)$$

Критерій згоди Пірсона дозволяє виконувати перевірку гіпотези (8) в умовах, коли значення параметрів $\theta_1,\theta_2,....\theta_s$ модельної функції розподілу невідомі. Для прийняття гіпотези H_o необхідно виконання відношення:

$$\chi^2(1-\alpha/2,k-s-1) \le \sum_{i=1}^{k}\frac{(v_i-np_i)^2}{np_i} < \chi^2(\alpha/2,k-s-1), \qquad (15)$$

де n – об'єм вибірки;

k – кількість інтервалів групування спостережень ($k \ge max(8,s+1)$);

s – кількість параметрів модельного закону розподілу;

v_i – кількість спостережень, які потрапили в *i*-й інтервал;

$$p_i = F_{mod}(x_i^o;\hat{\theta}_1,\hat{\theta}_2,....\hat{\theta}_s) - F_{mod}(x_{i-1}^o;\hat{\theta}_1,\hat{\theta}_2,....\hat{\theta}_s),$$

де $\hat{\theta}_j, j=\overline{1-s}$ – вибіркові оцінки параметрів модельного закону розподілу;

x_i^o, x_{i-1}^o – правий і лівий кінці *i*-го інтервалу групування;

α – рівень значимості;

$k-s-1$ – кількість ступенів свободи;

$\chi^2(\beta,v)$ – табличне значення χ^2-розподілу для рівня значимості β і кількості ступенів свободи v.

Критерій згоди *Колмогорова-Смирнова* дозволяє виконувати перевірку гіпотези (8) в умовах, коли модельна функція $F_{mod}(x)-F_o(x)$ відома повністю, тобто не залежить від невідомих параметрів.

Позначимо через $F^{(n)}(x)$ і $F_o(x)$:

$$D_n = \underset{x\in R^1}{SUP}\left|F^{(n)}(x) - F_o(x)\right|;$$
$$D_n^+ = \underset{x\in R^1}{SUP}(F^{(n)}(x) - F_o(x)); \qquad (16)$$
$$D_n^- = \underset{x\in R^1}{SUP}(F_o(x) - F^{(n)}(x)).$$

Статистика $\sqrt{n}D$ є статистикою критерія Колмогорова, а статистика $\sqrt{n}D_n^-$ – статистикою критерія Смирнова.

Очевидно, що $D_n = max(D_n^+, D_n^-)$.

На практиці статистики Колмагорова-Смирнова використовуються у вигляді:

$$D_n = \underset{1\le s\le n}{max}(D_n^+, D_n^-);$$
$$D_n^+ = \underset{1\le s\le n}{max}(\frac{i}{n} - t_i); \qquad (17)$$
$$D_n^- = \underset{1\le s\le n}{max}(t_i - \frac{i-1}{n}),$$

де $t_i = F_0(x_i)$ – значення гіпотетичної функції розподілу в *i*-й позиції варіаційного ряду.

Для вищенаведених статистик відомі точні закони розподілу.

Статистикою критерія ω^2 **(Крамера-Мізеса-Смирнова)** є величина:

$$W_n^2 = n\omega_n^2 = n\int(F_o(x) - F^{(n)}(x))^2 dF_o(x), \qquad (18)$$

для обчислення якої на практиці використовують таку залежність:

$$W_n^2 = \sum_{i=1}^{n}\left(F_o(x_i) - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}.$$

У випадку справедливості гіпотези H_0 функція розподілу статистики W_n^2 збігається при $n\to\infty$ до граничного розподілу $a_1(x)$, значення яких наведені у спеціальних таблицях;

– *визначення раціонального об'єму вибірки залежно від закону розподілу похибки спостережень.*

Мінімальний об'єм вибірки n, необхідний для оцінювання параметрів багатофакторної моделі, залежить від кількості апріорної інформації про властивості досліджує мого процесу (I), структурних особливостей моделі (S), необхідної точності оцінювання (ε), кількості невідомих параметрів моделі (m), імовірності довіри ($1-\alpha$), коефіцієнта множинної кореляції результатів спостереження (R), має такий вигляд:

$$n = \varphi(I, S, \varepsilon, m, \alpha, R). \qquad (19)$$

На практиці використовують формули для обчислення n, які базуються на основі нерівностей Чебишева. Наприклад, мінімальний об'єм вибірки n, що забезпечує задану точність моделювання ε, визначається із співвідношення:

$$n = \frac{t^2(\alpha, v) C_v^2}{\Delta^2}, \qquad (20)$$

де $t^2(\alpha, v)$ – статистика, яка при $\alpha = 0{,}05$ має такі значення:

4,46 – для довільного закону розподілу;

2,96 – для унімонального симетричного закону розподілу;

1,96 – для нормального закону розподілу;

$v = n - 1$;

$C_v = \frac{S_x^2(n)}{\bar{x}(n)}$ – вибіркова оцінка коефіцієнта варіації

$(S_x^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)$;

$\Delta = \frac{\varepsilon}{\bar{x}(n)}$ – відносна похибка;

ε – задана похибка оцінки математичного сподівання x, тобто $m_x - \varepsilon \le \bar{x}(n) \le m_x + \varepsilon$.

Якщо структура математичної моделі представлена рівнянням лінійної багатофакторної моделі з m оцінюваними парамет-

рами за допустимої точності ε і коефіцієнта множинної кореляції R, то мінімальний об'єм вибірки визначається таким співвідношенням:

$$n = \frac{m(1-\varepsilon^2 R^2) + (1+R^2)}{(1-\varepsilon^2)R^2}. \qquad (21)$$

За повної апріорної невизначеності відносно властивостей досліджуваного процесу та структури моделі прийнято вважати, що

$$n = f(m) = (10 \div 30)m, \qquad (22)$$

тобто об'єм вибірки повинен в 10–30 разів перевищувати кількість оцінених параметрів.

3. Дисперсійний аналіз

Дисперсійний аналіз є одним із найбільш поширених і загальних методів статистичного аналізу. Він дозволяє оцінити наявність впливу досліджуваних факторів на результативну змінну. Цей метод базується на лінійній математичній моделі й дозволяє аналізувати не тільки кількісні, а і якісні фактори.

Сутність методу полягає в тому, що загальна варіація результуючого показника поділяється на частини, які відповідають роздільному та сукупному впливу різних якісних факторів, і залишкову варіацію, яка збирає вплив усіх інших факторів. Статистичне вивчення цих частин дозволяє робити висновки про вплив того чи іншого якісного фактору на результуючий показник.

У випадку однофакторного дисперсійного аналізу вивчається наявність чи відсутність впливу на результуючий показник одного якісного фактору. В основі однофакторного дисперсійного аналізу лежить наступна теоретико-ймовірносна схема:

$$Y_{ji} = a_i + \varepsilon_{ji};\ j = 1,...,n;\ i = 1,...,I, n = \sum_{i=1}^{I} n_i, \qquad (23)$$

де Y_{ji} – випадкові величини, які демонструють результуючу ознаку;

a_i − середнє (математичне сподівання) результуючої ознаки при *i*-ому значені якісної ознаки;

ε_{ji} − випадкові, нормально розподілені відхилення результуючої ознаки від середніх;

n_i − число спостережень при *i*-ому значенні якісного фактору.

Після проведення вибіркового експерименту отримаємо I груп вибіркових значень результуючої ознаки Y_{ji}, $j = 1,..., n_i$; $i = 1,2,...,I$. За цією вибіркою треба перевірити правильність гіпотези H_0: $\alpha_i = 0$; $i = 1, 2, ...,I$, або $a_1 = a_2 = ... = a_I = a$, тобто, що якісний фактор не впливає на результуючу ознаку.

Позначимо загальне та групове вибіркове середнє:

$$y^{сер} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\ ,\ y_i^{сер} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Вибіркові групові середні є незсуненими ($My^{сер} = a_i$), та обґрунтованими оцінками середніх a_i. Якщо, згідно з гіпотезою H_0 усі середні однакові, то загальне вибіркове середнє y не повинне статистично відрізнятися від групових середніх $y^{сер}{}_i$. В іншому випадку відмінність повинна бути статистично важливою.

Представимо повну суму квадратів відхилень результуючої ознаки від загального середнього у вигляді двох сум квадратів відхилень.

$$S^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y^{сер})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y_i^{сер} + y_i^{сер} - y^{сер})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y^{сер}{}_i)^2 +$$

$$+ 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y_i^{сер})(y_i^{сер} - y^{сер}) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_i^{сер} - y^{сер})^2 =$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - y_i^{сер})^2 + \sum_{i=1}^{I} n_i (y_i^{сер} - y^{сер})^2 = S_R^2 + S_A^2$$

Квадрат подвійної суми $(y_{ij} - y^{сер} + y^{сер} - y^{сер})^2$ призводе до трьох подвійних сум, які зводяться до двох, так як проміжкова сума обертається в нуль, тобто

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-y_i^{сер})(y_i^{сер}-y^{сер})=\sum_{i=1}^{I}(y_i-y^{ср})\sum_{j=1}^{n_i}(y_{ij}-y_i^{сер})=0,$$

оскільки $\sum_{j=1}^{n_i}(y_{ji}-y_i^{сер})=\sum_{j=1}^{n_i}y_{ji}-n_i y_i^{сер}=\sum_{j=1}^{n_i}y_{ji}-n_i\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}=0.$

Із тих сум, що залишилися, одна $S^2_A=\sum_{i=1}^{I}n_i(y_i^{сер}-y^{сер})^2$ є сумою квадратів відхилень між групами, тобто варіація обумовлена якісним фактором, а інша $S^2_R=\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-y_i^{сер})^2$ сума квадратів відхилень усередині груп, тобто залишкова варіація, що обумовлена випадковими відхиленнями від групових середніх.

$\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(y_{ji}-y_i^{сер})^2$ має розподіл X^2 з n_i-1 степенями вільності, відповідно, $\frac{S^2_R}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ji}-y_i^{сер})^2$ має розподіл X^2 с n-I степенями вільності. При $a_1=\ldots=a_I$ S^2_A и S^2_R незалежні та $\frac{S^2_A}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{I_i}n_i(y_i^{сер}-y^{сер})^2$ має розподіл X^2 з I-1 степенями вільності.

Так, у випадку вірності справедливості гіпотези H_0 F-відношення

$$F=\frac{S^2_A/(I-1)}{S^2_R/(n-I)}$$

має розподіл Фішера з I-1 та n-I степенями вільності.

Якщо гіпотеза правильна, то $y_i^{сер}$ та $y^{сер}$ є обґрунтованими оцінками одного й того ж математичного сподівання та, відповідно, близькі між собою, тому S^2_A мала. Якщо гіпотеза H_0 хибна, тобто a_i різні, тому $y_i^{сер}$ та $y^{сер}$ зближаються з різними математичними сподіваннями, при цьому S^2_A повинна приймати більші значення. Тобто для перевірки слушності гіпотези H_0 отримаємо такий статистичний критерій: якщо $F\le F_\alpha(I-1,n-I)$,

то гіпотеза приймається, в іншому випадку – уважається хибною. У цьому критерії α – помилка першого роду.

Приклад 1. Нехай проведено чотири види дослідів на кожному із трьох рівнів фактору *F*. Результати досліду занесені в табл. 3. Необхідно на рівні значущості 0,05 перевірити нульову гіпотезу про рівність групових середніх (нехай вибірки взято з нормальних сукупностей з однаковими дисперсіями).

Таблиця 3 – Результуюча таблиця

Номер досліду	**Рівень фактору**		
I	F_1	F_2	F_3
1	38	20	21
2	36	24	22
3	35	26	31
4	31	30	34
y_i^{cep}	35	25	27

Для нашого прикладу *n = 12, I = 3*.

Знайдемо відповідні групові середні y_i^{cep} та міжгрупову середню $y^{cep} = 29$ і занесемо їх до таблиці.

Відшукаємо S_R^2 S_A^2.

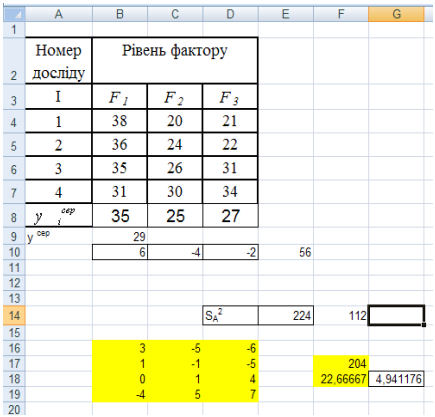
$$S_A^2 = \sum_{i=1}^{I} n_i (y_i^{cep} - y^{cep})^2 = 224. \qquad S_R^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} {}_i (y_{ij} - y_i^{cep})^2 = 204.$$

Визначимо вибіркову характеристику *F*:

$$F = \frac{S_A^2 / (I-1)}{S_R^2 / (n-I)} = \frac{112}{22{,}67} = 4{,}94.$$

Ураховуючи кількість степенів вільності 2 і 9, та рівень значущості 0,05, за таблицею розподілу Фішера знайдемо критичну точку $F_{кр}(0{,}05; 2; 9) = 4{,}26$. Оскільки $F > F_{кр}$, нульову гіпотезу відкидаємо.

Приклад виконання в ***Excel*** (рис. 6):



	A	B	C	D	E	F	G
1							
2	Номер досліду	Рівень фактору					
3	I	F_1	F_2	F_3			
4	1	38	20	21			
5	2	36	24	22			
6	3	35	26	31			
7	4	31	30	34			
8	$y_i^{сер}$	35	25	27			
9	$y^{сер}$	29					
10		6	-4	-2	56		
11							
12							
13							
14				S_A^2	224	112	
15							
16		3	-5	-6			
17		1	-1	-5		204	
18		0	1	4		22,66667	4,941176
19		-4	5	7			
20							

Рисунок 6 – Фрагмент виконання прикладу 1 у Excel

Або використати ***Данные/Анализ данных/Однофакторный дисперсионный анализ*** (рис. 7):

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	4	140	35	8,666666667
Столбец 2	4	100	25	17,33333333
Столбец 3	4	108	27	42

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	224	2	112	4,941176471	0,035631885	4,256494729
Внутри групп	204	9	22,66667			
Итого	428	11				

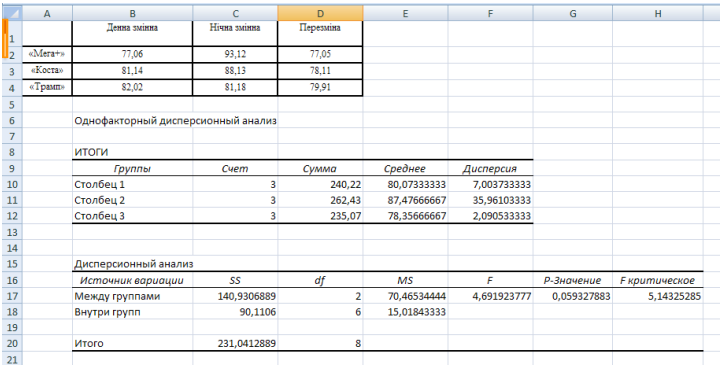
Рисунок 7 – Фрагмент виконання прикладу 1 у Excel

Приклад 2. Поставки продукції для компанії «Сігма» відбувається трьома постачальниками («Мега+», «Коста» та «Трамп») в різний час: у денний час, нічні зміни та навіть у перезміну. Контроль за якість продукції у денний час вище, ніж в інший. Зібрані данні з оцінки якості (у балах) наведені у табл. 4. Необхідно з'ясувати, чи є відмінність у якості продукції, яка поставляється в різний час?

Таблиця 4 – Вихідні данні

Постачальники	Денна змінна	Нічна змінна	Перезміна
«Мега+»	77,06	93,12	77,05
«Коста»	81,14	88,13	78,11
«Трамп»	82,02	81,18	79,91

Приклад виконання (рис. 8):



	A	B	C	D	E	F	G	H
1		Денна змінна	Нічна змінна	Перезміна				
2	«Мега+»	77,06	93,12	77,05				
3	«Коста»	81,14	88,13	78,11				
4	«Трамп»	82,02	81,18	79,91				
5								
6		Однофакторный дисперсионный анализ						
7								
8		ИТОГИ						
9		*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
10		Столбец 1	3	240,22	80,07333333	7,003733333		
11		Столбец 2	3	262,43	87,47666667	35,96103333		
12		Столбец 3	3	235,07	78,35666667	2,090533333		
13								
14								
15		Дисперсионный анализ						
16		*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
17		Между группами	140,9306889	2	70,46534444	4,691923777	0,059327883	5,14325285
18		Внутри групп	90,1106	6	15,01843333			
19								
20		Итого	231,0412889	8				
21								

Рисунок 8 – Фрагмент виконання прикладу 2 у Excel

З розрахунку бачимо, що что $F_{стат} < F_{критич}$ (4,691 < 5,14), тобто, відмінності у якості отриманої продукції у різний час відсутня. Крім того, *P-значение* (ймовірність істинності нульової гіпотези про рівність середніх) перевищує 0,05, тобто її не можна відхилити.

Отже, доведено, що якість отриманої продукції не залежить від часу її постачання і є однаковим у різний час.

Дисперсійний двофакторний аналіз застосовується в тих випадках, коли досліджується одночасна дія двох факторів на різні вибірки об'єктів, тобто коли різні вибірки опиняються під впливом різних поєднань двох факторів. Може статися, що одна змінна значущо діє на досліджувану ознаку тільки за певних значень іншої змінної. Наприклад, посилення мотивації може підвищувати швидкість рішення завдань у високоінтелектуальних осіб і знижувати її в низькоінтелектуальних. Отже, диспер-

сійний двофакторний аналіз дозволяє оцінити не лише вплив кожного з факторів, але й їхню взаємодію.

Суть методу залишається тією ж, як і за однофакторної моделі, але у двофакторному дисперсійному аналізі можна перевірити більшу кількість гіпотез, проте розрахунки дещо складніші, ніж в однофакторних комплексах.

Нехай необхідно визначити вплив двох факторів *A* і *B* на певну ознаку *X*. Для цього необхідно, щоб дослід здійснювався за фіксованих рівнів факторів *A* і *B*, а також їх одночасної дії на ознаку. При цьому дослід здійснюватимемо *n* раз для кожного з рівнів факторів *A* і *B*.

Позначимо через *k* конкретне значення ознаки *X*, якого вона набуває за *i*-го експерименту, *j*-го рівня фактора *A* і *k*-го рівня фактора *B*. Результат експерименту зручно подати у вигляді таблиці, яка поділена на блоки, в кожному з яких ураховується на певних рівнях факторів *A* і *B* їх вплив на конкретні значення ознаки (табл. 5).

Вивчають вплив на процес одночасно двох факторів *A* та *B*. Фактор *A* вивчається на рівнях a_1, a_2, ..., ak, фактор *B* – на рівнях b_1, b_2,…, b_m. Під час проведення дисперсійного аналізу в умовах лінійної моделі зручно використовувати нижченаведений алгоритм розрахунку. Знаходимо:

1) суми за стовпцями:

$$A_i = \sum_{j=1}^{m} y_{ij}, i = 1,2,\cdots,k; \tag{24}$$

2) суми за рядками:

$$B_i = \sum_{i=1}^{k} y_{ij}, j = 1,2,\cdots,m; \tag{25}$$

3) суму квадратів всіх дослідів:

$$SS_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}^2 \tag{26}$$

4) суму квадратів сум по стовпцях, поділену на число дослідів у стовпці:

$$SS_2 = \frac{1}{m}\sum_{i=1}^{k} A_i^2 \tag{27}$$

5) суму квадратів сум за стрічками, поділену на число дослідів у стрічці:

$$SS_3 = \frac{1}{k}\sum_{j=1}^{m} B_j^2 \tag{28}$$

6) квадрат загальної суми, поділений на число всіх дослідів (корегуючий член):

$$SS_4 = \frac{1}{mk}\left(\sum_{i=1}^{k} A_i\right)^2 = \frac{1}{mk}\left(\sum_{j=1}^{m} B_j\right)^2 \tag{29}$$

7) суму квадратів для стовпця:

$$SS_A = SS_2 - SS_4; \tag{30}$$

8) суму квадратів для стрічки:

$$SS_B = SS_3 - SS_4; \tag{31}$$

9) загальну суму квадратів, рівну різниці між сумою квадратів усіх дослідів та корегуючим членом:

$$SS_{\text{заг.}} = SS_1 - SS_4; \tag{32}$$

10) залишкову суму квадратів:

$$SS_{\text{зал.}} = SS_{\text{заг.}} - SS_A - SS_B = SS_1 - SS_2 - SS_3 + SS_4; \tag{33}$$

11) дисперсію S_A^2:

$$S_A^2 = \frac{SS_A}{k-1}; \tag{34}$$

12) дисперсію S_B^2:

$$S_B^2 = \frac{SS_B}{m-1}; \tag{35}$$

13) дисперсію $S_{\text{пом.}}^2$:

$$S_{\text{пом.}}^2 = \frac{SS_{\text{зал.}}}{(k-1)(m-1)}. \tag{36}$$

Результати дисперсного аналізу зручно представляти у вигляді таблиці дисперсного аналізу (табл. 5).

Таблиця 5 – Двофакторний дисперсійний аналіз (без паралельних дослідів)

Джерело дисперсії	Число ступеня вільності	Сума квадратів	Середній квадрат	Математичне сподівання середнього квадрату
A	$k-1$	SS_A	S_A^2	$m\sigma_A^2 + \sigma_{пом.}^2$
B	$m-1$	SS_B	S_B^2	$k\sigma_B^2 + \sigma_{пом.}^2$
Залишок	$(k-1)(m-1)$	$SS_{зал.}$	$S_{пом.}^2$	$S_{пом.}^2$
Загальна сума	$km-1$	$SS_{заг.}$	–	–

Установивши за допомогою дисперсійного аналізу значення впливу даного фактора, потім за допомогою критерію Стьюдента чи рангового критерію Дункана з'ясовуть, які саме середні значення різняться.

Лінійна модель справедлива, коли між факторами A та B немає взаємодії. У протилежному випадку цій взаємодії як фактору присутня своя дисперсія S_{AB}^2. Взаємодія AB, S_{AB}^2 є мірою того, наскільки вплив фактора A залежить від рівня фактора B та, навпаки, наскільки вплив фактора B залежить від рівня A.

У ході проведення дисперсійного аналізу за нелінійної моделі зручно користуватись нижченаведеним алгоритмом розрахунку. Знаходимо:

1) суми спостережень у кожній комірці:

$$y_{ij} = \sum_{u=1}^{n} y_{iju}, \quad j = 1,2,\cdots,m, \quad i = 1,2,\cdots,k; \tag{37}$$

2) квадрат суми спостережень у кожній комірці:

$$y_{ij}^2 = \left(\sum_{u=1}^{n} y_{iju} \right)^2; \tag{38}$$

3) суми за стовпцями:

$$A_i = \sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju} \; ; \tag{39}$$

4) суми за рядами:

$$B_j = \sum_{i=1}^{k}\sum_{u=1}^{n} y_{iju} \; ; \tag{40}$$

5) суму всіх спостережень (загальна сума):

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju} = \sum_{i=1}^{k} A_i = \sum_{j=1}^{m} B_j \; ; \tag{41}$$

6) суму квадратів усіх спостережень:

$$SS_1 = \sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju}^2 \; ; \tag{42}$$

7) суму квадратів сум за стовпцями, поділену на число спостережень у стовпці:

$$SS_2 = \frac{1}{mn}\sum_{i=1}^{k} A_i^2 \; ; \tag{43}$$

8) суму квадратів сум за рядами, поділену на число спостережень у стрічці:

$$SS_3 = \frac{1}{kn}\sum_{j=1}^{m} B_j^2 \; ; \tag{44}$$

9) квадрат загальної суми, поділений на число всіх спостережень (корегуючий член):

$$SS_4 = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{u=1}^{n} y_{iju}\right)^2}{N} = \frac{1}{mkn}\left(\sum_{i=1}^{k} A_i\right)^2 = \frac{1}{mkn}\left(\sum_{j=1}^{m} B_j\right)^2 ; \tag{45}$$

10) суму квадратів для стовпця:

$$SS_A = SS_2 - SS_4; \tag{46}$$

11) суму квадратів для ряду:

$$SS_B = SS_3 - SS_4; \tag{47}$$

12) суму квадратів для дисперсії відтворення:

$$SS_{пом.} = SS_1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij}^2}{n}; \quad (48)$$

13) загальну суму квадратів, рівну різниці між сумою квадратів усіх спостережень і коректуючим членом:

$$SS_{заг.} = SS_1 - SS_4; \quad (49)$$

14) залишкову суму квадратів відхилень для ефекту взаємодії *AB*:

$$SS_{AB} = SS_{заг.} - SS_A - SS_B - SS_{пом.}; \quad (50)$$

15) дисперсію S_A^2:

$$s_A^2 = \frac{SS_A}{k-1}; \quad (51)$$

16) дисперсію S_B^2:

$$s_B^2 = \frac{SS_B}{m-1}; \quad (52)$$

17) дисперсію S_{AB}^2:

$$s_{AB}^2 = \frac{SS_{AB}}{(k-1)(m-1)}; \quad (53)$$

18) дисперсію відтворення:

$$s_{пом.}^2 = \frac{SS_{пом.}}{mk(n-1)}. \quad (54)$$

Перевірку гіпотези про значимість взаємодії факторів *A* та *B* проводять за *F*-критерієм однаково для моделей із випадковими й фіксованими рівнями. Але перевірку гіпотези про значимість факторів *A* і *B* проводять неоднаково для різних моделей. У табл. 6 приведений двофакторний дисперсійний аналіз із паралельними дослідами для моделі з випадковими рівнями.

Таблиця 6 – Двофакторний дисперсійний аналіз для моделі з випадковими рівнями (з паралельними дослідами)

Джерело дисперсії	Число ступеня вільності	Сума квадратів	Середній квадрат	Математичне сподівання середнього квадрату
A	$k-1$	SS_A	S_A^2	$nms^2_A + ns^2_{AB} + s^2_{\text{пом.}}$
B	$m-1$	SS_B	S_B^2	$nks^2_B + ns^2_{AB} + s^2_{\text{пом.}}$
AB	$(k-1)(m-1)$	SS_{AB}	S_{AB}^2	$ns^2_{AB} + s^2_{\text{пом.}}$
Залишок (помилка)	$mk(n-1)$	$SS_{\text{пом.}}$	$S^2_{\text{пом.}}$	$s^2_{\text{пом.}}$
Загальна сума	$mkn-1$	$SS_{\text{заг.}}$	–	–

З табл. 6 видно, що для оцінки значимості фактора A необхідно скласти дисперсійне відношення виду:

$$F = \frac{S_A^2}{S_{AB}^2}. \tag{55}$$

Вплив фактора A признається значимим, коли:

$$\frac{S_A^2}{S_{AB}^2} > F_{1-p}(f_1, f_2), \tag{56}$$

де p –рівень значимості; $f_1 = k-1$; $f_2 = (k-1)(m-1)$.

Аналогічно, вплив фактора B уважається значимим, коли:

$$\frac{S_B^2}{S_{AB}^2} > F_{1-p}(f_1, f_2), \tag{57}$$

де p – рівень значимості; $f_1 = m-1$; $f_2 = (k-1)(m-1)$.

Коли нерівності (56) та (57) не виконуються, вплив факторів A та B слід рахувати незначним.

Для математичної моделі з фіксованими рівнями члени, що відповідають взаємодії, зникають із сум квадратів відхилень SS_A та SS_B.

Унаслідок цього для оцінки значимості фактора A складають дисперсійне відношення:

$$F = \frac{s_A^2}{s_{пом.}^2}, \tag{58}$$

у знаменнику котрого стоїть оцінка для дисперсії відтворення. Одержане дисперсійне відношення порівнюється з табличним $F_{1-p}(f_1, f_2)$ для чисел степенів вільності $f_1 = k - 1$, $f_2 = mk(n - 1)$. Аналогічно, для оцінки фактора В розглядають відношення

$$F = \frac{s_B^2}{s_{пом.}^2}, \tag{59}$$

яке порівнюють з табличним $F_{1-p}(f_1, f_2)$ для чисел степенів вільності $f_1 = m - 1$, $f_2 = mk(n - 1)$.

Якщо дисперсійні відношення (58) і (59) більше табличних

$$\frac{s_A^2}{s_{AB}^2} > F_{1-p}(f_1, f_2) \quad \text{та} \quad \frac{s_B^2}{s_{AB}^2} > F_{1-p}(f_1, f_2), \tag{60}$$

вплив факторів A та B слід рахувати значним.

Для перевірки значимості ефекту взаємодії складають дисперсійне відношення:

$$F = \frac{s_{AB}^2}{s_{пом.}^2} \tag{61}$$

і порівнюють його з табличним $F_{1-p}(f_1, f_2)$ для рівня значимості p та чисел степенів вільності $f_1 = (m - 1)(k - 1)$, $f_2 = mk(n - 1)$. Якщо одержане дисперсійне відношення більше табличного, то вплив ефекту взаємодії факторів слід уважати значним, в іншому випадку – вплив ефекту взаємодії вважають несуттєвим.

4. Кореляційний та регресійний аналіз

Будь-який соціально-економічний об’єкт або явище зазвичай характеризується за декількома ознаками, тобто різними властивостями. Ці ознаки взаємозв’язані та впливають одна на одну. Крім того, може існувати зв’язок між ознаками різних об’єктів і явищ. Тому в математичній статистиці розроблений апарат для

виявлення таких зв'язків та оцінки їх сили (тісноти). Цей математичний апарат називається кореляційним аналізом.

У багатьох прикладних задачах необхідно виявити залежність між двома властивостями (ознаками) X і Y одного й 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**, або просто фактором, а Y – залежною змінною або **результативною ознакою.**

Якщо кожному значенню факторної ознаки X відповідає одне й 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 $X = x_1, x_2, ..., x_l$ і кожному її значенню x_i відповідає множина значень Y, тобто:

значенню x_1 відповідає множина $y_{11}, y_{12}, ..., y_{1m_1}$;

значенню x_2 відповідає множина $y_{21}, y_{22}, ..., y_{2m_2}$;

…

значенню x_l відповідає множина $y_{l1}, y_{l2}, ..., y_{lm_l}$,

то між X та Y існує статистичний зв'язок.

Вивчення статистичного зв'язку дуже складний і трудомісткий процес, у якому потрібно аналізувати багатомірні таблиці даних. Тому зазвичай 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 $X = x_1, x_2, ..., x_l$ і кожному її значенню x_i відповідає середнє множини значень Y, тобто:

значенню x_1 відповідає $\bar{y}_{x_1} = \dfrac{y_{11} + y_{12} + ... + y_{1m_1}}{m_1}$;

значенню x_2 відповідає $\bar{y}_{x_2} = \dfrac{y_{21} + y_{22} + ... + y_{1m_2}}{m_2}$;

…

значенню x_l відповідає $\bar{y}_{x_l} = \dfrac{y_{l1} + y_{l2} + ... + y_{lm_l}}{m_l}$;

то між X та Y існує кореляційний зв'язок.

Основними завданнями кореляційного аналізу є:

– вивчення сили зв'язку між двома й більше ознаками досліджуваного об'єкта;

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

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

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

Коефіцієнт кореляції Пірсона розраховується за формулою:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}, \qquad (62)$$

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.

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

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

$$S_x=\sqrt{\frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i-\left(\frac{1}{n}\sum_{i=1}^{k}x_i\,n_i\right)^2}\ ;\ S_y=\sqrt{\frac{1}{n}\sum_{j=1}^{m}y_j^2 n_j-\left(\frac{1}{n}\sum_{j=1}^{m}y_j n_j\right)^2},$$

отримують більш зручну для розрахунків формулу:

$$r=\frac{n\sum_{i=1}^{k}\sum_{j=1}^{m}x_i y_j n_{ij}-\left(\sum_{i=1}^{k}x_i n_i\right)\left(\sum_{j=1}^{m}y_j n_j\right)}{\sqrt{n\sum_{i=1}^{k}x_i^2 n_i-\left(\sum_{i=1}^{k}x_i n_i\right)^2}\sqrt{n\sum_{j=1}^{m}y_j^2 n_j-\left(\sum_{j=1}^{m}y_j n_j\right)^2}}. \qquad (63)$$

У випадку незгрупованих даних розрахункова формула суттєво спрощується:

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}. \qquad (64)$$

Властивості коефіцієнта кореляції Пірсона

1) коефіцієнт кореляції Пірсона приймає значення на проміжку $-1;1$, тобто $-1\le r\le 1$;

2) якщо $|r|\le 0{,}5$, то зв'язок вважається слабким; якщо $|r|\le 0{,}7$, то зв'язок вважається середнім; $|r|>0{,}7$, 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. Слід пам'ятати, що коефіцієнт кореляції Пірсона показує силу лінійного зв'язку. Якщо між X та Y існує сильний нелінійний зв'язок, 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:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (65)$$

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

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

Приклад 3. За наявними даними про рівень механізації праці X (%) і продуктивності праці Y (од. продукції/год) для 14 однотипних підприємств (табл. 7) оцінити тісноту зв'язку між X і Y. Визначити можливість розповсюдження результатів розрахунків на всі підприємства такого типу.

Таблиця 7 – Вихідна таблиця

Підприємство	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Рівень механізації праці, %	32	30	36	40	41	47	56	54	60	55	61	67	69	76
Продуктивність праці, од. продукції/год	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Розв'язок. Дані табл. 7 є вибіркою значень X і відповідних значень Y. Оскільки кількість даних невелика ($n = 14$), то їх можна не групувати. Для оцінки тісноти зв'язку між X і Y розрахуємо коефіцієнт кореляції Пірсона за формулою (63) для незгрупованих даних. Розрахунки для зручності оформимо у вигляді табл. 8.

Таблиця 8 – Розрахункова таблиця

x_i	y_i	x_i^2	y_i^2	$x_i\ y_i$
32	20	1 024	400	640
30	24	900	576	720
36	28	1 296	784	1 008
40	30	1 600	900	1 200
41	31	1 681	961	1 271
47	33	2 209	1 089	1 551
56	34	3 136	1 156	1 904
54	37	2 916	1 369	1 998
60	38	3 600	1 444	2 280
55	40	3 025	1 600	2 200
61	41	3 721	1 681	2 501
67	43	4 489	1 849	2 881
69	45	4 791	2 025	3 105
76	48	5 779	2 304	3 848
Суми				
724	492	40 134	18 138	26 907

Отже,

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2}} =$$

$$\frac{14 \cdot 26\,907 - 724 \cdot 492}{\sqrt{14 \cdot 40\,134 - 724^2}\sqrt{14 \cdot 18\,138 - 492^2}} = \frac{20\,490}{\sqrt{37\,700}\sqrt{11\,868}} \approx 0{,}969.$$

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 (3.4): $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}969\sqrt{14-2}}{\sqrt{1-0{,}969^2}} \approx 13{,}59$. Знайдемо $t_{\text{крит}}$, ураховуючи, що $l = n - 2 = 14 - 2 = 12$. Оберемо рівень значущості $\alpha = 0{,}01$. Тоді $t_{\text{крит}}$ = СТЬЮДРАСПОБР (0,01; 12) = 3,055.

Оскільки розраховане значення *t*-статистики більше критичного $13{,}59 > 3{,}055$, то коефіцієнт кореляції можна вважати значимим на обраному рівні $\alpha = 0{,}01$.

Висновок. Між рівнем механізації праці та її продуктивністю на підприємствах, що досліджувалися, існує сильний додатній зв'язок: чим більше рівень механізації праці, тим вище її продуктивність. Висновок дійсний для всіх підприємств такого типу.

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, слугує коефіцієнт кореляції Спірмена.

Коефіцієнт кореляції Спірмена розраховується за формулою:

$$r_S(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2 + T_X + T_Y}{n(n^2-1)}, \qquad (66)$$

де n – кількість пар вибіркових даних;

d_i – різність між рангами *i*-го значення *X* та відповідного значення *Y*;

T_X, T_Y – поправки, що пов'язані з однаковими рангами; розраховуються за формулами:

$$T_X = \frac{\sum_{i=1}^{L_X} T_{Xi}^3 - T_{Xi}}{12}; \; T_Y = \frac{\sum_{i=1}^{L_Y} T_{Yi}^3 - T_{Yi}}{12}, \qquad (67)$$

де L_X, L_Y – кількість зв'язок (груп однакових рангів);

T_{Xi}, T_{Yi} – розміри *i*-тих зв'язок (кількість елементів в них).

У випадку, коли досліджуваний об'єкт або явище характеризується більш ніж двома ознаками X_1, X_2, … , X_k, необхідно вивчати множинні залежності. Для оцінки сили зв'язку між певною ознакою X_i та усіма іншими ознаками слугує **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:

$$A=\begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix}; \qquad (68)$$

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

3) розрахувати множинний коефіцієнт кореляції за формулою:

$$R_i=\sqrt{1-\frac{|A|}{A_{ii}}}. \qquad (69)$$

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

$$t=\frac{R^2 \quad n-k}{1-R^2 \quad k-1}, \qquad (70)$$

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

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

Якщо розраховане значення *t*-статистики більше критичного $|t| > F_{\text{крит}}$, то множинний коефіцієнт кореляції вважається значимим 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, вільний від впливу всіх 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_{ii}, A_{jj}, A_{ij} елементів r_{ii}, r_{jj}, r_{ij} відповідно;

3) розрахувати частинний коефіцієнт кореляції за формулою:

$$R_{ij} = \frac{-A_{ij}}{\sqrt{A_{ii} A_{jj}}}. \qquad (71)$$

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

$$t = \frac{R_{ij}\sqrt{n-k+2}}{\sqrt{1-R_{ij}{}^2}}, \qquad (72)$$

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

Розраховане значення *t*-статистики порівнюється із критичним значенням $t_{\text{крит}}$. Табличне значення розподілу Стьюдента ($t_{\text{крит}}$) визначається за допомогою вбудованої статистичної функції Excel СТЬЮДРАСПОБР (α ; l), де α – обраний дослідником рівень значущості, l – ступені волі, $l = n - k + 2$.

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

Зауваження:

1. Уважається, 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 в розрахунках.

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

3. 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_i.

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) обрати ***Сервис – Анализ данных***;

2) у діалоговому вікні для вибору інструменту аналізу обрати інструмент ***Корреляция***. З'явиться вікно для надання параметрів (рис. 9);

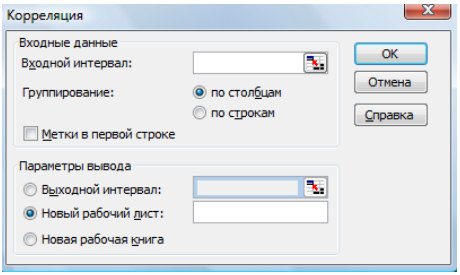


Рисунок 9 – Вікно надання параметрів кореляційного аналізу

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 клітину, починаючи з якої

будуть надаватися вихідні дані – парні коефіцієнти кореляції. Натиснути ***ОК***.

Приклад 4 і результати розрахунків парних коефіцієнтів кореляції надано на рис. 10.

	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					

Рисунок 10 – Фрагмент розрахунку коефіцієнтів кореляції у Excel

Зауваження.

1. У результаті роботи інструменту 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, звичайно надаються 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 знайти визначник заданої матриці.

Під час вивчення стохастичних зв'язків між різними ознаками економічного об'єкта головним завданням є встановлення виду кореляційної залежності результативної ознаки (Y) від

факторної (X), тобто виду функціональної залежності $\overline{Y} = f(X)$. У 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, – **рівнянням регресії**.

Регресійний аналіз проводиться за такими етапами:

1. Установлення 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: $x_1; y_1$, $x_2; y_2$, …, $x_n; y_n$, 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 і будується статистичний ряд, що містить значення X, відповідні середні значення Y та частоти n. Згруповані дані (табл. 9) зображуються графічно, що часто дозволяє визначити вид залежності Y від X.

Таблиця 9 – Статистичний ряд вхідних даних

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

Ламана лінія, що сполучає крапки з координатами $(x_i; \overline{y_{x_i}})$, 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. 11).

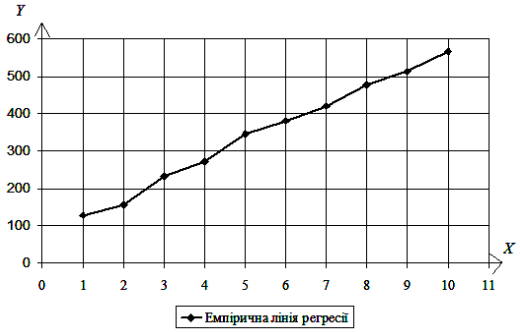


Рисунок 11 – Гіпотетична лінійна залежність

В іншому випадку висувається гіпотеза про наявність нелінійного зв'язку (рис. 12).

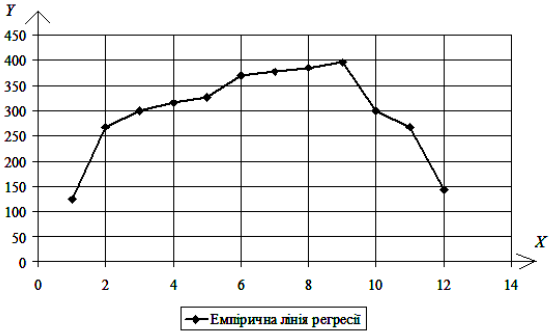


Рисунок 12 – Гіпотетична нелінійна залежність

Лінійна регресія

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

$$\overline{y_x} = ax + b, \qquad (73)$$

де a, b – параметри моделі.

Побудова лінійної регресійної моделі – це знаходження параметрів рівняння (73). Параметри рівняння регресії, зазвичай, знаходяться за **методом найменших квадратів**.

Ідея методу найменших квадратів

Нехай під час вивчення залежності Y від X було отримано вибіркові дані: $x_1, x_2, ..., x_n$ – значення величини X, $y_1, y_2, ..., y_n$ – відповідні значення Y. За вибірковими даними було побудовано рівняння регресії $y = ax + b$. Якщо в рівняння підставити замість x значення $x_1, x_2, ..., x_n$, то будуть отримані теоретичні значення Y: $y_{1,теор}, y_{2,теор}, ..., y_{n,теор}$, які відрізняються від $y_1, y_2, ..., y_n$. Різниця значень $y_{i,теор} - y_i$ називається помилкою регресійної моделі й позначається e_i. Якщо параметри рівняння підбираються так, щоб сума квадратів помилок була мінімальною, то говорять, що вони отримані за методом найменших квадратів.

У випадку лінійної регресії параметри рівняння регресії за методом найменших квадратів знаходяться із системи лінійних алгебраїчних рівнянь:

$$\begin{cases} a\sum\limits_{i=1}^{k} x_i^2 n_i + b\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum\limits_{i=1}^{k} x_i n_i + b\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}. \qquad (74)$$

Якщо вибіркові дані не згруповані, то система (74) значно спрощується:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} x_i + bn = \sum\limits_{i=1}^{n} y_i \end{cases}. \qquad (75)$$

Перевірка правильності побудови рівняння регресії здійснюється за основним варіаційним рівнянням:

$$Q = Q_p + Q_o, \qquad (76)$$

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, $\bar{y} = \frac{\sum_{i=1}^{k} \overline{y_{x_i}} n_i}{n}$;

$Q_p = \sum_{i=1}^{k} y_{i,\text{теор}} - \bar{y}^{\,2} n_i$ – варіація регресії, тобто сума квадратів відхилень теоретичних значень Y від середнього, що обумовлена регресією; $Q_o = \sum_{i=1}^{k} y_{i,\text{теор}} - \bar{y}_{x_i}^{\,2} n_i$ – варіація залишків, тобто сума квадратів відхилень теоретичних значень Y від емпіричних. У випадку незгрупованих даних загальна варіація, варіації регресії і залишків знаходяться за формулами: $Q = \sum_{i=1}^{n} y_i - \bar{y}^{\,2}$;

$Q_p = \sum_{i=1}^{n} y_{i,\text{теор}} - \bar{y}^{\,2}$; $Q_o = \sum_{i=1}^{n} y_{i,\text{теор}} - y_i^{\,2}$, а середнє значення за формулою $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

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

$$F = \frac{Q_p \;\; n-l}{Q_o \;\; l-1}, \qquad (77)$$

де n – кількість наглядів,

l – кількість груп у кореляційній таблиці або кількість параметрів моделі у випадку незгрупованих даних.

Розраховане значення F-статистики порівнюється із критичним значенням $F_{\text{кр}}$ розподілу Фішера, яке можна знайти за статистичними таблицями або за допомогою вбудованої функції Excel *FРАСПОБР* α, k_1, k_2, де $k_1 = l-1$; $k_2 = n-l$ – ступені волі, α – рівень значущості.

Адекватність моделі вибірковим даним можна оцінити за коефіцієнтом детермінації R^2, який показує частину варіації

значень результативної ознаки Y, що пояснюється рівнянням регресії. Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{Q_o}{Q} = \frac{Q_p}{Q}. \qquad (78)$$

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 $R^2 = 0{,}98$, то 98 % варіації результативної ознаки Y пояснюється рівнянням регресії.

Нелінійна регресія

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, наведено нижче.

Рівняння параболічної регресії:

$$\overline{y_x} = ax^2 + bx + c$$

Система для знаходження параметрів

для згрупованих вибіркових даних:

$$\begin{cases} a\sum\limits_{i=1}^{k} x_i^4 n_i + b\sum\limits_{i=1}^{k} x_i^3 n_i + c\sum\limits_{i=1}^{k} x_i^2 n_i = \sum\limits_{i=1}^{k} x_i^2 n_i \overline{y_{x_i}} \\ a\sum\limits_{i=1}^{k} x_i^3 n_i + b\sum\limits_{i=1}^{k} x_i^2 n_i + c\sum\limits_{i=1}^{k} x_i n_i = \sum\limits_{i=1}^{k} x_i n_i \overline{y_{x_i}} \\ a\sum\limits_{i=1}^{k} x_i^2 n_i + b\sum\limits_{i=1}^{k} x_i n_i + c\sum\limits_{i=1}^{k} n_i = \sum\limits_{i=1}^{k} n_i \overline{y_{x_i}} \end{cases}$$

для незгрупованих вибіркових даних:

$$\begin{cases} a\sum\limits_{i=1}^{n} x_i^4 + b\sum\limits_{i=1}^{n} x_i^3 + c\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i^2 y_i \\ a\sum\limits_{i=1}^{n} x_i^3 + b\sum\limits_{i=1}^{n} x_i^2 + c\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i \\ a\sum\limits_{i=1}^{n} x_i^2 + b\sum\limits_{i=1}^{n} x_i + cn = \sum\limits_{i=1}^{n} y_i \end{cases}$$

Рівняння гіперболічної регресії: $\overline{y_x} = \dfrac{a}{x} + b$

Продовж. рівняння регресії

Система для знаходження параметрів	
для згрупованих вибіркових даних:	для незгрупованих вибіркових даних:
$\begin{cases} a\sum\limits_{i=1}^{k}\frac{1}{x_i^2}n_i + b\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i = \sum\limits_{i=1}^{k}\frac{1}{x_i}\overline{y_{x_i}}n_i \\ a\sum\limits_{i=1}^{k}\frac{1}{x_i}n_i + b\sum\limits_{i=1}^{k}n_i = \sum\limits_{i=1}^{k}\overline{y_{x_i}}n_i \end{cases}$	$\begin{cases} a\sum\limits_{i=1}^{n}\frac{1}{x_i^2} + b\sum\limits_{i=1}^{n}\frac{1}{x_i} = \sum\limits_{i=1}^{n}\frac{1}{x_i}y_i \\ a\sum\limits_{i=1}^{n}\frac{1}{x_i} + bn = \sum\limits_{i=1}^{n}y_i \end{cases}$
Рівняння показникової регресії: $\overline{y_x} = ba^x$	
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 – $l = 2$.

Регресія в 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) викликати ***Сервис – Анализ данных – Регрессия – ОК***. З'явиться вікно для надання вхідних даних (рис. 10);

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

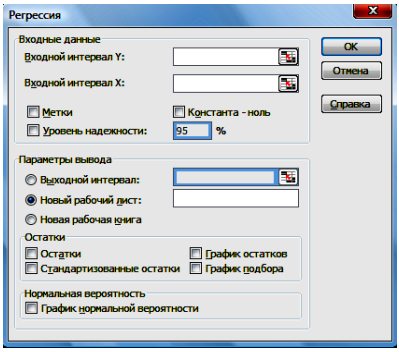


Рисунок 10 – Діалогове вікно функції ***Регрессия***

Приклад 5 і результати роботи функції ***Регрессия*** надано на рис. 11.

На рис. 11 у графі ***Коэффициенты*** вказані значення параметрів моделі a та b: b – у графі **Y-пересечение**, a – у 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.

Коефіцієнт детермінації моделі R^2 надається у графі ***R*-квадрат,** R^2=0,97.

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	

Рисунок 11 – Результати регресійного аналізу (приклад 5)

Крім того, може бути надано: графік підбору – порівняльна діаграма, що містить емпіричну й теоретичну лінії регресії; таблиця залишків – різниць емпіричних і теоретичних значень *Y* (рис. 12).

	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

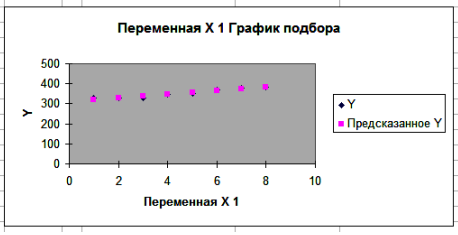


Рисунок 12 – Додаткові результати регресійного аналізу (приклад 5)

5. Дискримінантний аналіз

Дискримінантний аналіз застосовується *для розділення або класифікації об'єктів*. Це виконується за допомогою аналізу кількісних характеристик і врахування дискримінантної функції, з якою пов'язано прийняття рішення щодо проведення класифікації.

Крім факторів конкурентноздатності, у багатьох ситуаціях (під час проведення бенчмаркінгу, дослідження ринків збуту, видів товарів, джерел ресурсів тощо) необхідно правильно розподілити об'єкти, що вивчаються, на окремі групи (категорії, класи) відповідно до ключових стратегічних пріоритетів. Оскільки основною метою стратегічного аналізу є глибоке дослі-

дження саме стратегічних, а не поточних, проблем, необхідно правильно окреслювати сферу першочергової уваги й не витрачати час на вирішення другорядних завдань.

Діагностика та прогнозування зовнішніх та внутрішніх процесів підприємства передбачає застосування різних методів одно- і багатовимірного групування об'єктів та виділення серед них тих груп (категорій, класів), які є стратегічно важливими. Серед цих методів можна назвати дискримінантний, кластерний, факторний, компонентний та регресійний аналіз.

Розглянемо спочатку техніку дискримінантного аналізу на прикладі розподілу конкурентів умовної компанії «Сонячний берег» на окремі групи з урахуванням стратегічних пріоритетів. Необхідно зауважити, що дискримінантний аналіз можна застосовувати лише тоді, коли є певне уявлення про характер тих груп об'єктів, які планується виділити з генеральної сукупності.

Наприклад, якщо аналітик повинен згрупувати конкурентів, він повинен знати, на які групи їх можна поділити. У разі, коли стратегічним пріоритетом компанії є боротьба за лідерство в галузі, то найпростішим варіантом поділу конкурентів буде виділення групи лідерів і групи інших фірм, які не претендують на лідерство.

Деталізацією цього групування може бути виділення окремих груп «челенджерів» (послідовників) і фірм-новачків.

Прилад 6. Припустимо, що в галузі, яка досліджується, працює 10 фірм, які можуть бути визначені як конкуренти фірми «Сонячний берег». Критичні процеси, які було визначено у процесі бенчмаркінгу, характеризуються рядом факторів: 1) обсяг продажу; 2) рентабельність продажу; 3) співвідношення «якість-ціна» (формується на основі експертних оцінок за шкалою від 0 до 10 балів). Дані, які будуть використані для розрахунків, наведені у табл. 10.

Таблиця 10 – Показники конкурентів компанії «Сонячний берег»

Фірми-конкуренти	Обсяг продажу, млн грн	Рентабельність, %	«Якість-ціна», балів
1	48	6	5,0
2	70	9	8,0
3	76	8	5,0

Продовж. табл. 10

Фірми-конкуренти	Обсяг продажу, млн грн	Рентабельність, %	«Якість-ціна», балів
4	33	13	6,0
5	53	7	4,0
6	24	12	8,0
7	40	11	9,0
8	79	10	5,0
9	39	8	7,0
10	68	6	8,0
Середнє значення	53	9	6,5

Необхідно визначити правило або, точніше, кількісну межу, яка б відділяла підприємства групи лідерів від усіх інших підприємств.

Для того, щоб розпочати процедуру визначення цієї межі, яка називається дискримінантною лінією, потрібно отримати дві вибірки (відповідно до кількості груп, які ми плануємо сформувати), причому у першу мають увійти представники лідерів, а у другу – представники інших фірм.

Кількість об'єктів у кожній із двох вибірок, які називаються зразковими, має бути не меншою, ніж кількість ознак об'єктів, обраних для дискримінантного аналізу. Отже, у нашому прикладі в обидві вибірки необхідно включити не менше, ніж по три підприємства.

Припустимо, що ми вибрали по три підприємства, ураховуючи їх ринкові позиції (кількісним критерієм у нашому прикладі може бути обсяг продажу):

1-ша вибірка: підприємства з номерами 2, 3, 8 (найбільші обсяги продажу);

2-га вибірка: підприємства з номерами 4, 6, 9 (найменші обсяги продажу).

Очевидно, що в основі формування зразкових вибірок мають бути евристичні прийоми оцінки об'єктів, ураховуючи їх складність і багатовимірність.

На першому етапі дискримінантного аналізу визначаються середні значення ознак кожного підприємства спочатку за пер-

шою, а потім за другою вибіркою. При цьому застосовуються формули:

$$\overline{X_i}^{(1)} = \sum_{j=1}^{n_1} X_{ij}^{(1)} / n_1 \qquad \overline{X_i}^{(2)} = \sum_{j=1}^{n_2} X_{ij}^{(2)} / n_2 \tag{79}$$

де $X_i^{(1)}$, $X_i^{(2)}$- елементи векторів-стовбців X_1 і X_2 при i=1,j,m;

Після цього потрібно визначити коваріаційні матриці *S*1 і *S*2 розміром ($m \times m$) для першої і другої зразкових вибірок відповідно.

Матриця сумарної внутрішньовибіркової дисперсії *S*’ може бути використана лише в оберненому вигляді; тобто наступним етапом розрахунків є знаходження оберненої матриці *S*’ – 1.

Тепер, маючи значення оберненої матриці, можна розраховувати значення вектора дискримінантних множників і побудувати дискримінаційну функцію.

За цією функцією визначаються відповідні значення *Z* для кожної з 10-ти фірм, які ми намагаємося згрупувати (табл. 11). На основі цих показників розраховується межа дискримінації, причому можуть застосовуватися різні способи.

Найпростішим із них є визначення середніх значень дискримінаційних функцій для першої і другої виборок окремо, а потім знаходження середнього арифметичного цих значень.

Таблиця 11 – Значення дискримінантної функції *Z* для фірм-конкурентів

Фірми-конкуренти	Дискримінантна функція *Z*
1	69,4541
2	103,7341
3	99,1919
4	68,4882
5	72,7951
6	64,2531
7	80,9785
8	105,0411
9	69,6929
10	97,3490

Ураховуючи те, що до першої вибірки було обрано 2-й, 3-й і 8-й об'єкти, а до другої – 4-й, 6-й і 9-й, можна визначити межу дискримінації. Вона дорівнюватиме 86. Отже, отримано кількісний критерій для поділу підприємств-конкурентів на дві групи.

Оскільки в першу зразкову вибірку включали великі підприємства, а у другу – усі інші, то до першої групи – групи лідерів – належатимуть підприємства, у яких значення дискримінантної функції перевищують межу дискримінації (табл. 11). Номера цих підприємств: 2, 3, 8, 10.

Відповідно, підприємства з номерами 1, 4, 5, 6, 7, 9 увійдуть до другої групи.

Цінність дискримінантного аналізу для стратегічного аналітика полягає в тому, що з його допомогою можна зосередитися на дослідженні тих об'єктів, які є дійсно стратегічно важливими. Якщо підприємство намагається увійти до групи лідерів у своїй галузі, необхідно точно знати, хто з конкурентів реально може з нами змагатися і які необхідно запланувати заходи для боротьби з ними.

6. Кластерний аналіз

Кластерний аналіз – це метод багатомірного статистичного дослідження, до якого належать збір даних, що містять інформацію про вибіркові об'єкти, та упорядкування їх в порівняно однорідні, схожі між собою групи. Отже, сутність кластерного аналізу полягає у здійсненні класифікації об'єктів дослідження за допомогою численних обчислювальних процедур. У результаті цього утворюються «кластери» або групи дуже схожих об'єктів. На відміну від інших методів, цей вид аналізу дає можливість класифікувати об'єкти не за однією ознакою, а за декількома одночасно. Для цього вводяться відповідні показники, що характеризують певну міру близькості за всіма класифікаційними параметрами.

Мета кластерного аналізу полягає в пошуку наявних структур, що виражається в утворенні груп схожих між собою об'єктів – кластерів. Водночас його дія полягає й у привнесенні структури в досліджувані об'єкти. Це означає, що методи кластеризації необхідні для виявлення структури в даних, яку

нелегко знайти за візуального обстеження або за допомогою експертів.

Основними завданнями кластерного аналізу є:

– розробка типології або класифікації досліджуваних об'єктів;

– дослідження та визначення прийнятних концептуальних схем групування об'єктів;

– висунення гіпотез на підставі результатів дослідження даних;

– перевірка гіпотез, чи справді типи (групи), які були виділені певним чином, мають місце в наявних даних.

Кластерний аналіз потребує здійснення таких послідовних кроків:

1) проведення вибірки об'єктів для кластеризації;

2) визначення множини ознак, за якими будуть оцінюватися відібрані об'єкти;

3) оцінка міри подібності об'єктів;

4) застосування кластерного аналізу для створення груп подібних об'єктів;

5) перевірка достовірності результатів кластерного рішення.

Кожен із цих кроків відіграє значну роль у практичному здійсненні аналізу.

Визначення множини ознак, які покладаються в основу оцінки об'єктів ($x_1, x_2, x_3, \ldots, x_n$), у кластерному аналізі є одним із найважливіших завдань дослідження. Мета цього кроку повинна полягати у визначенні сукупності змінних ознак, яка найкраще відображає поняття подібності. Ці ознаки мають вибиратися з урахуванням теоретичних положень, покладених в основу класифікації, а також мети дослідження.

Під час визначення міри подібності об'єктів кластерного аналізу використовуються чотири види коефіцієнтів: коефіцієнти кореляції, показники віддалей, коефіцієнти асоціативності та ймовірносні, коефіцієнти подібності. Кожен із цих показників має свої переваги й недоліки, які попередньо потрібно врахувати. На практиці найбільшого розповсюдження у сфері соціальних та економічних наук здобули коефіцієнти кореляції та віддалей.

У результаті аналізу сукупності вхідних даних створюються однорідні групи у такий спосіб, що об'єкти всередині цих груп подібні між собою за деяким критерієм, а об'єкти з різних груп відрізняються один від одного.

Кластеризація може здійснюватися двома основними способами, зокрема за допомогою ієрархічних чи ітераційних процедур.

Ієрархічні процедури – послідовні дії щодо формування кластерів різного рангу, підпорядкованих між собою за чітко встановленою ієрархією. Найчастіше ієрархічні процедури здійснюються шляхом агломеративних (об'єднувальних) дій. Вони передбачають такі операції:

– послідовне об'єднання подібних об'єктів з утворенням матриці подібності об'єктів;

– побудова дендрограми (деревоподібної діаграми), яка відображає послідовне об'єднання об'єктів у кластери;

– формування з досліджуваної сукупності окремих кластерів на першому початковому етапі аналізу та об'єднання всіх об'єктів в одну велику групу на завершальному етапі аналізу.

Ітераційні процедури полягають в утворенні з первинних даних однорівневих (одного рангу) ієрархічно не підпорядкованих між собою кластерів.

Одним із найбільш поширених способів проведення ітераційних процедур є метод *k*-середніх (розроблений у 1967 р. Дж. МакКуіном). Застосування його потребує здійснення таких кроків:

– розділення вихідних даних досліджуваної сукупності на задану кількість кластерів;

– обчислення багатовимірних середніх (центрів тяжіння) виділених кластерів;

– розрахунку Евклідової відстані кожної одиниці сукупності до визначених центрів тяжіння кластерів та побудова матриці відстаней, яка ґрунтується на метриці відстаней. Використовують різні метрики відстаней, наприклад: Евклідова відстань (проста та зважена), Манхеттенська, Чебишева, Мінковського, Махалонобіса D^2 тощо;

– визначення нових центів тяжіння та нових кластерів.

Найбільш відомими та широко застосовуваними методами формування кластерів є: одиничного зв'язку; повного зв'язку; середнього зв'язку; метод Уорда.

Метод одиничного зв'язку (метод близького сусіда) передбачає приєднання одиниці сукупності до кластера, якщо вона близька (знаходиться на одному рівні схожості) хоча б до одного представника цього кластера.

Метод повного зв'язку (далекого сусіда) вимагає певного рівня подібності об'єкта (не менше граничного рівня), що передбачається включити у кластер, з будь-яким іншим.

Метод середнього зв'язку ґрунтується на використанні середньої відстані між кандидатом на включення у кластер і представниками наявного кластера.

Згідно з методом Уорда приєднання об'єктів до кластерів здійснюється у випадку мінімального приросту внутрішньогрупової суми квадратів відхилень. Завдяки цьому утворюються кластери приблизно одного розміру, які мають форму гіперсфер.

Оптимальною прийнято вважати кількість кластерів, яка визначається як різниця кількості спостережень і кількості кроків, після якої відстань об'єднання збільшується стрибкоподібно.

Кластерний аналіз, як і інші методи вивчення стохастичного зв'язку, вимагає численних складних розрахунків, які краще здійснювати за допомогою сучасних інформаційних систем, зокрема з використанням програмного продукту Statistica 6.0.

Отже, кластерний аналіз, за оцінкою науковців, має велике значення у проведенні аналітичних досліджень завдяки можливості перетворити великий обсяг різнобічної інформації в упорядкований, компактний вигляд. Це сприяє підвищенню рівня наочності, зрозумілості та сприйняття результатів аналізу, а також створює підґрунтя для прогнозування.

7. Факторний аналіз

Факторний аналіз концептуально тісно пов'язаний із методом головних компонент і використовується для вивчення співвідношення між випадковими змінними, зумовленими загальними причинами або факторами, а також із метою кількісного виразу цих співвідношень.

Усі явища та процеси господарської діяльності підприємств знаходяться у взаємозв'язку, взаємозалежності та взаємообумовленості. Деякі з них безпосередньо пов'язані між собою, а інші – опосередковано. Кожний результативний показник залежить від численних і різноманітних факторів. Звідси, важливим методологічним питанням економічного аналізу є вивчення й вимірювання впливу факторів на величину досліджуваних економічних показників.

Під факторним аналізом розуміють методику комплексного та системного вивчення і вимірювання впливу факторів на величину результативних показників.

Розрізняють такі види факторного аналізу:

– детермінований (функціональний) і стохастичний (кореляційний);
– прямий (дедуктивний) і зворотний (індуктивний);
– одноступеневий і багатоступеневий;
– статичний і динамічний;
– ретроспективний і перспективний (прогнозний).

Детермінований факторний аналіз – це методика дослідження впливу факторів, зв'язок яких із результативним показником має функціональний характер, тобто результативний показник може бути представлений у вигляді алгебраїчної суми, добутку або частки показників, що є факторами детермінованої моделі.

Основні властивості детермінованого факторного аналізу:

– визначення детермінованої моделі шляхом логічного аналізу;
– наявність повного зв'язку між показниками;
– неможливість розподілити результати впливу одночасно діючих факторів, які не підлягають об'єднанню в одну модель;
– вивчення взаємозв'язків у короткостроковому періоді.

Стохастичний аналіз – це методика дослідження факторів, зв'язок яких із результативним показником, на відміну від функціонального, є неповним, імовірним, кореляційним. Якщо за функціональної залежності зі зміною аргументу завжди відповідно змінюється функція, то за кореляційного зв'язку зміна аргументу може дати декілька значень приросту функції залежно від поєднання інших факторів, що визначають цей показник.

Наприклад, немає можливості функціонально показати зв'язок між рентабельністю роботи підприємства та середнім рівнем освіти управлінського персоналу або між курсом національної валюти й рівнем інфляції у країні. Для проведення стохастичного факторного аналізу використовуються спеціальні прийоми і способи, у тому числі й економіко-математичні.

Прямий (дедуктивний) факторний аналіз передбачає дослідження дедуктивним способом – від загального до часткового. При зворотному (індуктивному) факторному аналізі вивчення причиново-наслідкових зв'язків виконується способом логічної індукції – від часткових, окремих факторів до узагальнюючих.

Одноступеневий факторний аналіз використовується для дослідження факторів тільки одного рівня підпорядкування без їх деталізації на складові частини. За багатоступеневого факторного аналізу проводиться деталізація факторів на складові елементи з метою вивчення їх поведінки.

Статичний факторний аналіз використовується під час дослідження впливу факторів на результативні показники на відповідну дату, а динамічний факторний аналіз представляє собою методику дослідження причинно-наслідкових зв'язків у динаміці.

Ретроспективний факторний аналіз – це системне комплексне дослідження результатів господарської діяльності, яке проводиться за даними певних минулих аналітичних періодів. Перспективний факторний аналіз дозволяє дослідити поведінку факторів і результативних показників у майбутньому.

Основні етапи проведення факторного аналізу:

1. Постановка мети аналізу, вибір факторів, які здійснюють вплив на досліджувані результативні показники.

2. Класифікація і систематизація факторів із метою забезпечення можливостей системного підходу.

3. Визначення форми залежності між факторами й результативним показником.

4. Моделювання взаємозв'язків між результативним та факторними показниками, побудова економічно обґрунтованої (з позиції факторного аналізу) факторної моделі.

5. Розрахунок впливу факторів та оцінка ролі кожного з них у зміні величини результативного показника. Проводиться вибір прийому факторного аналізу й підготовка умов для його виконання, реалізація розрахункових процедур.

6. Формулювання висновків за результатами проведених досліджень, підготовка відповідних управлінських рішень.

8. Аналіз часових рядів

Часовий ряд (time series) – це сукупність вимірювань деякої змінної величини, які проводяться у часі. Характерною особливістю часових рядів є те, що спостереження за деяким об'єктом проводяться послідовно в часі. Наприклад, температура повітря в середині кожної години доби, щорічна врожайність зернових, щоденний об'єм продажів якого-небудь товару, вартість акції підприємства, рівень інфляції, обмінний курс валют – усе це часові ряди.

Для аналізу часового ряду порядок у послідовності є суттєвим, тобто час виступає одним із визначальних чинників. Це відрізняє часовий ряд від звичайної випадкової вибірки, де індекси вводять лише для зручності ідентифікації. Принциповою відмінністю часового ряду від простих статистичних сукупностей є:

по-перше, рівні часового ряду не є незалежними. Інакше кажучи, якщо майбутні значення змінної можна визначити, то вони є функцією від минулих значень цієї змінної;

по-друге, рівні часового ряду неоднаково розподілені. Закон розподілу ймовірностей цих випадкових величин і, зокрема, їхні математичні сподівання та дисперсії можуть залежати від часу.

Незалежно від природи кожного часового ряду можна виділити такі основні типи задач, які звичайно вирішують під час проведення аналізу початкових даних. На першому етапі намагаються побудувати просту математичну систему або модель, яка описує поведінку часового ряду в короткій формі.

Потім виконується спроба пояснити його поведінку за допомогою інших змінних і з'ясувати ступінь зв'язку як між спостереженнями одного ряду, так і між різними рядами. Виходячи з цілей дослідження, кожний часовий ряд звичайно розглядають як суміш таких компонент:

– тренд або довгострокова тенденція в розвитку ряду;

– сезонна компонента або, іншими словами, деякий ефект у динаміці ряду, який повторюється через певний період.

Розподіл динаміки часового ряду на вищезгадані компоненти визначає і групи математичних методів, які використовуються для аналізу відповідної компоненти. Так для виявлення й аналізу тренда використовують апарат регресійного аналізу (regression analysis) й ковзних середніх. Для аналізу сезонного ефекту застосовують спеціальні моделі сезонного згладжування і сезонної авторегресії. Існує навіть спеціальний клас моделей, призначений для побудови і прогнозування (prediction) наслідків інтервенцій. Коливання щодо тренда виявляються за допомогою спектрального аналізу, а для опису і прогнозування таких процесів використовують гармонійні моделі або моделі авторегресії – наприклад, метод ковзного середнього.

У задачах прогнозування часові ряди використовуються за наявності значної кількості реальних значень даного показника з минулого і за умови, що тенденція, яка намітилася у минулому, чітка і відносно стабільна. При цьому неявно передбачається, що минуле є хорошим провідником у майбутнє. Аналіз часових рядів дозволяє зумовити, що повинно відбутися за відсутності втручання ззовні, і, значить, не може передбачити зміни тенденції. Тим самим, подібним аналізом переважно користуються при складанні короткострокових прогнозів.

Криві тренда згладжують динамічний ряд значень показника, виділяючи загальну тенденцію. Саме вибір кривої тренда, що сам по собі є досить важкою задачею, багато в чому визначає результати прогнозування. У більшості випадків динамічний ряд, окрім тренда і випадкових відхилень від нього, характеризується ще сезонними і циклічними складовими. Циклічні складові відрізняються від сезонних більшою тривалістю і непостійністю амплітуди. Звичайна тривалість сезонної компоненти вимірюється днями, тижнями або місяцями, а циклічної – роками або десятками років. До основних методів аналізу часових рядів можна віднести метод ковзного середнього, експоненціального згладжування і проектування тренда. Для розгляду, як працюють ці методи, користуватимемося одним і тим же часовим рядом.

Приклад 7. Припустимо, що об'єми продажів деякого товару описуються в перебігу тижня часовим рядом:

День тижня	**Кількість проданої продукції**
Понеділок	10
Вівторок	6
Середа	5
Четвер	11
П'ятниця	9
Субота	8
Неділя	7

або дещо по-іншому

x	1	2	3	4	5	6	7
t	10	6	5	11	9	8	7

Розв'яжемо приклад 7 декількома методами.

Метод рухомого (ковзного) середнього

Цей метод розділяють на два типи: метод рухомого (ковзного) середнього й метод зваженого (ковзного) середнього.

Метод рухомого (ковзного) середнього.

Цей метод полягає в тому, що розрахунок показника на прогнозований момент часу будується шляхом усереднювання значень цього показника за декілька попередніх днів. Припустимо, що у нас є дані показника тільки за перші три дні. Обчислимо прогнозоване число дефектів на четвертий день тижня (четвер). Для цього визначимо середнє значення числа дефектів за попередні три дні – понеділок, вівторок і середу – і знайдемо їх середнє арифметичне: $f_4 = \frac{10+6+5}{3} = \frac{21}{3} = 7$.

Прогнозований об'єм продажів на п'ятницю обчислюється аналогічним чином за реальними показниками за три попередні дні – вівторок, середу й четвер: $f_5 = \frac{6+5+11}{3} = \frac{22}{3} \approx 7.33$.

Подібним способом розраховуються прогнози на суботу, неділю і черговий понеділок:

$$f_6 = \frac{5+11+9}{3} = \frac{25}{3} \approx 8.33$$

$$f_7 = \frac{11+9+8}{3} = \frac{28}{3} \approx 9.33\,.$$

$$f_6 = \frac{9+8+7}{3} = \frac{24}{3} \approx 8$$

У результаті можна отримати табл. 12.

Таблиця 12 – Розрахункова таблиця

x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t	10	6	5	11	9	8	7	–
f	–	–	–	7	7,33	8,33	9,33	8

Порівняльні результати наведені на рис. 13: темними точками відзначені реальні значення, а світлими – прогнозовані.

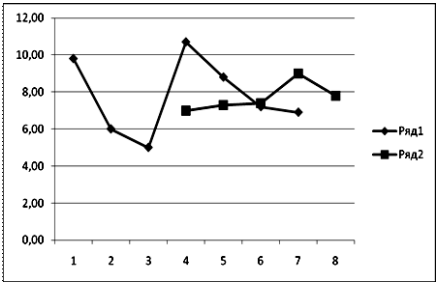


Рисунок 13 – Порівняльні результати прогнозу методом ковзного середнього

Для загального випадку розрахункова формула виглядає так:

$$f_k = \frac{x_{k-N} + x_{k-N+1} + \ldots + x_{k-1}}{N} \qquad (80)$$

або

$$f_k = \frac{1}{N}\sum_{i=1}^{N} x_{k-i}\,, \qquad (81)$$

де x_{k-i} – реальне значення показника в момент часу *tk-i*;

N – число попередніх моментів часу;

fk – прогноз на момент часу *tk*.

Метод зваженого (ковзного) середнього

При прогнозуванні методом усереднювання часто доводиться спостерігати, що ступінь впливу використаних при розрахунку реальних показників виявляється неоднаковим, при цьому звичайно більш «свіжі» дані мають більшу вагу [7, 8]. Математично метод зваженого рухомого середнього можна записати як

$$f_k = \frac{\sum_{i=1}^{N} w_{k-i} x_{k-i}}{\sum_{i=1}^{N} w_{k-i}}, \qquad (82)$$

де x_{k-i} – реальне значення показника в момент часу;

N – кількість попередніх моментів часу, що використовуються при розрахунку;

fk – прогноз на момент часу *tk*;

wi – вага, з якою використовується показник x_{k-i} під час розрахунку.

Для розрахунків звернемося до початкового часового ряду (прикладу 7), уважаючи, що під час складання прогнозу на завтрашній день об'єм сьогоднішніх продажів береться з вагою 60, вчорашніх – із вагою 30, а позавчорашніх – із вагою 10. Маємо:

$$f_4 = \frac{10 \cdot 10 + 30 \cdot 6 + 60 \cdot 5}{10 + 30 + 60} = \frac{580}{100} \approx 5.80;$$

$$f_5 = \frac{10 \cdot 6 + 30 \cdot 5 + 60 \cdot 11}{100} = \frac{870}{100} \approx 8.70;$$

$$f_6 = \frac{10 \cdot 5 + 30 \cdot 11 + 60 \cdot 9}{100} = \frac{920}{100} \approx 9.20;$$

$$f_7 = \frac{10 \cdot 11 + 30 \cdot 9 + 60 \cdot 8}{100} = \frac{860}{100} \approx 8.60;$$

$$f_8 = \frac{10 \cdot 9 + 30 \cdot 8 + 60 \cdot 7}{100} = \frac{750}{100} \approx 7.50.$$

Результати розрахунків наведені в таблиці13:

Таблиця 13 – Результативна таблиця

x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t	10	6	5	11	9	8	7	–
f	–	–	–	5,8	8,7	9,2	8,6	7,5

На рис. 14 темними точками відзначені реальні значення, а світлими – прогнозовані.

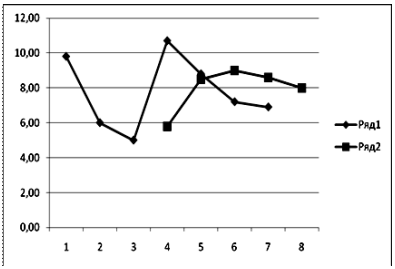


Рисунок 14 – Порівняльні результати прогнозу методом зваженого ковзного середнього

Метод експоненціального згладжування

Експоненціальне згладжування – це дуже популярний метод прогнозування багатьох часових рядів. Історично метод був незалежно відкритий Броуном і Холтом. Під час розрахунку прогнозу методом експоненціального згладжування (exponential smoothing) враховується відхилення попереднього прогнозу від реального показника, а сам розрахунок проводиться за такою формулою:

$$f_k = f_{k-1} + \alpha(x_{k-1} - f_{k-1}), \qquad (83)$$

де α – стала згладжування ($0 < \alpha < 1$).

Коли ця формула застосовується рекурсивно, то кожне нове згладжене значення (яке є також прогнозом) обчислюється як зважене середнє поточного спостереження і згладженого ряду. Рекомендується як *f0* брати початкове значення, що дає якнай-

кращий прогноз. Очевидно, результат згладжування залежить від параметра α. Якщо α дорівнює 1, то попередні спостереження повністю ігноруються. Якщо α дорівнює 0, то ігноруються поточні спостереження. Значення α між 0 і 1 дають проміжні результати. З формули (83) виходить, що α має потрапляти в інтервал між 0 і 1. На практиці звичайно рекомендується брати α менше 0,3. Крім того, параметр згладжування α часто знаходиться пошуком на сітці. Можливі значення параметра α розбиваються сіткою з певним кроком.

Слід мати на увазі, що під час розв'язання реальної задачі прогнозування часовий ряд складається в декілька етапів, і реальне значення показника на момент часу, що розраховується, зазвичай, наперед невідоме. Проте, перш ніж заглянути в майбутнє за допомогою одного з вищезазначених методів, звичайно проводяться розрахунки з повним часовим рядом, що описує деякий проміжок часу в минулому. Це робиться для того, щоб підібрати відповідне значення *f0* і порівняти результати прогнозу з реальними даними (метод простого ковзного середнього), підібрати відповідні значення *f0* та ваги й порівняти результати прогнозу з реальними даними (метод зваженого ковзного середнього).

Метод проектування тренда

Основною ідеєю методу проектування тренда є побудова прямої, яка «в середньому» якнайменше відхиляється від масиву точок (t, x) заданого часового ряду, що описується рівнянням:

$$x = at + b\,, \tag{84}$$

де a, b – сталі коефіцієнти.

Розрахунок коефіцієнтів a і b ведеться за методом найменших квадратів:

$$\varphi(a,b) = \sum_{i=1}^{n}(at_i + b - x_i)^2 \to \min\,,$$

$$\begin{cases} a\sum_{i=1}^{n} t_i + bn = \sum_{i=1}^{n} x_i \\ a\sum_{i=1}^{n} t_i^{\,2} + b\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} t_i x_i \end{cases} \tag{85}$$