МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ СЕМЕНА КУЗНЕЦЯ

БІЗНЕС-АНАЛІТИКА БАГАТОВИМІРНИХ ПРОЦЕСІВ

Навчальний посібник

Харків
ХНЕУ ім. С. Кузнеця
2018

УДК 303.722.3(075)
Б59

Авторський колектив: д-р екон. наук, професор Т. С. Клебанова – розділи 1 – 9; д-р екон. наук, професор Л. С. Гур'янова – розділи 6, 7, 9; канд. екон. наук, доцент Л. О. Чаговець – розділи 1, 5, 8, 9; канд. екон. наук, доцент О. В. Панасенко – розділи 3, 4, 9; канд. екон. наук, доцент О. А. Сергієнко – розділи 2, 8, 9.

Рецензенти: завідувач кафедри економічної кібернетики та маркетингового менеджменту Національного технічного університету "Харківський політехнічний інститут", д-р екон. наук, професор *В. Я. Заруба*; завідувач кафедри економічної кібернетики Запорізького національного університету, д-р екон. наук, професор *Н. К. Максишко*.

Рекомендовано до видання рішенням ученої ради Харківського національного економічного університету імені Семена Кузнеця.
Протокол № 3 від 20.11.2017 р.

Самостійне електронне текстове мережеве видання

Бізнес-аналітика багатовимірних процесів : навчальний по-
Б59 сібник [Електронний ресурс] / Т. С. Клебанова, Л. С. Гур'янова, Л. О. Чаговець та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2018. – 272 с.
ISBN 978-966-676-734-2

Розглянуто особливості застосування економіко-математичних методів і моделей для розв'язання широкого класу прикладних задач багатовимірного аналізу даних: формування вибірки для дослідження та оцінювання її якості; застосування методів кластерного та дискримінантного аналізу; розрахунок таксономічного показника рівня розвитку, використання методу "центру ваги" в економічних дослідженнях; застосування методів факторного аналізу багатовимірних процесів. Подано теоретичний матеріал і приклади, що дозволяють засвоїти зміст і методику застосування методів аналізу багатовимірних економічних процесів. Наведено завдання для самостійного вирішення, глосарій, лабораторний практикум.

Рекомендовано для студентів економічних спеціальностей і практичних спеціалістів.

УДК 303.722.3(075)

ISBN 978-966-676-734-2

Зміст

Вступ

В умовах, коли рішення приймаються на підставі стохастичної, неповної інформації, використання методів багатовимірного статистичного аналізу є необхідним. Соціально-економічні процеси і явища залежать від великої кількості параметрів, які їх характеризують. Це обумовлює труднощі, пов'язані з виявленням структури взаємозв'язків цих параметрів. Методи багатовимірного аналізу даних використовуються під час розв'язання завдань, пов'язаних з дослідженням поводження індивідуума, родини, іншої соціально-економічної чи виробничої одиниці як представника великої сукупності об'єктів. Сучасний спектр методів багатовимірного статистичного аналізу досить широкий. Так, вивчення взаємозв'язків у багатовимірних сукупностях можна здійснювати за допомогою кореляційного та регресійного аналізу. Для оцінювання тісноти зв'язку між системами показників можуть бути використовують канонічні кореляції.

Методи багатомірної класифікації призначені для розподілу сукупності об'єктів на визначені змістовні однорідні групи. Кожний з об'єктів характеризується великою кількістю різних стохастично зв'язаних ознак. Для розв'язання задач класифікації застосовують кластерний і дискримінантний аналіз.

Наявність множини вихідних ознак, що характеризують багатовимірні об'єкти, викликає необхідність визначати найбільш істотні та вивчати менший набір показників. Для відбору, угрупування змінних і рейтингування використовують методи таксономії та вибору репрезентантів. Розв'язання задач зниження розмірності простору ознак забезпечують методи факторного, компонентного аналізу, багатовимірного шкалювання. Ці методи розкривають об'єктивно існуючі закономірності, що безпосередньо не спостерігаються, за допомогою факторів, головних компонент і шкал. Стиснення інформації здійснюється за рахунок того, що кількість факторів або головних компонент значно менша, ніж кількість вихідних ознак.

Сучасні ринкові умови потребують від суб'єктів господарювання вміння аналізувати багатовимірні процеси та практично використовувати методи та моделі багатовимірного аналізу даних для прийняття ефективних управлінських рішень. У зв'язку з цим навчальна дисципліна "Бізнес-аналітика багатовимірних процесів" є однією з необхідних дисциплін економіко-математичного циклу. Метою дисципліни є вивчення теоретичних основ і можливостей практичного застосування методів багатовимірного статистичного аналізу для дослідження економічних систем різного призначення. *Завданням* навчальної дисципліни є засвоєння студентами основних понять бізнес-аналітики багатовимірних процесів, методологічних основ формування і оцінювання якості вибірки з метою проведення аналізу, оволодіння навичками використання методів і моделей аналізу багатовимірних процесів для розроблення та прийняття

управлінських рішень та інформаційних технологій і програмних засобів – для проведення необхідних розрахунків.

Навчальний посібник рекомендовано для використання під час вивчення дисциплін "Бізнес-аналітика багатовимірних процесів", "Багатомірний аналіз даних", "Методи економіко-статистичних досліджень", "Управління проектами інформатизації", "Математичні методи, моделі та інформаційні технології у наукових дослідженнях", "Методи та моделі дослідження економічних процесів та управління проектами в туризмі", "Методи та моделі дослідження економічних процесів". У свою чергу, знання з цієї дисципліни забезпечують успішне засвоєння інших дисциплін економіко-математичного циклу, виконання тренінгів, міждисциплінарних комплексних курсових і магістерських дипломних робіт. Пререквізитами до вивчення зазначених дисциплін є знання та навички з навчальних дисциплін "Вища математика", "Теорія ймовірності та математична статистика", "Макроекономіка", "Мікроекономіка", "Прикладна економетрика".

Опрацювання матеріалів навчального посібника *"Бізнес-аналітика багатовимірних процесів"* передбачає формування у студентів таких *компетентностей:* здатність виділяти й аналізувати багатовимірні об'єкти в економіці; будувати алгоритми кластер-процедур; здійснювати класифікацію об'єктів на основі методу дискримінантного аналізу; здійснювати лінійне впорядкування багатовимірних об'єктів на основі методів таксономії; застосовувати метод дендритів для нелінійного впорядкування об'єктів; виділяти об'єкти-репрезентанти в однорідних групах; визначати агрегатні діагностичні ознаки; застосовувати методи дисперсійного аналізу; досліджувати взаємозалежності на основі методу канонічних кореляцій; застосовувати моделі факторного аналізу для зниження розмірності ознакового простору; використовувати методи багатовимірного шкалювання.

Навчальний посібник містить інформацію за темами: "Базові поняття бізнес-аналітики багатовимірних процесів"; "Вимірювання і типи вимірювальних шкал. Методи оцінювання вибірки"; "Особливості класифікації багатовимірних процесів"; "Методи кластерного аналізу. Класифікація без навчання"; "Методи дискримінантного аналізу. Класифікація з навчанням"; "Методи повної редукції. Таксономічний показник рівня розвитку"; "Методи неповної редукції. Метод центру ваги"; "Методи факторного аналізу".

У рамках кожної теми поданий теоретичний матеріал; демонстраційні приклади, що дозволяють засвоїти зміст і методику застосування методів і моделей аналізу багатовимірних економічних процесів; запитання для самодіагностики; тести; задачі для самостійного розв'язання; ключові слова.

Навчальний посібник призначений для студентів спеціальності 051 "Економіка", а також для студентів інших спеціальностей, які виконують дослідження, пов'язані із бізнес-аналітикою багатовимірних процесів.

Розділ 1. Базові поняття бізнес-аналітики багатовимірних процесів

1.1. Сутність багатовимірного статистичного аналізу. Можливості застосування багатовимірного статистичного аналізу в бізнес-аналітиці.

1.2. Історичні аспекти використання багатовимірного статистичного аналізу. Методи багатовимірного статистичного аналізу.

1.3. Особливості обробки багатовимірних статистичних даних. Види простору ознак. Етапи дослідження за допомогою багатовимірного статистичного аналізу.

Ключові слова: багатовимірний статистичний аналіз; методи багатовимірного аналізу; ознака; матриця; простір ознак; модель; моделювання; бізнес-аналітика.

Література: [14; 15; 23; 34; 36].

1.1. Сутність багатовимірного статистичного аналізу. Можливості застосування багатовимірного статистичного аналізу в бізнес-аналітиці

Методи багатовимірного статистичного аналізу (БСА) є інтелектуальним інструментом дослідження. Зростаючий інтерес до БСА пояснюється перш за все його можливостями у відображенні та моделюванні реальних явищ і процесів, що найчастіше мають багатоознакову природу.

Без базових знань з обробки багатовимірних даних не можуть розвиватися математика та статистика. Усі новітні розробки, присвячені проблемам нечітких множин, моделюванню катастроф, розпізнаванню образів, сценарного прогнозування тощо, припускають багатовимірне подання спостережуваних об'єктів. Основні можливості БСА наведено на рис. 1.1.

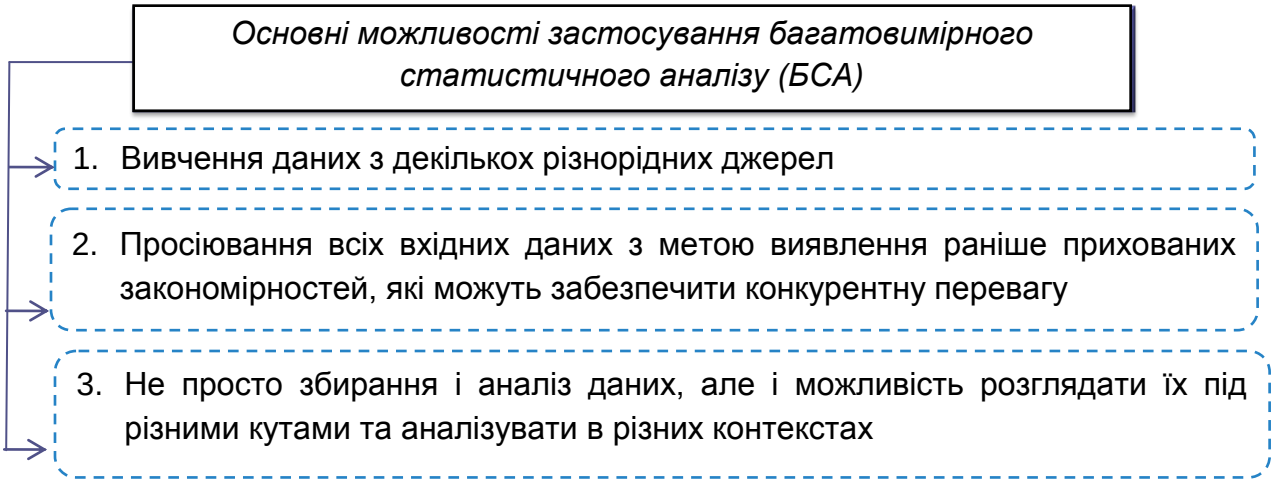


Рис. 1.1. **Основні можливості БСА**

БСА – це сукупність формалізованих статистичних методів, які базуються на поданні результативної інформації в багатовимірному геометричному просторі та дозволяють визначати приховані (латентні), але об'єктивно існуючі закономірності в організаційній структурі та тенденціях розвитку соціально-економічних явищ і процесів

Застосування багатовимірного статистичного аналізу в економіці

Промислові та торговельні підприємства:

оцінювання класу інвестиційної привабливості;

прогнозування класу банкрутства;

аналіз економічної безпеки підприємства;

позиціонування та оцінювання рівня економічного розвитку підприємства.

Роздрібна торгівля:

аналіз купівельної корзини для виявлення товарів, які покупці прагнуть купувати, розроблення стратегії створення запасів товарів і способів їх розміщення в торгових залах;

дослідження характеру потреб різних категорій клієнтів з певною поведінкою (наприклад, тих, які купують товари відомих дизайнерів або відвідують розпродажі), для розроблення заходів із просування товарів.

Банківська справа:

виявлення шахрайства з кредитними картками шляхом аналізу минулих транзакцій;

сегментація клієнтів у маркетинговій політиці, прогнозування змін клієнтури, прогнозні моделі цінності своїх клієнтів.

Страхування:

оцінювання стереотипів у заявах про виплату страхового відшкодування;

аналіз ризику шляхом виявлення поєднань факторів, пов'язаних з оплаченими заявами.

Відомий випадок, коли в США велика страхова компанія виявила, що суми, виплачені за заявами людей, які перебували у шлюбі, вдвічі перевищує суми за заявами самітних людей. Компанія відреагувала переглядом своєї загальної політики надання знижок сімейним клієнтам.

Інтернет-продажі:

розроблення ефективних підходів для очищення, обробки, агрегування і дослідження даних.

1.2. Історичні аспекти використання багатовимірного статистичного аналізу. Методи багатовимірного статистичного аналізу

Багатопараметричний опис об'єктів, явищ і процесів широко використовувався в роботах таких вчених, як:

Ч. Дарвін (60-ті рр. XIX ст., англійський натураліст) у селекції видів і для визначення чинників еволюції органічного світу

Д. І. Менделєєв (60 – 70 рр. XIX ст.) для систематизації якісних характеристик хімічних елементів

А. П. Шлікевич, І. Ф. Анненський та ін. (початок XX ст.) для класифікації земельних господарств

Історію БСА як самостійного наукового напряму розвитку статистичної теорії пов'язують з публікаціями на початку XX ст. англійських вчених К. Пірсона і У. Спірмена, присвячених основам побудови алгоритмів стиснення статистичних даних. Фундамент теорії БСА заклали такі вчені, як Л. Л. Терстоун, Л. Р. Такер, Р. Хорст, Г. Кайзер, Т.Келлі, Г. Харман, Р. Тріонон, Р. Сокал, М. Жамбю, Р. В. Хемінг, Л. Заде, Р. Фішер та ін. (рис. 1.2).

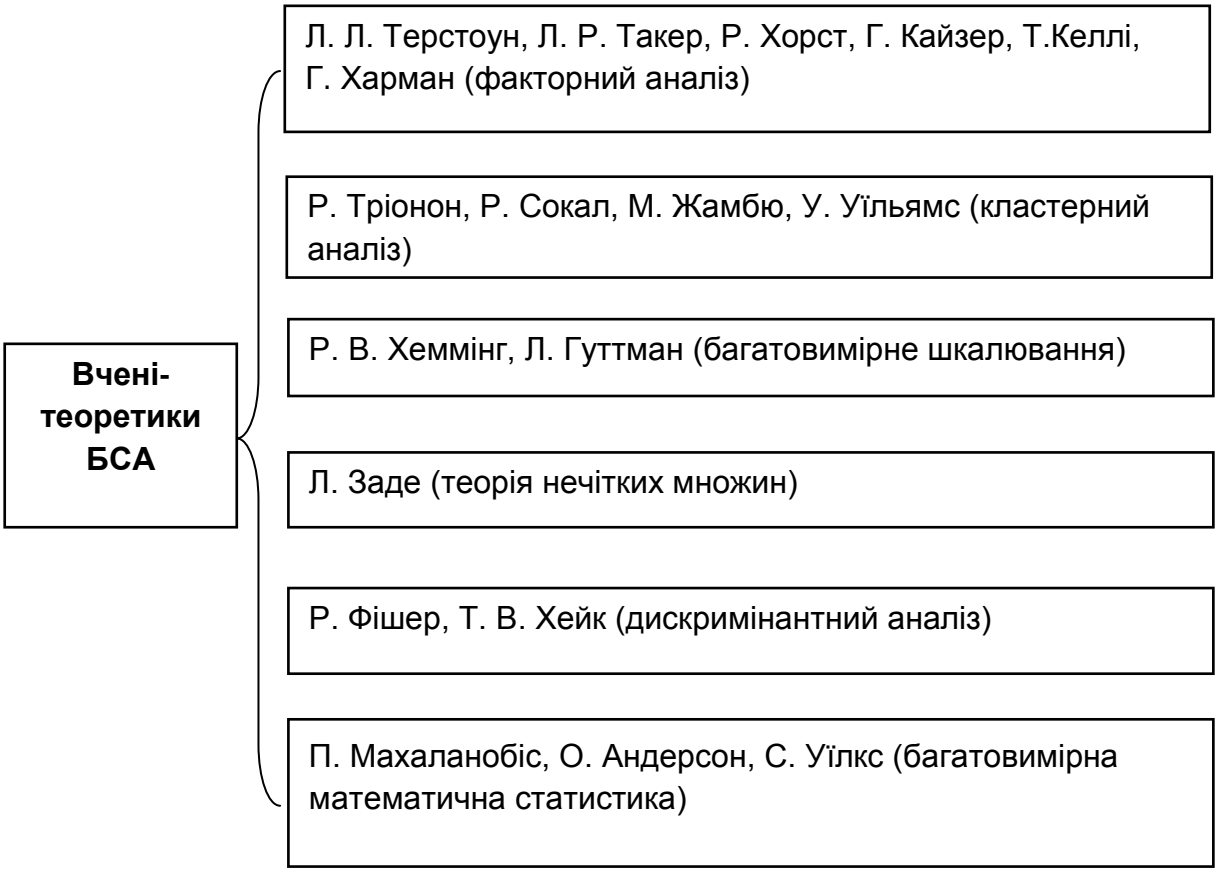


Рис. 1.2. **Вчені-теоретики, які заклали фундамент теорії БСА**

Методи багатовимірного статистичного аналізу базуються на методах статистики (теорія ймовірностей, математична статистика, загальна теорія статистики) та методах вищої математики (аналітична геометрія, матрична алгебра, багатомірний математичний аналіз) і використовують методи ймовірнісного аналізу даних і методи логіко-геометричного напряму (рис. 1.3).

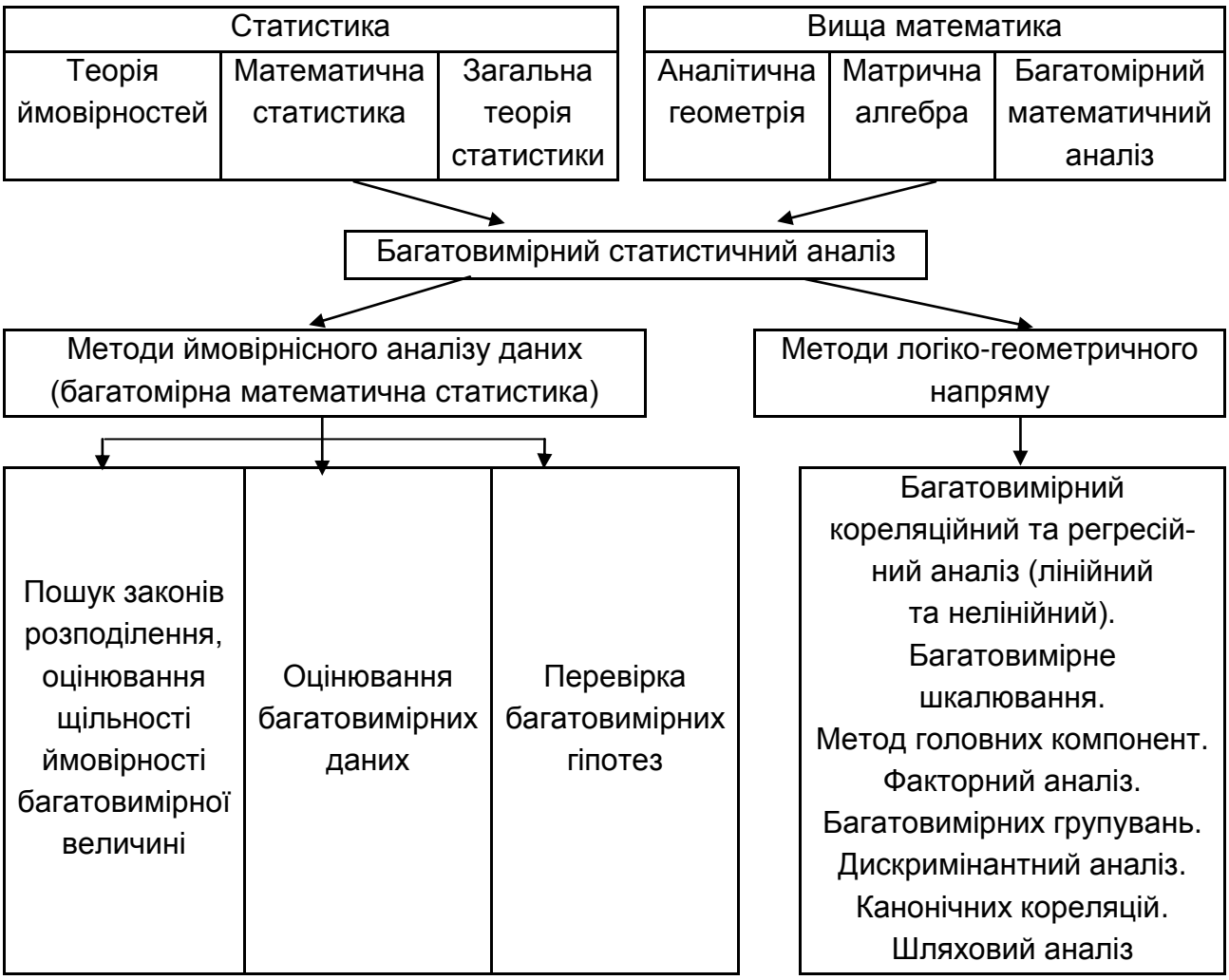


Рис. 1.3. **Методи багатовимірного статистичного аналізу**

У бізнес-аналізі багатовимірних процесів виникають різні задачі, для дослідження та розв'язання яких застосовують відповідні методи (табл. 1.1).

Таблиця 1.1

Методи та задачі багатовимірного статистичного аналізу

Методи	Типи задач	Коментарі
1	2	3
Статистичного оцінювання багатовимірної випадкової величини	Оцінка параметрів багатовимірної сукупності	Оцінка: багатовимірної середньої, матриці коваріацій, імовірнісних оцінок, робастне оцінювання

Продовження табл. 1.1

1	2	3
Перевірка багатовимірних гіпотез	Перевірка гіпотез про рівність параметрів багатовимірних сукупностей і відповідність законам розподілу	–
Множинний кореляційно-регресійний аналіз	Вимірювання і моделювання зв'язків досліджуваних ознак або об'єктів	–
Багатовимірне шкалювання	Візуалізація даних, моделювання складних систем	Подання даних у теоретичному просторі; опис процесів і явищ, які через свою складність або нестабільність не піддаються моделюванню традиційними методами
Метод головних компонент Факторний аналіз	Стиснення даних	Зведення множини елементарних ознак до малої кількості значущих "узагальнених ознак" і виявлення латентних факторів. Це ж завдання може вирішуватися відносно не тільки ознак, але й об'єктів
Багатовимірне групування (кластерний аналіз)	Групування багатовимірних об'єктів	Групування багатовимірних об'єктів
Дискримінантний аналіз	Угрупування з "навчанням"	Пошук еталонних груп, класифікація нових об'єктів за відомими еталонними групами
Канонічних кореляцій	Стиснення даних і моделювання зв'язків узагальнених ознак	Установлюється форма зв'язку наборів залежних змінних з незалежними факторними змінними, які можуть бути узагальненими ознаками

Закінчення табл. 1.1

1	2	3
Багатовимірний дисперсійний аналіз	Оцінювання та дослідження дисперсій комплексів ознак	–
Багатовимірний коваріаційний аналіз	Оцінювання залежності варіації результативної ознаки від факторної	Передбачає попередню класифікацію даних і пошук регресійних зв'язків для кожного класу. Потім обчислюються і аналізуються оцінки коваріацій (Тхх, Туу, Тух)

Отже, методи багатовимірного статистичного аналізу дозволяють розв'язувати різноманітні задачі дослідження багатовимірних процесів: групування, стиснення даних, моделювання складних систем, оцінювання параметрів багатовимірних сукупностей та ін. Це дозволяє використовувати отримані результати для прийняття ефективних управлінських рішень.

1.3. Особливості обробки багатовимірних статистичних даних. Види простору ознак. Етапи дослідження за допомогою багатовимірного статистичного аналізу

Обробка багатовимірних статистичних даних має певні особливості: застосування методів БСА вимагає творчого підходу до розв'язання аналітичних задач, оскільки оброблюються багатовимірні сукупності даних. Для методів БСА характерна глибока формалізація та складна логіко-математична конструкція, їх практичне застосування потребує використання обчислювальної техніки (рис. 1.4).

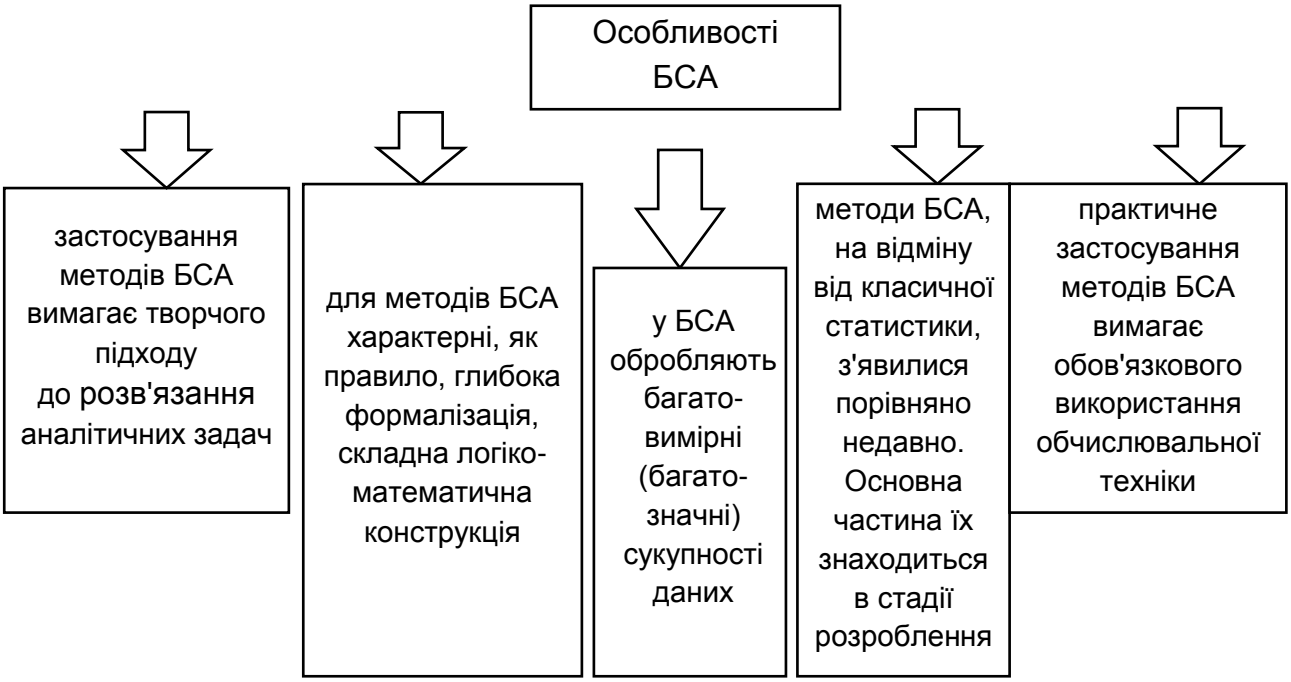


Рис. 1.4. **Особливості застосування методів БСА**

Найбільш структурованою формою подання вихідних даних про багатовимірний об'єкт є матрична.

Види матриць

1) матриця значень аналітичних ознак (x_j)

Об'єкт	X_1	X_2	X_3	X_4	…	X_m
n_1	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1m}
n_2	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2m}
n_3	x_{31}	x_{23}	x_{33}	x_{34}	…	x_{3m}
…	…	…	…	…	…	…
n_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	…	x_{nm}

2) матриця теоретичних відстаней між об'єктами (n_j)

Об'єкт	Об'єкт				
	n_1	n_2	n_3	…	n_n
n_1	c_{11}	c_{12}	c_{13}	…	c_{1n}
n_2	c_{21}	c_{22}	c_{23}	…	c_{2n}
…	…	…	…	…	…
n_n	c_{n1}	c_{n2}	c_{n3}	…	c_{nn}

Види простору ознак:

з нульовою розмірністю (об'єкти не мають характеристик);

одномірний простір ознак (об'єкти відображаються значеннями однієї ознаки);

багатовимірний простір (об'єкти подані значеннями двох і більше ознак).

Приклад одновимірного простору ознак (рис. 1.5).

Підприємство № п/п	Середньочасовий рівень вироблення одного робочого, дол. США (X)
1	6
2	4
3	9
4	7
5	3

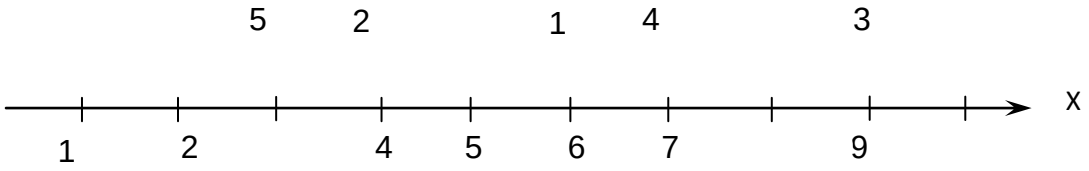


Рис. 1.5. **Одновимірний простір ознак**

Приклад двовимірного простору ознак (рис. 1.6).

Підприємство № п/п	Середньочасовий рівень вироблення одного працівника, дол. США (X_1)	Середній стаж роботи на підприємстві робочого, років (X_2)
1	6	5
2	4	7
3	9	12
4	7	14
5	3	11

Рис. 1.6. **Двовимірний простір ознак** (початок)

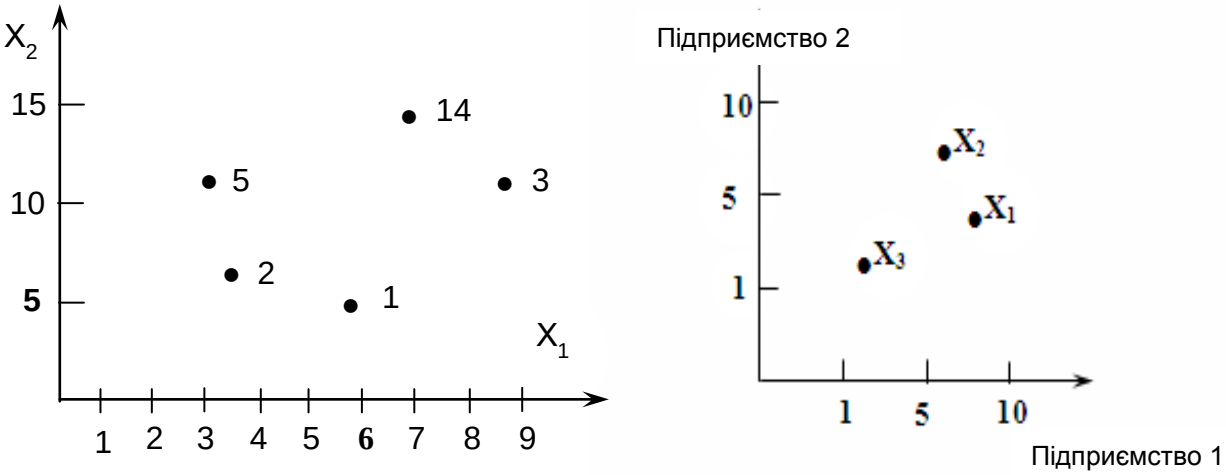


Рис. 1.6. **Двовимірний простір ознак** (закінчення)

Приклад тривимірного простору ознак (рис.1.7).

Підприємство № п/п	Середньочасовий рівень вироблення одного працівника, дол. США (X_1)	Середній стаж працівника роботи на підприємстві, років (X_2)	Середній рівень класифікації працівників за тарифним розрядом (X_3)
1	6	5	1,2
2	4	7	1,9
3	9	12	3,5
4	7	14	2,7

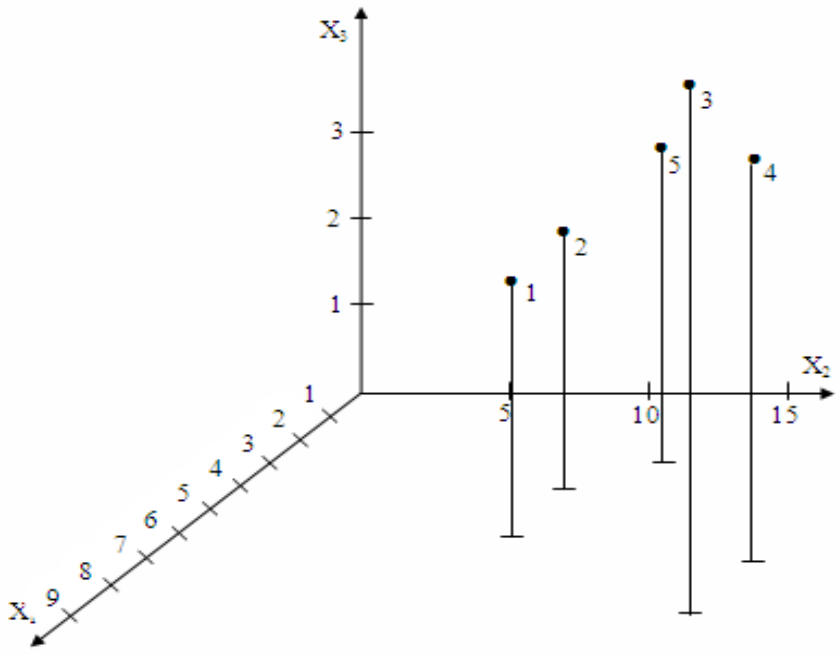


Рис. 1.7. **Тривимірний простір ознак**

Сучасні економічні умови потребують від суб'єктів господарювання вміння аналізувати багатовимірні процеси та практично використовувати результати аналізу даних для прийняття ефективних управлінських рішень. Для того щоб орган управління вибрав ту або іншу процедуру прийняття рішень (або механізм управління, тобто залежність своїх дій від цілей організації і дій суб'єктів управління – агентів), він має вміти передбачати поведінку агентів – їх реакцію на ті або інші управлінські дії (рис. 1.8).

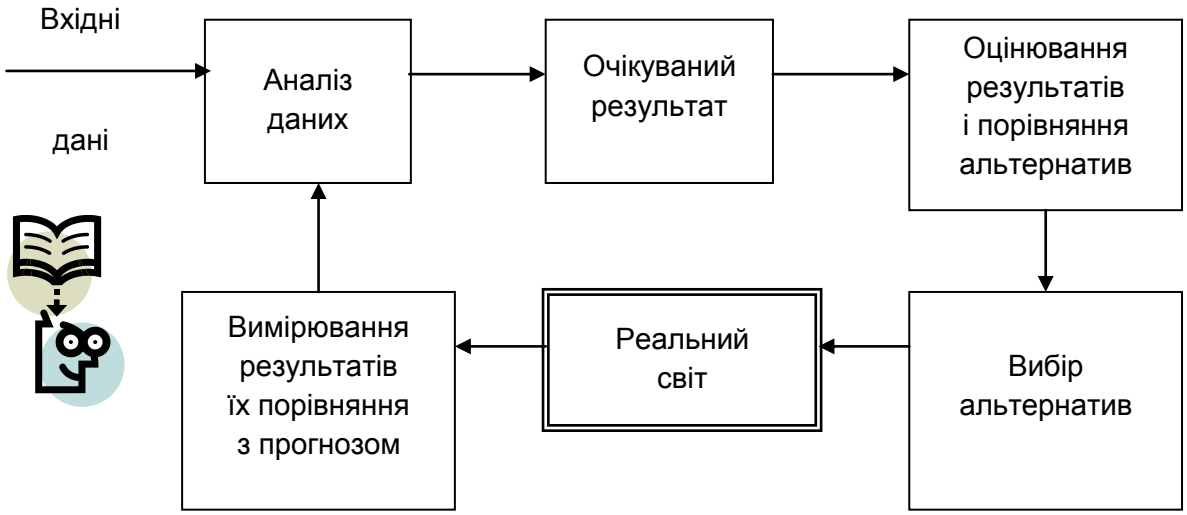


Рис. 1.8. **Загальна схема процесу формування рішень**

Експериментувати в житті, застосовуючи різні управлінські дії і вивчаючи реакцію підлеглих, неефективно та практично є неможливим. Тут на допомогу приходить моделювання – метод дослідження, що полягає в побудові й аналізі моделей – аналогів об'єктів дослідження. Маючи адекватну модель, можна з її допомогою проаналізувати реакції системи управління (етап аналізу на рис. 1.8), а потім вибрати (на етапі синтезу) та використовувати на практиці той управлінський вплив, який призводить до необхідної реакції. Наявність моделей і механізмів управління корисно як з точки зору органу управління (тому що дозволяє передбачити поведінку суб'єктів управління), так і самих суб'єктів (оскільки робить передбачуваною поведінку керівного органу). Тобто зниження невизначеності за рахунок використання механізмів управління є однією з істотних властивостей будь-якої організації як соціального інституту.

Вивчення певного економічного багатовимірного об'єкта будь-якої форми – це розкриття не тільки його якісних, але і кількісних закономірностей за допомогою моделей багатовимірних об'єктів. Під **моделлю** будемо розуміти образ реального об'єкта (процесу) в матеріальній або ідеальній формі (тобто описаний знаковими засобами будь-якою мовою), що відбиває істотні властивості змодельованого об'єкта (процесу) й управління в ході дослідження.

Проблема – це ситуація, що характеризується розходженням між необхідним (бажаним) виходом та існуючим входом. Вона має два стани: існуючий і бажаний.

Проблеми розподіляють на:

кількісні, для вирішення яких використовують способи маніпулювання числами;

якісні проблеми, пов'язані з перерахуванням майбутніх чи погано визначених ресурсів і їх властивостей;

слабоструктуровані проблеми, склад елементів, властивості та зв'язки яких відомі тільки частково.

Моделювання – це універсальний спосіб вивчення процесів і явищ реального світу. Особливого значення воно набуває у процесі вивчення об'єктів, не доступних прямому спостереженню та дослідженню. До них, зокрема, відносять соціально-економічні явища та процеси. Моделювання завжди має цільову спрямованість. Цілі та методи його можуть бути різноманітними.

Відповідно, **економіко-математична модель** – це вираз формальної залежності у вигляді математичних співвідношень (функцій, систем рівнянь або нерівностей, операторів тощо) економічного процесу, проблеми або об'єкта, що досліджуються. Надалі будемо говорити тільки про економіко-математичне моделювання методами багатовимірного аналізу, тобто про опис знаковими математичними засобами соціально-економічних систем.

Розрізняють вербальне, геометричне (предметне), фізичне й інформаційне моделювання:

вербальне моделювання – моделювання на основі використання розмовної мови;

геометричне моделювання здійснюється на макетах або об'єктних моделях. Ці моделі передають просторові форми об'єкта, пропорції тощо;

фізичне моделювання застосовують для вивчення фізико-хімічних, технологічних, біологічних, генних процесів, що відбуваються в оригіналі. Таке моделювання називають аналоговим;

інформаційне моделювання має фундаментальне значення в усіх галузях науки (схеми, графіки, креслення, формули, рівняння, нерівності).

Найважливіша роль серед методів інформаційного моделювання належить логіко-математичному моделюванню, тобто моделюванню за допомогою застосування математичного апарата багатовимірного статистичного аналізу.

Під час побудови економіко-математичної моделі реалізується метод моделювання за принципом "чорного ящика", коли досліднику невідомий механізм процесів, що протікають у системі, вивчити який можна за вхідними та вихідними характеристиками системи. Вхідні та вихідні характеристики системи часто ототожнюють з екзогенними й ендогенними змінними, або вживають терміни "незалежні" (факторні) змінні або ознаки і "залежні" (результативні) змінні або ознаки. Графічно принцип "чорного ящика" зображено на рис. 1.9.

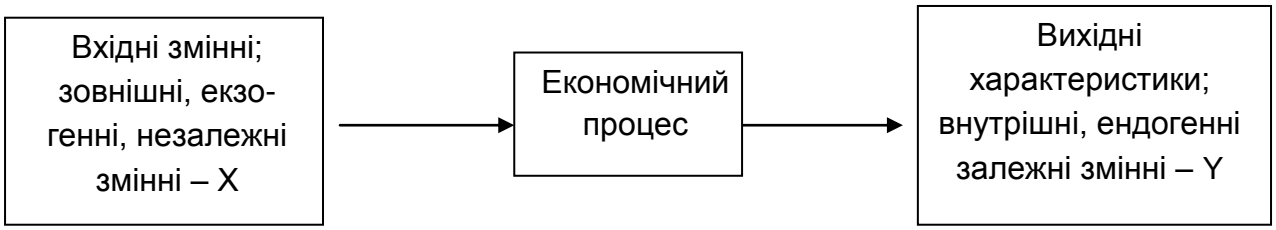


Рис. 1.9. **Дослідження системи за принципом "чорного ящика"**

Досліднику необхідно виділити вхідні та вихідні характеристики та на підставі методів установити характер причинно-наслідкових зв'язків, що закладені в основу механізму функціонування соціально-економічної системи.

Етапи дослідження за допомогою багатовимірного статистичного аналізу:

1) постановка задачі включає опис предметної області об'єктів, визначення обсягів виділених ресурсів (час, трудовитрати і т. д.);

2) визначення набору методів багатовимірного статистичного аналізу та порядку їх використання;

3) збирання вихідних даних, визначення способів збирання інформації, форм її подання;

4) аналіз даних (перевірка однорідності вибірки, відповідність законам розподілу, виявлення грубих помилок і т.д.);

5) уточнення математичної постановки задачі та визначення можливості застосування раніше відібраних методів (у разі необхідності набір методів змінюється);

6) реалізація – проведення обчислень, реалізація за допомогою програмного забезпечення математичного інструментарію;

7) оцінювання адекватності моделі, визначення несуперечності математичних результатів і економічних висновків.

Одним з головних принципів управління економічними об'єктами є принцип системності. Система може бути подана нескінченним числом структурних і функціональних інваріантів, які відбивають взаємозв'язки між різними процесами, що протікають у цій системі (економічними, соціальними, екологічними, демографічними і т. д.). Опис системи здійснюється за допомогою її якісних і кількісних характеристик, іменованих параметрами. Параметри складають основу мов опису систем, а шляхом формалізації ототожнюються з незалежними змінними математичного опису процесу функціонування систем.

Чим повніше й обґрунтованіше принципи моделювання, тим вище точність моделі та ймовірність досягнення позитивних результатів від її застосування. У моделюванні найбільшого значення мають такі основні принципи.

Принцип інтегратизма полягає в тому, що окремі частини цілого характеризуються сукупністю трьох елементів:

1) виникненням взаємодіючих систем – зв'язків між частинами цілого;

2) утратою деяких властивостей частини зі входженням у ціле;

3) появою нових властивостей у цілого, обумовлених властивостями складових.

Водночас обов'язковою є впорядкованість частин, детермінованість їх просторових і функціональних звязків, коли частина стає компонентом інтегрального цілого, внутрішньо об'єднаного. Цей принцип перетинається з відомим протилежним положенням У. Ешбі, який розглядає загальну теорію систем як загальну теорію спрощення.

Принцип невизначеності припускає, що "на краях" економічні процеси розпливчасті та невизначені. Протікаючи в часі, вони постійно змінюються, і якщо нам навіть удасться встановити яку-небудь якість чи властивість процесу, то вона дійсна тільки в розглянутий момент часу в даній ситуації. Інакше кажучи, на мікрорівні економічні процеси необхідно розглядати з урахуванням випадкової зміни факторів.

Принцип невизначеності дозволяє також стверджувати, що існує рівень факторів, коли їхні малі відхилення не призводять до змін у стані системи. Однак чим складніше модель системи, чим глибше ми намагаємося аналізувати її, тим більш невизначеним стає розв'язок задачі, а її результати тим далі від практичного змісту.

Принцип інваріантості полягає в тому, що модель системи повинна бути інваріантна для будь-яких організаційних форм виробництва, а зміна умов не має змінювати сутності моделі.

Принцип головних видів діяльності полягає в тому, що в різних системах існують "схожі" види діяльності (управління, регулювання, розподіл і т.п.), які можна виділити як стандартні. Вони бувають незмінні на певному проміжку часу; їх можна описати деякими схожими моделями.

Для моделювання об'єктів і подання їх у вигляді систем необхідно враховувати **властивості економічних систем** – такі, як:

емерджентність як прояв у найбільш яскравій формі властивості цілісності системи. Тобто наявність у економічної системи таких властивостей, які не притаманні жодному з елементів системи, взятому окремо, поза системою. Емерджентність є результатом виникнення між елементами системи так званих синергічних зв'язків, що забезпечують збільшення загального ефекту до величини, більшої за суму ефектів елементів системи, які діють незалежно;

подільність – цілісний об'єкт повинний бути зображений поділеним на елементи;

неможливість ізолювати процеси, що протікають в економічних системах, і явища від навколишнього середовища, щоб спостерігати та досліджувати їх у чистому вигляді;

стійкість – система має нормально функціонувати та бути нечутєвою до неминучих сторонніх дестабілізаційних впливів;

розмаїтість – кожному елементу системи властиве поводження та стан, відмінні від поводження та стану інших елементів і системи в цілому;

ідентифікованість – кожен елемент системи може бути відділений від інших складових;

стабілізація – система здійснює відновлення своїх елементів за рахунок їхнього регулювання;

спостережність – усі без винятку входи та виходи системи або контрольовані дослідником, або принаймні дослідник може за ними спостерігати;

невизначеність – дослідник одночасно не може фіксувати всі властивості та відношення елементів системи. Саме з метою їхнього виявлення він здійснює системне дослідження;

нетотожність відображення – знакова система дослідника відмінна від знакової системи прояву властивостей об'єктів і їхніх відношень. Втрата інформації визначає нетотожність системи досліджуваному об'єкту;

адаптація – система зберігає стан рухливої рівноваги та стійкість до дестабілізаційних впливів, яким вона постійно піддається, шляхом перебудови внутрішньої структури та функцій окремих елементів;

масовий характер економічних явищ і процесів – закономірності економічних процесів не виявляються на підставі незначної кількості спостережень;

динамічність економічних процесів полягає в зміні параметрів і структури економічних систем під впливом середовища (зовнішніх факторів);

випадковість і невизначеність у розвитку економічних явищ – закономірності економічних процесів виявляються на підставі випадкових величин і факторів, параметри розподілу яких можна встановити з деякою ймовірністю;

активна реакція на нові фактори, *що з'являються,* – здатність соціально-економічних систем до активних, не завжди передбачуваних дій залежно від відношення системи до цих факторів.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. У чому полягають особливості багатовимірного статистичного аналізу?

2. Як можна використовувати багатовимірний аналіз у бізнес-аналітиці?

3. Назвіть переваги методів багатовимірного аналізу.

4. Назвіть основні етапи дослідження за допомогою БСА.

5. Які існують особливості обробки багатовимірних статистичних даних?

6. Які існують методи багатовимірного статистичного аналізу?

7. У чому полягає відмінність методів багатовимірного статистичного аналізу від методів класичної статистики?

8. Назвіть методи багатомірного статистичного аналізу та задачі, які можна розв'язати з їх допомогою.

9. Охарактеризуйте методологічну та теоретичну основу багатовимірного статистичного аналізу.

10. Сформулюйте поняття простору ознак. Наведіть приклади одновимірного, двовимірного та багатовимірного простору ознак.

Тестові завдання

1. Факторний аналіз належить до:

а) методів логіко-геометричного напряму БСА;

б) методів ймовірнісного аналізу даних БСА.

2. Основи теорії нечітких множин у БСА започаткував:

а) Р. Фішер;

б) Л. Терстоун;

в) Л. Заде;

г) П. Махаланобіс.

3. Для групування багатовимірних об'єктів застосовується метод:

а) багатовимірного шкалювання;

б) кластерного аналізу;

в) канонічних кореляцій.

4. Формою подання вихідних даних про багатовимірний об'єкт є:

а) матриця значень аналітичних ознак;

б) матриця теоретичних відстаней між об'єктами;

в) усі відповіді правильні.

5. Економіко-математична модель – це:

а) вираз формальної залежності у вигляді математичних співвідношень (функцій, систем рівнянь або нерівностей, операторів тощо) економічного процесу, проблеми або об'єкта, що досліджуються;

б) ситуація, що характеризується розходженням між необхідним (бажаним) виходом та існуючим входом;

в) образ реального об'єкта (процесу) у матеріальній або ідеальній формі (тобто описаний; знаковими засобами на якій-небудь мові), що відбиває істотні властивості змодельованого об'єкта (процесу) й управління в ході дослідження.

6. Екзогенні змінні в дослідженнях економічних процесів – це:

а) вхідні змінні;

б) вихідні змінні.

7. Властивість цілісності економічної системи називають:

а) емерджентністю;

б) стійкістю;

в) ідентифікованістю;

г) усі відповіді правильні.

8. Властивість, коли система зберігає стан рухливої рівноваги та стійкість до дестабілізаційних впливів, яким вона постійно піддається, шляхом перебудови внутрішньої структури та функцій окремих елементів – це:

а) динамічність;

б) адаптація;

в) невизначеність.

9. Принцип інваріантності у моделюванні полягає в тому, що:

а) зміна умов у системі не повинна змінювати суті моделі;

б) встановлена властивість процесу дійсна тільки в розглянутий момент часу в даній ситуації;

10. *Оцінювання адекватності моделі, визначення несуперечності математичних результатів є обов'язковими етапом БСА:*

а) так;

б) ні.

Практичні завдання

Завдання 1. Досліджується рівень фінансової безпеки найбільших банків України (не менше десяти банків). Обґрунтуйте вибір показників фінансової безпеки банків і сформуйте матрицю значень аналітичних ознак (x_j) (табл. 1.2).

Таблиця 1.2

Матриця значень аналітичних ознак

Об'єкти	X_1	X_2	X_3	X_4	…	X_m
n_1	x_{11}	x_{12}	x_{13}	x_{14}	…	x_{1m}
n_2	x_{21}	x_{22}	x_{23}	x_{24}	…	x_{2m}
n_3	x_{31}	x_{23}	x_{33}	x_{34}	…	x_{3m}
…	…	…	…	…	…	…
n_n	x_{n1}	x_{n2}	x_{n3}	x_{n4}	…	x_{nm}

Завдання 2. Досліджується рівень життя населення в різних регіонах України. Треба сформулювати найбільш повний перелік змінних, що характеризують рівень життя населення. Виділіть серед них ендогенні й екзогенні змінні системи. Побудуйте причинно-наслідкові зв'язки між основними змінними у вигляді рекурсивної системи одночасних рівнянь.

Завдання 3. Опишіть модель залежності рівня загрози банкрутства підприємства від концентрації позикового капіталу. Серед залежностей (лінійна, показникова, модифікована показникова) необхідно вибрати тип функції та записати найбільш ймовірні для опису взаємозв'язку досліджуваних змінних.

Завдання 4. Оберіть для дослідження економічну систему (промислове підприємство, банк, туристична фірма, страхова компанія, регіон, країна тощо). Визначіть основні змінні, які характеризують, функціонування обраної економічної системи. Наведіть приклади й обґрунтуйте різні типи зв'язку між змінними (прямий зв'язок – зі збільшенням факторної змінної значення результативної ознаки збільшується, зворотний – зі збільшенням факторної змінної значення результативної ознаки зменшується).

Розділ 2. Вимірювання і типи вимірювальних шкал. Методи оцінювання вибірки

2.1. Поняття, сутність вимірювання та їх класифікація.

2.2. Вибіркова сукупність, оцінювання якості та формування вибірки.

2.3. Сутність та основи робастного оцінювання вибірки.

2.4. Статистичні критерії виявлення грубих помилок.

2.5. Основні методи визначення стійких статистичних оцінок.

Ключові слова: вимірювання; валідність; надійність; шкала вимірювань; генеральна сукупність; вибіркова сукупність; методи робастного оцінювання; статистичні критерії; стійкість статистичних оцінок.

Література: [14; 15; 23; 34; 36].

2.1. Поняття, сутність вимірювання та їх класифікація

Теорія вимірів (ТВ) є однією із складових частин прикладної статистики. Вона входить до складу статистики об'єктів нечислової природи. Застосування ТВ у соціально-економічних дослідженнях необхідне для аналізу якісних даних як теорія, що є основою формування баз даних для розроблення, вивчення і застосування конкретних методів розрахунку. Розглянемо основні поняття вимірювання.

Вимірювання – сукупність дій, виконуваних за допомогою засобів вимірювань з метою знаходження числового значення вимірюваної величини в прийнятих одиницях вимірювання

Вимірювання – кодування та співвіднесення ступеня вираженості ознак емпіричних об'єктів або подій за допомогою чисел відповідно до певних встановлених правил

Загальна класифікація вимірювань показників за чотирма визначальними напрямами наведена в на рис. 2.1.

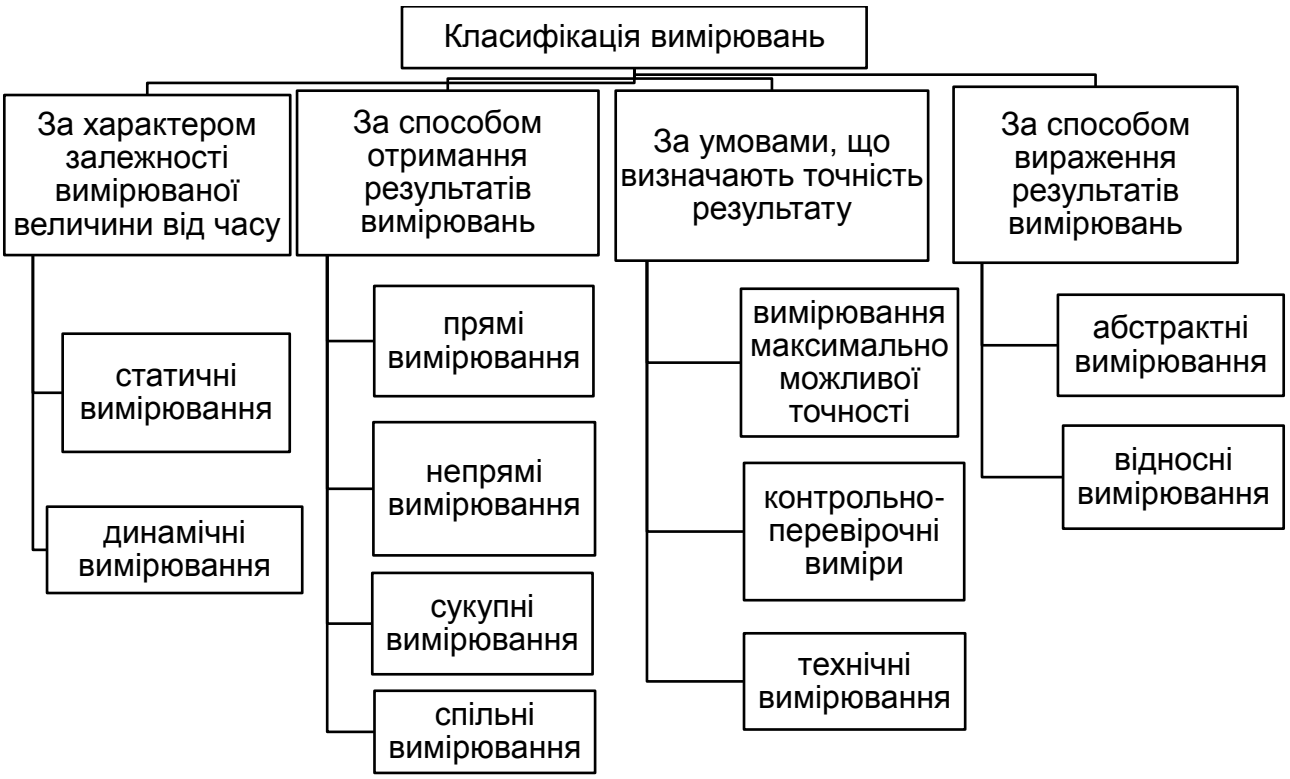


Рис. 2.1. **Загальна класифікація вимірювань**

Основні оцінки якості проведених вимірювань наведені на рис. 2.2.

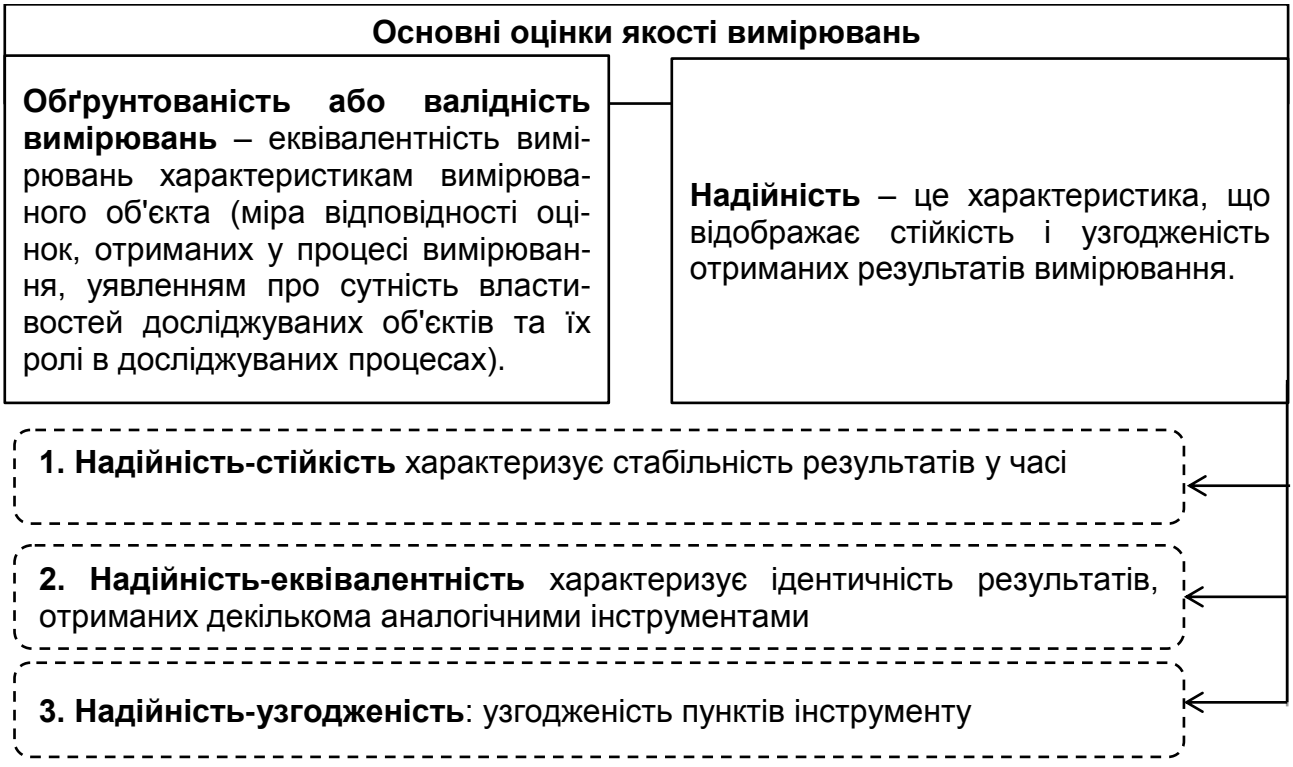


Рис. 2.2. **Основні оцінки якості вимірювань**

Розглянемо більш детально типи валідності вимірювань, що певним чином характеризують еквівалентність вимірювань характеристикам вимірюваного об'єкта (міра відповідності оцінок, одержуваних у процесі вимірювання, уявленням про сутність властивостей досліджуваних об'єктів та їх ролі в досліджуваних процесах) (рис. 2.3).

Рис. 2.3. **Типи валідності**

Для якісного дослідження сукупностей та побудови адекватних моделей залежностей показників і прогнозування необхідне чітке уявлення про поняття шкали як в загальному сенсі, так і відповідно до особливос-

тей досліджуваних даних і специфіки сукупностей. Отже, поняття шкали вимірювання формулюють таким чином:

Шкала вимірювань – це певний, заздалегідь установлений порядок визначення та позначення можливих значень конкретної величини або проявів якої-небудь властивості. Тип шкали задає групу допустимих перетворень шкали.

Розрізняють основні атрибути шкал:

упорядкованість шкали – коли одна позиція шкали, яка визначається числом і відповідна вираженості вимірюваної властивості, більше, менше або дорівнює іншій позиції;

інтервальність шкали означає, що інтервали між позиціями шкали дорівнені між собою;

нульова точка (точка відліку) шкали означає, що набір чисел, відповідних вираженості вимірюваної ознаки, має точку відліку, що позначеному 0, яка відповідає повній відсутності вимірюваної властивості.

Рис. 2.4. **Класифікація вимірювальних шкал**

Ця класифікація є найбільш поширеною для метричних і неметричних даних у сукупностях спостережень. На рис. 2.4 схематизована класифікація вимірюваних шкал.

2.2. Вибіркова сукупність, оцінювання якості та формування вибірки

Одним з визначальних факторів проведення якісних соціально-економічних досліджень є формування сукупності спостережень, тобто вибіркової сукупності, що буде визначати характеристики генеральної сукупності та буде задовольняти висунутим оцінкам якості.

Дамо визначення поняттям генеральної сукупності, вибіркової та вибірки спостережень.

Генеральна сукупність	сукупність усіх елементів з рядом спільних характеристик, яка охоплює повну множину елементів з точки зору вирішення певної проблеми
Вибіркова сукупність (вибірка)	відібране за строго визначеними правилами певне число елементів генеральної сукупності
Вибірка	своєрідна мікромодель (проекція) всієї генеральної сукупності, яка за всіма основними досліджуваними якісними характеристиками та контрольними ознаками повинна своєю структурою максимально повторювати структуру генеральної сукупності. **Мета формування вибірки** – отримання достовірних висновків про властивості досліджуваної генеральної сукупності.

Класифікація вибіркових сукупностей за визначальними ознаками наведена на рис. 2.5.

Отже, дослідження об'єктів вибірки називають вибірковим дослідженням. Після аналізу вибіркового дослідження перевіряють ряд гіпотез для обґрунтування можливості перенесення цих результатів на всю генеральну сукупність. Переваги вибіркового дослідження:

суттєве зниження витрат;

скорочення часу на отримання даних і їх обробку.

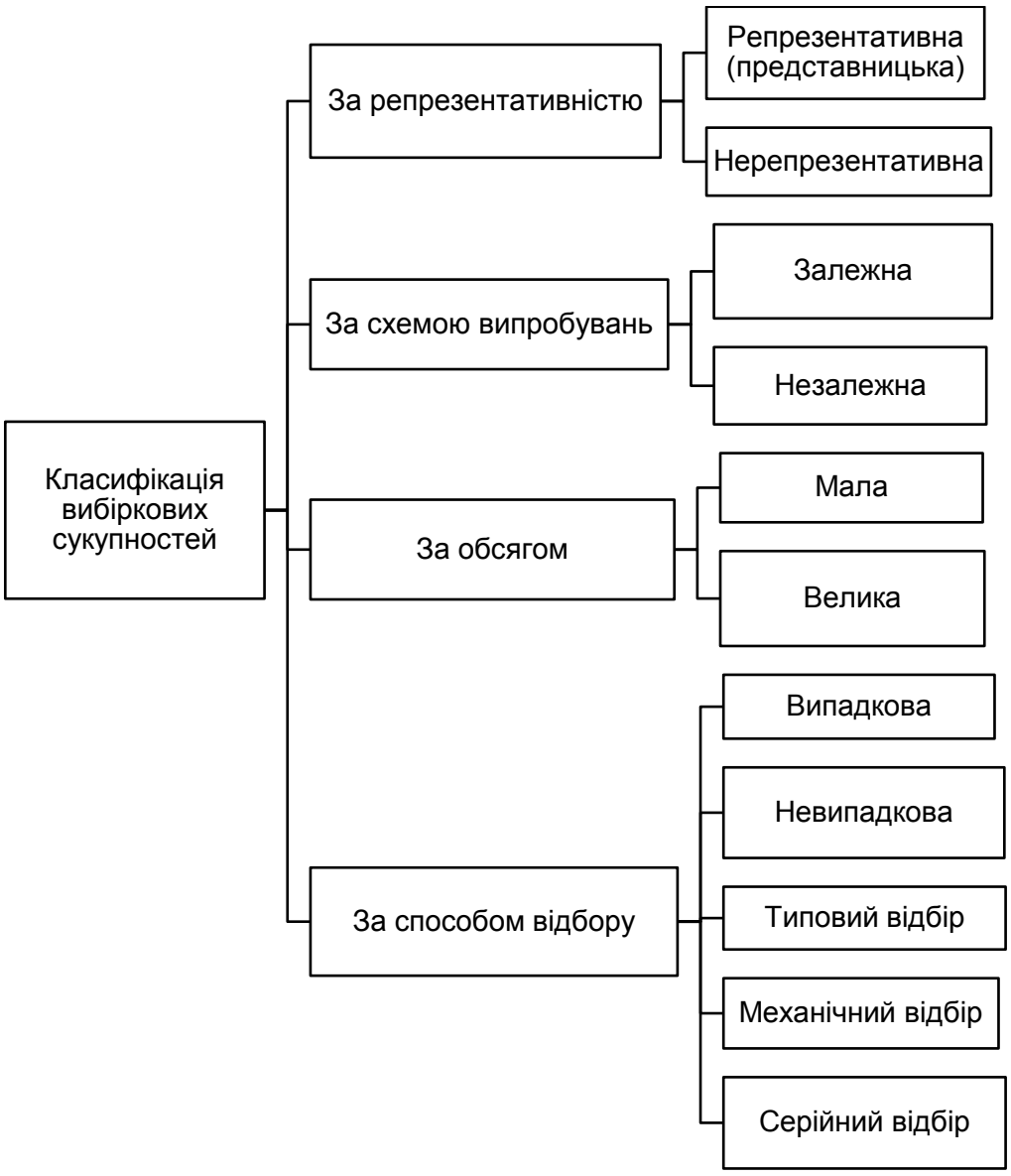


Рис. 2.5. **Класифікація вибіркових сукупностей**

Схема формування вибіркової сукупності подана на рис. 2.6.

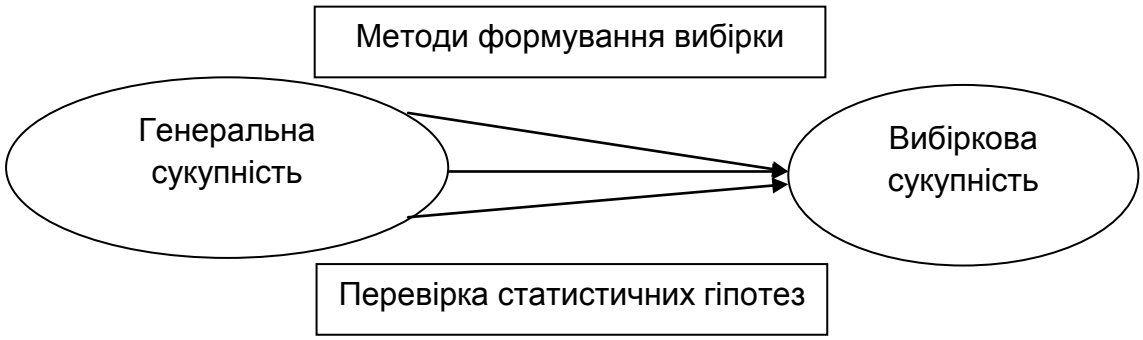


Рис. 2.6. **Схема формування вибіркової сукупності**

Визначальним фактором для подальшого застосування суб'єктів вибіркової сукупності є оцінювання її якості за показниками репрезентативності та надійності.

Репрезентативність – вибірка за своєю структурою максимально повторює структуру генеральної сукупності;

Надійність визначається за такими параметрами:

повнота вибірки (в ній представлені всі елементи генеральної сукупності);

точність інформації (у ній немає неіснуючих одиниць спостереження);

адекватність (вибірка співвідноситься з розв'язанням поставлених дослідженням задач).

Існує кілька методів формування вибірок із загальних генеральних сукупностей. У найбільш загальному випадку **формування вибірки можна розподілити на дві групи:**

ймовірнісні (випадкові) вибірки формуються за допомогою таких підходів, які передбачають ретельне проходження алгоритму, не залишаючи місце безсистемності та "випадковості";

умовно-ймовірнісні вибірки формуються "за бажанням", "за присутністю", "за доступністю" та іншими аналогічними принципами та критеріями.

Досить поширеною є думка про те, що збільшення точності результатів вибіркового дослідження пропорційно збільшенню числа елементів вибірки. Це думка не зовсім справедлива, оскільки ґрунтується на трьох помилках:

Помилка 1. Чим більше вибірка, тим вона репрезентативніша

Помилка 2. Вибірка повинна складатися з як мінімум 10 % елементів генеральної сукупності.

Помилка 3. Заздалегідь не можна відповісти на запитання про необхідну та достатню чисельність вибірки

Необхідний обсяг вибірки є функцією варіації змінних параметрів генеральної сукупності та точності оцінки цих параметрів, необхідної дослідникy.

Загальне правило формування вибірки: чим більше дисперсія оцінюваних параметрів генеральної сукупності, тим більший обсяг вибірки, потрібний для того, щоб забезпечити необхідну точність.

2.3. Сутність та основи робастного оцінювання вибірки

У процесі дослідження статистичних сукупностей часто спостерігаються дані, що різко відхиляються від основного масиву, тобто є похибками, або "викидами".

У ході виявлення подібних похибок ("викидів") виникають такі запитання: чи дійсно ці дані є похибками (наприклад, реєстрації) або це реальні значення; як отримати адекватні оцінки для параметрів досліджуваної сукупності? Дослідження цих проблем цих питань виконують у спеціальному розділі економіко-статистичних методів – робастне (стійке) оцінювання. Основи робастних методів оцінювання інформації були розроблені академіками: А. Н. Колмогоровим, Н. В. Смирновим, Б. С. Ястремським. Подальший розвиток методи отримали в роботах американського статистика Д. Тьюкі та швейцарського математика П. Хубера.

Методи робастного оцінювання – це статистичні методи, які дозволяють отримати досить надійні оцінки статистичної сукупності з урахуванням неявності закону розподілу та наявності значних відхилень у даних

У вирішенні задач робастного оцінювання у статистичній сукупності виділяють два типи даних (рис. 2.7).

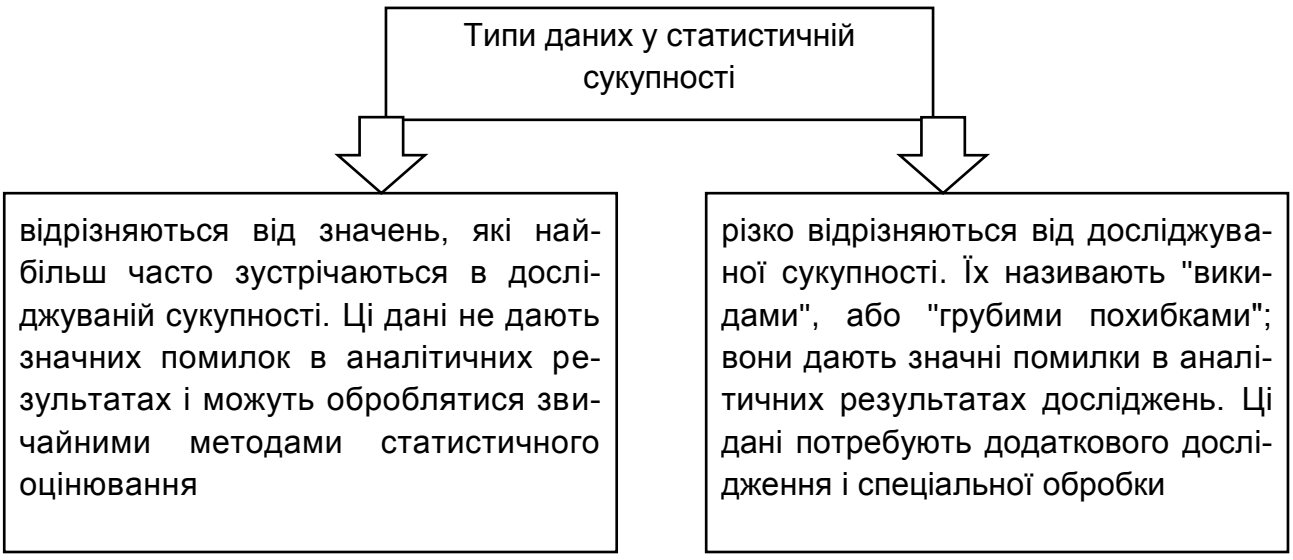


Рис. 2.7. **Типи даних у статистичній сукупності**

У практиці стійкого оцінювання виділяють такі основні причини наявності грубих помилок (рис. 2.8).

Рис. 2.8. **Основні причини наявності грубих помилок**

Алгоритм обробки "грубих похибок" включає такі основні кроки (рис. 2.9):

1) розпізнавання похибок у даних;

2) вибір методу та проведення робастного оцінювання;

3) критеріальна або логічна перевірка й інтерпретація результатів стійкого оцінювання.

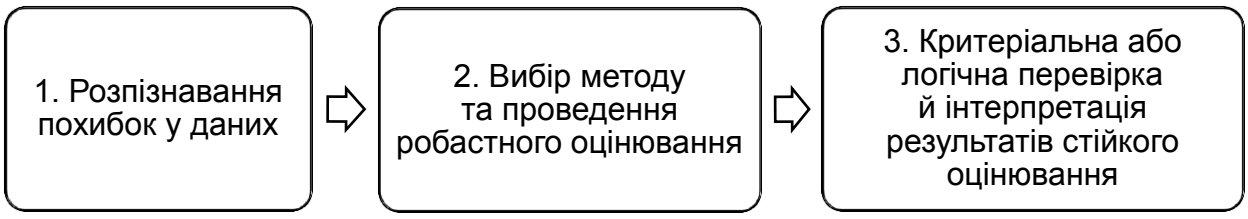


Рис. 2.9. **Алгоритм обробки "грубих похибок"**

Виявлення "грубих похибок" і оцінювання їх ступеня можливі за допомогою візуального аналізу даних або перевірки статистичної гіпотези на наявність похибок. Більш глибокий аналіз передбачає розрахунок спеціальних статистичних критеріїв.

2.4. Статистичні критерії виявлення грубих помилок

У практиці оцінювання розрізняють такі статистичні критерії виявлення грубих помилок для одновимірних і багатовимірних даних.

Одновимірна сукупність: Т-критерій Граббса; L-критерій; L'-критерій; Е-критерій. *Багатовимірна сукупність:* критерій Фішера (F). Розглянемо пропоновані критерії та алгоритми їх розрахунку більш детально.

1. Тест для виявлення похибок, заснований на розрахунку Т-критерію Граббса. Даний критерій є простим, що дозволяє легко застосовувати його в аналізі, але має ряд недоліків. Зокрема, недостатня точність (дає досить грубі оцінки); нечутливість до маскувальних ефектів (коли похибки групуються досить близько у віддаленості від основної маси спостережень). Етапи алгоритм Т-критерію Граббса наведено на рис. 2.10.

I етап	Розрахунок середнього значення $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ за вибіркою спостережень
II етап	Розрахунок середньоквадратичного $S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ відхилення за вибіркою
III етап	Розрахунок за формулою значення Т- критерію для кожного елемента вибірки $T_p = \frac{x_{ij} - \bar{x}_j}{s_j}$
IV етап	Визначення табличного значення критерієм Смірнова – Граббса $(\mathrm{T_{кр}} = \mathrm{T}_{\alpha,n})$.
V етап	Досліджуваний об'єкт є похибкою, якщо $\mathrm{T_р} > \mathrm{T_{кр}}$

Рис. 2.10. **Алгоритм Т-критерію Граббса**

2. Оцінювання грубих похибок на основі L- і E-критеріїв (Тітьєна та Мура).

L-критерій застосовується для виявлення грубих похибок у верхній частині ранжованого ряду даних, розраховується в такий спосіб:

$$L = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x}_k)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (2.1)$$

де x_i – вибірка і-го спостереження за j-ю ознакою;

n – обсяг вибірки;

k – число об'єктів, підозрюваних на похибку;

$\bar{x}$ – середнє значення для всієї сукупності;

$\bar{x}_k$ – середня, розрахована за n-k спостереженнями, що залишилися після відкидання k грубих похибок, "зверху" ранжованого ряду даних.

L'-критерій застосовується для виявлення грубих похибок у нижній частині ранжованого ряду даних, розраховується в такий спосіб:

$$L' = \frac{\sum_{i=k+1}^{n}(x_i - \bar{x}_k)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (2.2)$$

де $\bar{x}_k$ – середня, розрахована за n-k спостереженнях, що залишилися після відкидання k грубих помилок, "знизу" ранжованого ряду даних.

Алгоритм L-критерію наведено на рис. 2.11.

E-критерій використовується за наявності у вибірці грубих похибок, розташованих у верхній та нижній частинах ранжованого ряду даних, розраховується в такий спосіб:

$$E = \frac{\sum_{i=k+1}^{n-k'}(x_i - \bar{x}_{k'})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (2.3)$$

де $\bar{x}_{k'}$ – середня, розрахована за даними після відкидання з вибірки найменших (k) і найбільших (k') значень, які розглядаються як похибки.

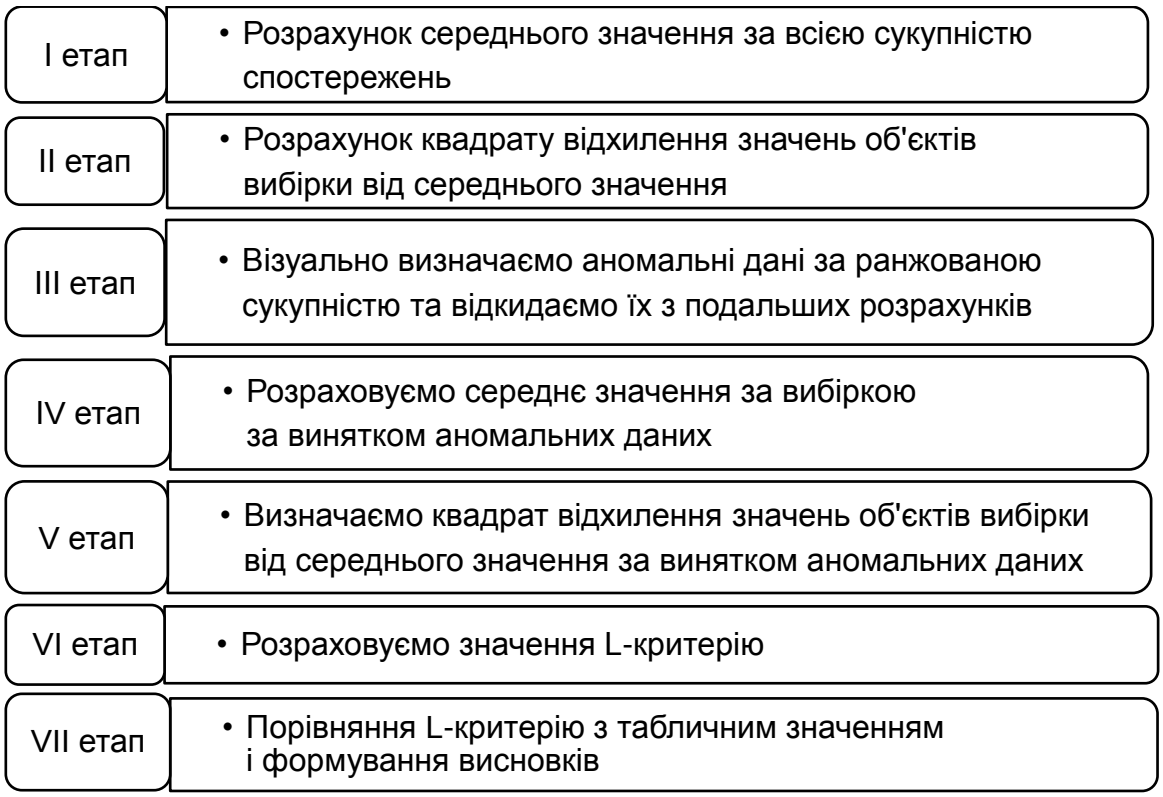


Рис. 2.11. **Алгоритм L-критерію**

Розглянуті критерії L, L' і E мають табульовані критичні значення для заданого рівня значущості (α) за відомого обсягу вибірки (n) і кількості похибок (k). Якщо розрахункові значення критеріїв менше за критичні $C_{\alpha,k}$, то похибки в даних, які перевіряються, є грубими, тобто істотно відхиляються від основного масиву даних. Якщо L, L', E > $C_{\alpha,k}$, дані є типовими для досліджуваної вибіркової сукупності.

3. Оцінювання багатовимірної сукупності даних на основі F-критерію Фішера. У багатовимірному випадку "похибками" сукупності даних вже є не окремі значення, а вектор значень – аномальний об'єкт. Для оцінювання багатомірного спостереження використовують відстань Махаланобіса, що розраховується за формулою:

$$d^m = (X - \bar{X})' \sum^{-1} (X - \bar{X}), \qquad (2.4)$$

де X – вектор значень, досліджуваних на "похибку";

$\bar{X}$ – вектор середніх значень для багатовимірної сукупності;

Σ – матриця коваріацій.

F-критерій для перевірки гіпотези про істотність відхилення випадкового вектора X розраховують в такий спосіб:

$$F_p = \frac{(n-m)n}{2(n-1)m}(X-\bar{X})\sum^{-1}(X-\bar{X}). \quad (2.5)$$

Алгоритм F-критерію Фішера наведено на рис. 2.12.

I етап	• Розрахунок вектора середніх значень за кожною змінною вибіркових даних у сукупності
II етап	• Розрахунок матриці коваріацій між змінними та оберненої до неї
III етап	• Розрахунок вектора відхилень показників від середніх значень і транспонованих значень даного вектора відхилень
IV етап	• Розрахунок відстані Махаланобіса
V етап	• Розрахунок F-критерію для перевірки гіпотези про істотність відхилення випадкового вектора X

Рис. 2.12. **Алгоритм F-критерію Фішера**

У випадку наявності значної кількості похибок, багатовимірна сукупність перевіряється ітеративним методом (рис. 2.13).

→ Перевіряється одне зі спостережень, що приблизно є "грубою похибкою"; якщо припущення доводиться, спостереження усувається з вибірки

→ За усіченою сукупністю багатомірних об'єктів визначається новий вектор середніх значень

→ Перевірці піддається наступний об'єкт

→ Повторення кроків 1 – 3

Рис. 2.13. **Ітеративний метод перевірки багатовимірної сукупності за наявності значної кількості похибок**

Для F-критерію число ступенів свободи дорівнює відповідно: $\nu_1 = m$ і $\nu_2 = n - m - 1$. Із заданим рівнем значущості (α), якщо $\mathrm{F_p} > \mathrm{F}_{\alpha,\nu_1,\nu_2}$, досліджуваний об'єкт є аномальним. У протилежному випадку, коли $\mathrm{F_p} \le \mathrm{F}_{\alpha,\nu_1,\nu_2}$ відхилення випадкового вектора від вектора середніх значень є допустимим, а гіпотеза про "похибки" сукупності відкидається.

2.5. Основні методи визначення стійких статистичних оцінок

Після виявлення похибок у даних ставиться задача оцінювання параметрів вибіркової сукупності. Для цього використовують: усунення з вибіркової сукупності похибок та оцінювання параметрів за усіченою сукупністю; модифікацію досліджуваних даних. Підходи до вирішення проблеми грубих помилок та їх характеристика наведені на рис. 2.14.

Перший підхід	**Усічення вибірки** – відкидання певної частини спостережень з мінімальними або максимальними значеннями та почастинне оцінювання параметрів розподілу
Другий підхід	**Вінзорування** – всім спостереженнями лівіше та/або правіше певних визначених критичних значень привласнюють однакові розраховані середні значення
Третій підхід	**Цензурування** – для спостережень, що потрапили лівіше та/або правіше певних значень, фіксують лише факт потрапляння у відповідний інтервал, опускаючи конкретні значення цих спостережень. За такою цензуруваною вибіркою оцінюють параметри
Четвертий підхід	**Функція правдоподібності** – розрахунок параметрів розподілу на основі максимальної правдоподібності

Рис. 2.14. **Підходи до вирішення проблеми грубих помилок**

Основні методи визначення стійких статистичних оцінок подані на рис. 2.15.

для симетричних розподілів	для асиметричних розподілів
• Метод Пуанкаре • Середня за Вінзором • Метод Хубера	• Джеккнайф-оцінка

Рис. 2.15. **Методи оцінювання параметрів вибіркової сукупності**

Оцінка Пуанкаре для розрахунку середньої за усіченою сукупністю (урізана середня):

$$T(\alpha) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i, \qquad (2.6)$$

де k – число грубих похибок; k ≤ αn – ціла частина від добутку αn;

n – обсяг вибірки;

α – деяка функція величини засмічення вибірки ξ (знаходять за таблицями).

Оцінка Вінзора передбачає заміну ознакових значень, що засмічують вибірку, на модифіковані (вінзоровані) значення з усунутими або зменшеними похибками. Середня за Вінзором визначається з відомим рівнем α $(0 < \alpha < 1/2)$ за формулою:

$$W(\alpha) = \frac{1}{n}\left(\sum_{i=k+2}^{n-k-1} x_i + k(x_{k+1} + x_{n-k})\right). \qquad (2.7)$$

Робастне оцінювання за Пуанкаре та Вінзором дають гарні результати на вибірках з симетричним розподілом засмічення, коли грубі похибки групуються приблизно на одній відстані від центра в нижній і верхній частинах статистичної сукупності.

Метод послідовного "виправлення" даних (метод Хубера) використовує вихідну величину (k), визначену з урахуванням ступеня "засміченості" статистичної сукупності (ξ), що визначає крок модифікації "хибних" спостережень. Оцінка середньої за методом Хубера визначається за формулою:

$$H = \frac{1}{n}\left(\sum_{|x_i - H| < k} x_i + (n_2 + n_1)k\right), \qquad (2.8)$$

де H – стійка оцінка середньої;

k – величина відхилення від центра сукупності, приймає значення з урахуванням питомої ваги "грубих похибок" у сукупності (ξ);

n_1 – чисельність групи спостережень із сукупності, що відрізняються найменшими значеннями: $x_i < H - k$, або значення в інтервалі$(-\infty;\ H - k)$;

n_2 – чисельність групи спостережень із сукупності, що відрізняються найбільшими значеннями: $x_i < H + k$, або значення в інтервалі$(H + k;\ \infty)$.

В якості початкової оцінки H може бути використана звичайна середня або медіана. Потім на кожній ітерації виконується групування вибіркової сукупності на три частини. В одну частину відносять "дійсні" признакові значення, які залишаються без зміни ($|x_i - H| < k$). У дві інші частини сукупності (для $x_i > H + k$ і $x_i < H - k$) відносять "похибки", які заміняються, відповідно, на величини $x_i - k$ і $x_i + k$. За "дійсними" та модифікованими даними на кожній ітерації визначається нова оцінка середньої Н; ітерація відновлюється. Ітерації повторюються поки всі спостереження не перейдуть в інтервал "дійсних" значень: $|x_i - H| < k$.

Оцінка Н, за методом Хубера, є досить ефективною, але швидко втрачає оптимальні властивості зі збільшенням засміченості вибірки (зростанням ξ).

Приклад 2.1. У табл. 2.1 наведені дані за показниками обсягу виробленої продукції та отриманого прибутку для вибіркових двадцяти машинобудівних підприємств банків України (в млн грн).

Таблиця 2.1

Вихідні дані

№ п/п	Обсяг виробленої продукції, тис. грн	Прибуток, тис. грн	№ п/п	Обсяг виробленої продукції, тис. грн	Прибуток, тис. грн
1	2515,1	986,8	11	4214,3	1130
2	669,7	263	12	333,4	200,2
3	89,2	123	13	158,6	167,6
4	155,4	49,2	14	611	312,2
5	150,1	150	15	134,5	137
6	129	43,2	16	128,5	590
7	128,3	143,1	17	1559,4	600
8	216,5	250,1	18	1599,2	70,3
9	772,6	276,4	19	374,5	136,2
10	863,5	303	20	146,2	200

Необхідно дослідити дану статистичну сукупність спостережень, виявити похибки в вибірці даних за допомогою методів робастного оцінювання. Для виявлення похибок використайте критерій Граббса, Тітьєна та Мура, F-критерій Фішера. Відповідно до розглянутого алгоритму подамо візуальний аналіз сукупності об'єктів.

Графік зміни показників підприємств машинобудування наведено на рис. 2.16 і рис. 2.17.

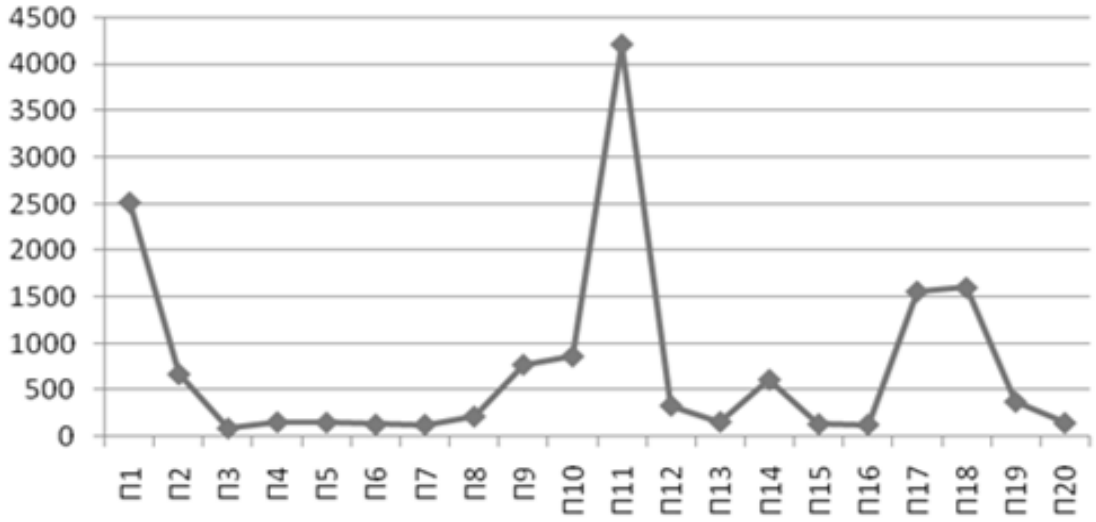


Рис. 2.16. **Обсяг виробленої продукції**

Слід зазначити, що на основі візуального аналізу можна висунути гіпотезу про неоднорідність вибірки та наявність похибок. Найбільш значні зрушення в обсягах виробленої продукції спостерігаються у підприємств 1, 11, 17, 18. Це дозволяє провести більш детальний та ґрунтовний аналіз сукупності на наявність аномальних помилок і спостережень у вибірці. Дослідимо наявність похибок у сукупності даних на основі таких критеріїв: T, L, L', E та F-критерію.

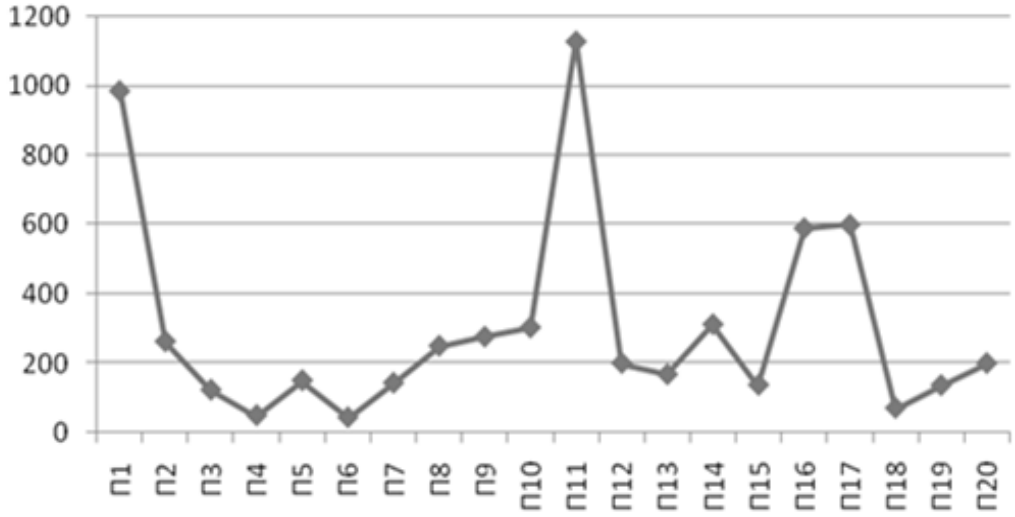


Рис. 2.17. **Прибуток**

Розрахунки на наявність похибок у сукупності даних за показником обсягу виробленої продукції подані в табл. 2.2.

Таблиця 2.2

Розрахунки на наявність похибок за показником обсягу виробленої продукції

№ п/п	Обсяг продукції	T_p	$(x_i - \bar{x})^2$	$(x_i - \bar{x}_k)^2$	$(x_i - \bar{x}_k)^2$	$(x_i - \bar{x}_{k'})^2$
П11	4214,30	3,42	12019048,92		11780032,97	
П1	2515,10	1,74	3124586,52		3003307,24	
П18	1599,20	0,84	725478,06	1305433,20	667661,01	1256509,12
П17	1559,40	0,80	659262,80	1216069,82	604203,47	1168866,24
П10	863,50	0,11	13467,60	165531,44	6626,82	148410,76
П9	772,60	0,02	632,52	99827,91	90,15	86636,73
П2	669,70	-0,08	6045,06	45392,67	12632,58	36649,72
П14	611,00	-0,13	18618,60	23825,64	29273,41	17620,22
П19	374,50	-0,37	139091,70	6747,71	166133,47	10765,89
П12	333,40	-0,41	171437,40	15189,19	201326,97	20984,08
П8	216,50	-0,52	281907,90	57669,35	319897,41	68517,68
П13	158,60	-0,58	346744,32	88830,49	388745,69	102181,76
П4	155,40	-0,58	350523,20	90748,22	392746,29	104237,82
П5	150,10	-0,59	356827,02	93969,50	399417,35	107688,21
П20	146,20	-0,59	361501,56	96375,75	404362,12	110263,06
П15	134,50	-0,60	375707,70	103777,04	419378,94	118170,13
П6	129,00	-0,61	382480,40	107350,88	426532,74	121981,73
П16	128,50	-0,61	383099,10	107678,78	427186,08	122331,23
П7	128,30	-0,61	383346,72	107810,07	427447,56	122471,18
П3	89,20	-0,65	433293,06	135015,42		
		$\sum$	**20533100,2**	**3867243,1**	**20077002,2**	**3724285,6**

Розрахунки на наявність похибок у сукупності даних за показником чистого прибутку наведені в табл. 2.3.

Таблиця 2.3

Розрахунки на наявність похибок за показником чистого прибутку

№ п/п	Чистий прибуток	T_p	$(x_i - \bar{x})^2$	$(x_i - \bar{x}_k)^2$	$(x_i - \bar{x}_k)^2$	$(x_i - \bar{x}_{k'})^2$
П11	1130,0	2,83	678045,20		631239,08	
П1	986,8	2,34	462719,66		424198,93	
П17	600,0	1,01	86104,10	142108,06	69963,19	125931,83
П16	590,0	0,98	80335,40	134668,61	64773,08	118934,45
П14	312,2	0,02	31,75	7951,69	542,63	4498,22
П10	303,0	-0,01	12,71	6395,56	1055,89	3348,79
П9	276,4	-0,10	909,93	2848,59	3492,15	977,73
П2	263,0	-0,15	1897,91	1597,78	5255,44	319,29
П8	250,1	-0,19	3188,30	732,91	7292,21	24,69
П12	200,2	-0,37	11313,51	521,11	18304,59	2018,82
П20	200,0	-0,37	11356,10	530,28	18358,74	2036,83
П13	167,6	-0,48	19311,27	3072,24	28188,54	6011,09
П5	150,0	-0,54	24512,60	5333,06	34408,19	9049,95
П7	143,1	-0,56	26720,81	6388,45	37015,62	10410,38
П15	137,0	-0,58	28752,29	7400,78	39400,04	11692,37
П19	136,2	-0,59	29024,23	7539,06	39718,28	11866,02
П3	123,0	-0,63	33696,11	10005,56	45153,89	14916,04
П18	70,3	-0,81	55821,15	23325,77	70328,09	30565,97
П4	49,2	-0,89	66236,74	30216,10		
П6	43,2	-0,91	69361,12	32338,03		
		$\sum$	**1689351**	**422973,6**	**1538689**	**352602,5**

Значення T, L, L', E-критеріїв та їх табличні значення наведені в табл. 2.4.

Таблиця 2.4

Значення критеріїв на виявлення похибок

Критерії	Обсяг виробленої продукції	Чистий прибуток	Табличні значення
$T_p = \frac{x_{ij} - \bar{x}_j}{s_j}$	$T_p = 3{,}42$	$T_p = 2{,}83$ $T_p = 2{,}34$	$T_t = 2{,}263$
	П11	П11 П1	
$L = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x}_k)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$L_p = 0{,}188$	$L_p = 0{,}25$	$L_t = 0{,}484$
	П11 П1	П11 П1	
$L' = \frac{\sum_{i=k+1}^{n}(x_i - \bar{x}_k)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$L'_p = 0{,}98$	$L'_p = 0{,}91$	$L'_t(k=1) = 0{,}639$ $L'_t(k=2) = 0{,}484$
$E = \frac{\sum_{i=k+1}^{n-k'}(x_i - \bar{x}_{k'})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$E_p = 0{,}18$	$E_p = 0{,}21$	$E(k=3) = 0{,}302$ $E(k=4) = 0{,}22$
	П11 П1 П3	П11 П1 П4 П6	

Проведений аналіз доводить, що розраховані значення критеріїв підтверджують гіпотезу про неоднорідність сукупності спостережень за досліджуваними показниками. Не відповідають загальній вибірці, тобто є похибками підприємства П11 і П1 за всіма визначеними критеріями, а також є достатньо відмінними П3, П4, П6 за відповідним показником.

Проведемо оцінювання багатовимірної сукупності даних на основі F-критерію Фішера. Матриця коваріацій та обернена до неї подана на рис. 2.18.

$$\sum = \begin{matrix} 1026655{,}0 & 246919{,}2 \\ 246919{,}2 & 84467{,}5 \end{matrix} \qquad \sum^{-1} = \begin{matrix} 3{,}28\text{E-}06 & -9{,}59\text{E-}06 \\ -9{,}59\text{E-}06 & 3{,}99\text{E-}05 \end{matrix}$$

Рис. 2.18. **Матриця коваріацій та обернена до неї**

Так, відстань Махаланобіса для першого спостереження визначається в такий спосіб:

$$d_1^m = \begin{pmatrix}1767{,}65\\680{,}235\end{pmatrix} \cdot \begin{pmatrix}3{,}28E-06 & -9{,}59E-06\\-9{,}59E-06 & 3{,}99E-05\end{pmatrix} \cdot (1767{,}65 \quad 680{,}235) = 5{,}64.$$

F-критерій для перевірки гіпотези про істотність відхилення випадкового вектора X розраховується в такий спосіб:

$$F_p = \frac{(20-2)\cdot 20}{2(20-1)\cdot 2} \cdot 5{,}64 = 26{,}71.$$

Розрахунок F-критерію Фішера за досліджуваною сукупністю наведено в табл. 2.5.

Таблиця 2.5

Розрахунок F-критерію Фішера

№ п/п	Обсяг виробленої продукції	Чистий прибуток	d^m	F_p
П1	2515,1	986,8	5,64	26,71
П2	669,7	263	0,03	0,14
П3	89,2	123	0,45	2,12
П4	155,4	49,2	0,87	4,11
П5	150,1	150	0,35	1,68
П6	129	43,2	0,90	4,25
П7	128,3	143,1	0,38	1,81
П8	216,5	250,1	0,48	2,26
П9	772,6	276,4	0,05	0,25
П10	863,5	303	0,05	0,25
П11	4214,3	1130	11,71	55,47
П12	333,4	200,2	0,17	0,80
П13	158,6	167,6	0,34	1,60
П14	611	312,2	0,08	0,37
П15	134,5	137	0,39	1,83
П16	128,5	590	7,82	37,06
П17	1559,4	600	1,03	4,86
П18	1599,2	70,3	8,46	40,10
П19	374,5	136,2	0,39	1,87
П20	146,2	200	0,41	1,94
$F_t = F(0{,}01; 2; 17) = 6{,}11$				
$F_t = F(0{,}05; 2; 17) = 3{,}59$				

За даним критерієм можна дійти висновку:

з рівнем значущості $(\alpha = 0{,}01)$ такі підприємства, як П1, П11, П16, П18, є аномальними для досліджуваної сукупності;

з рівнем значущості $(\alpha = 0{,}05)$ такі підприємства, як П4, П6 і П17, за досліджуваними показниками будуть виділятися від значень загальної вибірки. Подальший аналіз досліджуваної сукупності передбачає розрахунок стійких статистичних оцінок Пуанкаре, Вінзора, Хубера для побудови адекватних і статистично значущих моделей.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. Які типи шкал відносять до метричних, а які – до неметричних?
2. Які види систем вимірювань вам відомі?
3. Наведіть основну відмінність між вибірковою та генеральною сукупностями.
4. Як відбувається оцінювання параметрів розподілу випадкових величин?
5. Опишіть методологію формування і аналізу вибіркової сукупності.
6. Опишіть сутність методів робастного оцінювання.
7. Які існують критерії виявлення і дослідження "грубих похибок" одномірних та багатовимірних даних?
8. Назвіть стійкі методи оцінювання параметрів вибіркової сукупності.
9. Як відбувається оцінювання параметрів симетричних і асиметричних сукупностей?
10. Наведіть узагальнений алгоритм методів робастного оцінювання.

Тестові завдання

1. Вимірювання – це:

а) сукупність дій, виконуваних за допомогою засобів вимірювання з метою знаходження числового значення вимірюваної величини в прийнятих одиницях вимірювання;

б) сукупність дій, виконуваних за допомогою засобів спостережень з метою знаходження числового значення вимірюваної величини в прийнятих одиницях вимірювання.

2. За способом статистичного вираження результати вимірювання розподіляють на:

а) статичні та динамічні;

б) прямі, непрямі, сукупні, спільні;

в) абсолютні та відносні.

3. Перевагами вибіркового дослідження є:

а) зниження витрат на отримання даних та їх обробку;

б) зниження витрат і скорочення часу на обробку даних;

в) зниження витрат і скорочення часу на отримання даних.

4. Формовані вибірки можна розділити на:

а) безперервні та дискретні;

б) ймовірнісні та умовно-ймовірнісні вибірки;

в) абсолютні та відносні.

5. Стандартне відхилення вибіркових середніх за генеральною сукупністю визначають за формулою:

а) $\sigma_x = \frac{s}{\sqrt{n}}$;

б) $\sigma_x = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$;

в) $\sigma_x = \frac{\sigma}{\sqrt{n}}$.

6. Довірчий інтервал характеризує:

а) точність оцінки вимірюваної величини;

б) повноту оцінки вимірюваної величини;

в) складність оцінки вимірюваної величини.

7. Під робастністю розуміють:

а) оцінювання параметрів функції переваги;

б) нечутливість до різних відхилень у вибірці, обумовлена в загальному випадку невідомими причинами;

в) стійкість оцінок максимальної правдоподібності.

8. Неправильне віднесення елементів до досліджуваної сукупності є помилкою:

а) групування або організації спостереження;

б) помилкою реєстрації та обробки даних;

в) помилкою окремих елементів, яка не призводить до появи "аномальних" відхилень.

9. Методи робастного оцінювання – це:

а) статистичні методи, які дозволяють отримувати надійні оцінки статистичної сукупності з урахуванням наявності істотних відхилень у значеннях даних;

б) статистичні методи, які дозволяють отримувати надійні оцінки статистичної сукупності в умовах відсутності "засмічення" даних.

10. Для виявлення грубих помилок у вибірці даних використовують критерії:

а) Граббса, Вінзора, *L*- та *E*-критерії;

б) Граббса, Пуанкаре, *L*- та *E*-критерії;

в) Граббса, *L*- і *E*-критерії, Хубера;

г) Граббса, *L*- і *E*-критерії.

11. Виберіть правильну послідовність етапів обробки "засмічень" у даних:

а) критеріальна та логічна перевірка результатів стійкого оцінювання; вибір методу, проведення робастного оцінювання; розпізнавання помилок у даних;

б) вибір методу, проведення робастного оцінювання; розпізнавання помилок у даних; критеріальна та логічна перевірка результатів стійкого оцінювання;

в) розпізнавання помилок у даних; вибір методу, проведення робастного оцінювання; критеріальна та логічна перевірка результатів стійкого оцінювання.

12. Для визначення "викидів" у багатовимірних даних використовують критерії:

а) *E*-критерій;

б) Критерій Фішера;

в) *L*-критерій;

г) усі відповіді правильні.

13. Критерій Фішера розраховується за формулою:

а) $F_p = \frac{(n-m)n}{2(n-1)m}(X-\bar{X})\sum^{-1}(X-\bar{X})$;

б) $F_p = \frac{nm}{2(n-1)}(X-\bar{X})\sum^{-1}(X-\bar{X})$;

в) $F_p = (X-\bar{X})\sum^{-1}(X-\bar{X})$.

14. Для оцінювання параметрів вибіркової сукупності для симетричних розподілів використовують методи:

а) критерій Граббса, Критерій Хубера;

б) Джеккнайф-оцінка;

в) критерій Хубера; оцінки Вінзора, критерій Пуанкаре;

г) критерій Фішера, критерій Тітьєна – Мура.

15. Стійка середня може бути розрахована на основі методів:

а) Вінзора, Тітьєна – Мура, Пуанкаре, Хубера;

б) Вінзора, Граббса, Пуанкаре;

в) Вінзора, Пуанкаре, Хубера;

г) Вінзора, Тітьєна – Мура, Пуанкаре.

Практичні завдання

Завдання 1. Результати вибіркового обстеження ста домогосподарств характеризуються наведеними даними (табл. 2.6). Для кожного показника необхідно визначити відносну похибку вибірки з ймовірністю 0,954. Порівняйте отримані результати та зробіть висновки.

Таблиця 2.6

Дані вибіркового обстеження

Показники	Середній рівень	Коефіцієнт варіації, %
Середня кількість членів домогосподарств, осіб	3,1	35
Середня кількість працюючих, осіб	2,2	30
Місячний середньодушовий дохід, грн	231	52
Розмір житла на члена сім'ї, $м^2$	7,0	66
Споживання хліба, кг/людину за рік	80	18

Завдання 2. Результати 20 % вибіркового обстеження ста сімей переселенців із зони суворого радіаційного контролю за кількістю дітей у сім'ях наведено в табл. 2.7.

Таблиця 2.7

Дані вибіркового обстеження

Кількість дітей	0	1	2	3	4
Кількість сімей	11	32	30	20	7

Необхідно визначити середню кількість дітей у сім'ях переселенців і довірчі інтервали для середньої з ймовірністю 0,954. З тією ж ймовірністю визначте граничну помилку та довірчій інтервал для частки сімей, які мають більше трьох дітей. Зробіть висновки.

Завдання 3. У табл. 2.8 наведені дані про показники діяльності вибіркових двадцяти підприємств. За статистичною сукупністю спостережень виявіть похибки в вибірці даних. Для виявлення похибок використайте критерій Граббса, Тітьєна та Мура, F-критерій Фішера. Розрахуйте стійкі статистичні оцінки Пуанкаре, Вінзора, Хубера.

Таблиця 2.8

Вихідні дані

№ підприємства	Обсяг продукції (млн грн)	Чистий прибуток (млн грн)	№ підприємства	Обсяг продукції (млн грн)	Чистий прибуток (млн грн)
1	14772,9	1064,1	11	4095,1	134,3
2	11854,0	641,4	12	5001,9	315,0
3	10735,2	575,8	13	1264,7	65,2
4	8028,5	331,3	14	1993,8	86,8
5	5446,4	316,6	15	1362,3	59,0
6	5812,2	244,3	16	867,4	59,3
7	2374,3	184,7	17	2652,4	118,2
8	1938,5	73,4	18	757,4	30,7
9	2565,9	97,5	19	1719,4	55,5
10	3688,8	239,1	20	1012,3	28,6

Завдання 4. У табл. 2.9 наведені значення показника середньої заробітної плати для європейських країн за три роки.

Таблиця 2.9

Вхідні дані

№ п/п	Країни	Рік			№ п/п	Країни	Рік		
		2014	2015	2016			2014	2015	2016
1	2	3	4	5	6	7	8	9	10
1	Бельгія	1501,8	1501,8	1501,8	9	Угорщина	335,2	341,7	332,8

Закінчення табл. 2.9

1	2	3	4	5	6	7	8	9	10
2	Болгарія	158,5	173,8	184,1	10	Мальта	702,8	718	720,5
3	Чехія	318,1	309,9	331,7	11	Нідерланди	1469,4	1485,6	1501,8
4	Ірландія	1461,9	1461,9	1461,9	12	Португалія	565,8	565,8	589,2
5	Греція	683,8	683,8	683,8	13	Румунія	157,5	190,1	217,5
6	Іспанія	752,9	752,9	756,7	14	Словаччина	337,7	352	380
7	Франція	1430,2	1445,4	1457,5	15	Англія	1249,9	1251,1	1378,9
8	Люксембург	1874,2	1921	1923					

Необхідно на основі статистичної оцінки й аналізу даної сукупності розрахувати основні статистичні характеристики варіаційного ряду, дослідити вибірку на однорідність. Побудуйте графіки динаміки зміни показників, визначте загальні тенденції. Зробіть висновки щодо угрупування країн за досліджуваним показником на основі критеріїв робастності.

Завдання 5. У табл. 2.10 наведені значення показника ВВП на душу населення (дол. США) для країн колишнього СНГ за три роки.

Таблиця 2.10

Вихідні дані

№ п/п	Країни	Рік			№ п/п	Країни	Рік		
		2014	2015	2016			2014	2015	2016
1	Азербайджан	7 857	8 771	9 579	8	Литва	18 186	19 877	16 669
2	Вірменія	5 324	5 808	5 014	9	Молдова	2 721	3 003	2 845
3	Білорусь	10 917	12 319	12 496	10	Таджикистан	1 653	1 783	1956
4	Грузія	4 683	4 899	4 762	11	Туркменістан	5 398	6 011	6 370
5	Казахстан	10 896	11 355	12 283	12	Узбекистан	2 396	2 639	2 855
6	Киргизстан	2 029	2 228	2 307	13	Україна	6 987	7 340	6 326
7	Латвія	17 080	18 302	15 226	14	Естонія	21 378	21 680	19 638

Необхідно на основі статистичної оцінки й аналізу даної сукупності розрахувати основні статистичні характеристики варіаційного ряду, дослідити вибірку на однорідність. Побудуйте графіки динаміки зміни показників, визначте загальні тенденції.

Розділ 3. Особливості класифікації багатовимірних об'єктів

3.1. Особливості застосування методів кластерного аналізу.
3.2. Термінологія кластерного аналізу.
3.3. Міри подібності.
3.4. Приклади розрахунку мір подібностей.

Ключові слова: кластер, кластерний аналіз, кластерний метод, багатовимірна класифікація, об'єкт, ознака, подібність, коефіцієнти подібності, властивості кластера, міри подібності, міри відстані.

Література: [1; 15; 17; 34; 36; 45].

3.1. Особливості застосування методів кластерного аналізу

В економічних дослідженнях число об'єктів дослідження може досягати кількох десятків чи навіть сотень; число ознак, що їх характеризують, також може обчислюватися десятками. Очевидно, безпосередній (візуальний) аналіз матриці даних за великої кількості об'єктів і ознак практично малоефективний. Через це виникають завдання укрупнення, концентрації вихідних даних, аналізу структури об'єктів дослідження. Вирішення цих завдань може здійснюватися за допомогою сучасних методів багатовимірної класифікації.

Методи багатовимірної класифікації дозволяють групувати об'єкти з урахуванням усіх істотних структурно-типологічних ознак і характеру розподілу об'єктів у заданій системі ознак. Така класифікація проводиться на основі прагнення зібрати в одну групу в деякому сенсі схожі об'єкти, причому так, щоб об'єкти з різних груп були за можливості несхожими.

Отже, методи багатовимірної класифікації використовуються для розділення сукупності об'єктів на однорідні групи. Водночас кожен об'єкт характеризується великим числом різних стохастично пов'язаних ознак.

В економічних дослідженнях використовують певні підходи до класифікації об'єктів і відповідні методи класифікації (рис. 3.1).

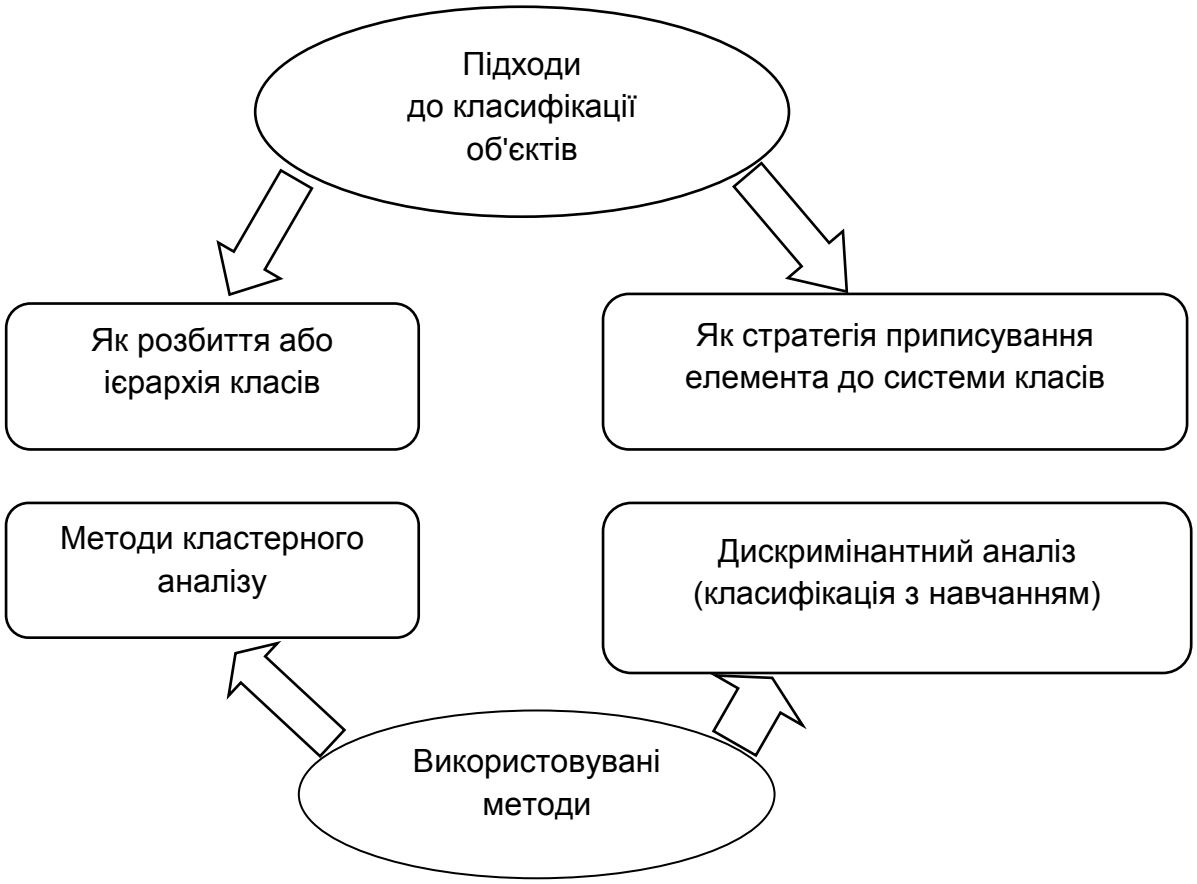


Рис. 3.1. **Підходи до класифікації об'єктів**

Розглянемо основні поняття кластерного аналізу.

Кластер (англ. – cluster) – група елементів, які характеризуються будь-якою загальною властивістю

Кластерний аналіз (КА) – множина обчислювальних процедур, що використовуються для класифікації (методи знаходження кластерів)

Кластерний метод – багатовимірна статистична процедура, що виконує збирання даних, які містять інформацію про вибір об'єктів, і потім упорядковує об'єкти в порівняно однорідні групи

Хронологія розвитку кластерного аналізу наведена на рис. 3.2.

1913 р. – польським антропологом К. Чекановським висунута ідея "структурної класифікації"

↓

1925 р. – радянський гідробіолог П. В. Терентьєв розробив "метод кореляційних плеяд" для групування корелюючих ознак, який дав поштовх розвитку методів групування за допомогою графів

↓

1939 р. – Р. Тріоном уперше був запропонований термін "кластерний аналіз". Слово "cluster" перекладається з англійської мови як "ґроно, кисть, пучок, група", тому спочатку цей вид аналізу називали "гроновим аналізом"

↓

поч. 1950-х рр. – спроби розроблення ієрархічного КА Р. Люїсом, Е.Фіксом і Дж. Ходжесом;
розроблення Р. Розенблаттом розпізнавального пристрою (персептрона), що поклало початок розвитку теорії "розпізнавання образів без вчителя"

↓

1963 р. – розробляється чисельна таксономія Р. Сокелом і П. Снітом;
розроблення алгоритмів за методами k-середніх Дж. Мак-Кіном, Г. Боллом і Д. Холлом;
розроблення алгоритмів за ієрархічним КА Г. Лансом, У. Уільямсом, Н. Джардайном

↓

1965 р. – М. І. Шлізенгером вперше точно була сформульована статистична й оптимізаційна задача навчання без вчителя і запропонований функціонал якості класифікації загального виду

↓

1960 – 1970 рр. – розроблення новосибірськими математиками Н. Г. Загоруйко, В. Н. Йолкіною і Г. С. Лбовим алгоритмів FOREL, BIGFOR, KRAB, NTTP, DRET, TRF та ін. (пакет ОТЕКС);
московськими математиками С. А. Айвазяном, І. С. Єнюковим і Б. Г. Міркіним створені програмні продукти ППСА та Клас-Майстер;
упроваджені алгоритми КА в пакети: SIGAMD, DataScope, STADIA, СОМИ, ПНП-БИМ, СОРРА-2, СИТО, SAS, SPSS, STATISTICA, BMDP, STATGRAPHICS, GENSTAT, S-PLUS

Рис. 3.2. **Хронологія розвитку кластерного аналізу**

Кластерний аналіз допомагає розв'язувати ряд задач у ході проведення економічних досліджень на основі багатовимірної класифікації. Основні задачі кластерного аналізу наведено на рис. 3.3.

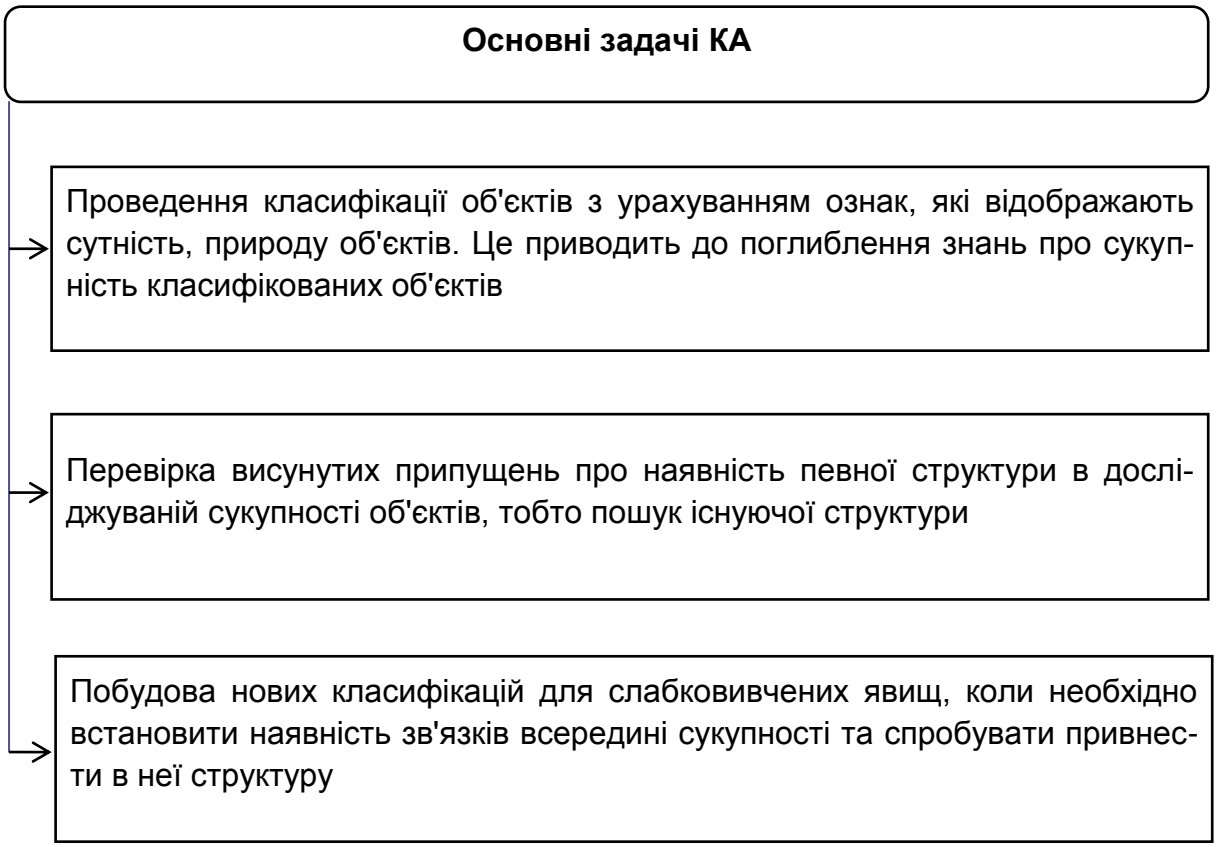


Рис. 3.3. **Основні задачі кластерного аналізу**

Кластерний аналіз застосовується у різноманітних економічних дослідженнях. Наприклад, у маркетингу це сегментація конкурентів і споживачів. У менеджменті: розбиття персоналу на різні за рівнем мотивації групи, класифікація постачальників, виявлення схожих виробничих ситуацій, за яких виникає брак. У фінансовому аналізі – для класифікації досліджуваних підприємств за рівнем фінансового стану та ін. Також кластерний аналіз успішно застосовується для проведення макроекономічних досліджень. Наприклад, для класифікації країн чи регіонів окремої країни за рівнем життя населення, за станом трудових ресурсів, за рівнем туристичної привабливості тощо. Він працює навіть тоді, коли даних мало та не виконуються вимоги нормальності розподілу випадкових величин та інші вимоги класичних методів статистичного аналізу.

3.2. Термінологія кластерного аналізу

Кластерний аналіз передбачає виділення компактних, віддалених один від одного груп об'єктів, відшукує "природне" розбиття сукупності на області скупчення об'єктів. Він використовується, коли вихідні дані подані у вигляді матриць близькості, або відстаней між об'єктами, або у вигляді точок у багатовимірному просторі. Найбільш поширені дані другого виду, для яких кластерний аналіз орієнтований на виділення деяких геометрично віддалених груп, всередині яких об'єкти близькі.

У кластерному аналізі використовується така **термінологія**.

Кластер – клас, таксон, згущення, група, пучок.

КА – таксономія, автоматична класифікація, стратифікація, класифікація без вчителя, розпізнавання з самонавчанням.

Об'єкт – подія, предмет, таксономічна одиниця.

Ознака – змінна, характеристика, властивість.

Матриця вихідних даних – матриця X розмірністю $n * m$:

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}$$

n – об'єкти (рядки матриці); m – ознаки (стовпці матриці).

Подібність – подоба, близькість, зв'язність, асоціативність.

Коефіцієнти подібності – міра подібності (коефіцієнт кореляції, міри відстані, коефіцієнт асоціативності, ймовірнісний коефіцієнт подібності).

Матриця подібності або матриця близькості – матриця D розмірністю $n * n$ або R розмірністю $m * m$.

У результаті застосування методів кластерного аналізу досліджувані об'єкти розбивають на певні кластери, які мають такі **властивості** (рис. 3.4): щільність; дисперсія; розмір; форма; віддільність.

Типи кластерних структур. Поняття "тип кластерної структури" не має формального визначення і залежить від нормування ознак і методу кластеризації. Приклади різних типів кластерних структур (кластери з центром; стрічкові кластери; кластери, що з'єднані перемичками; кластери, що перекриваються; кластери, що утворені не за подібністю, а за іншими типами регулярності; відсутність кластерів) наведено на рис. 3.5.

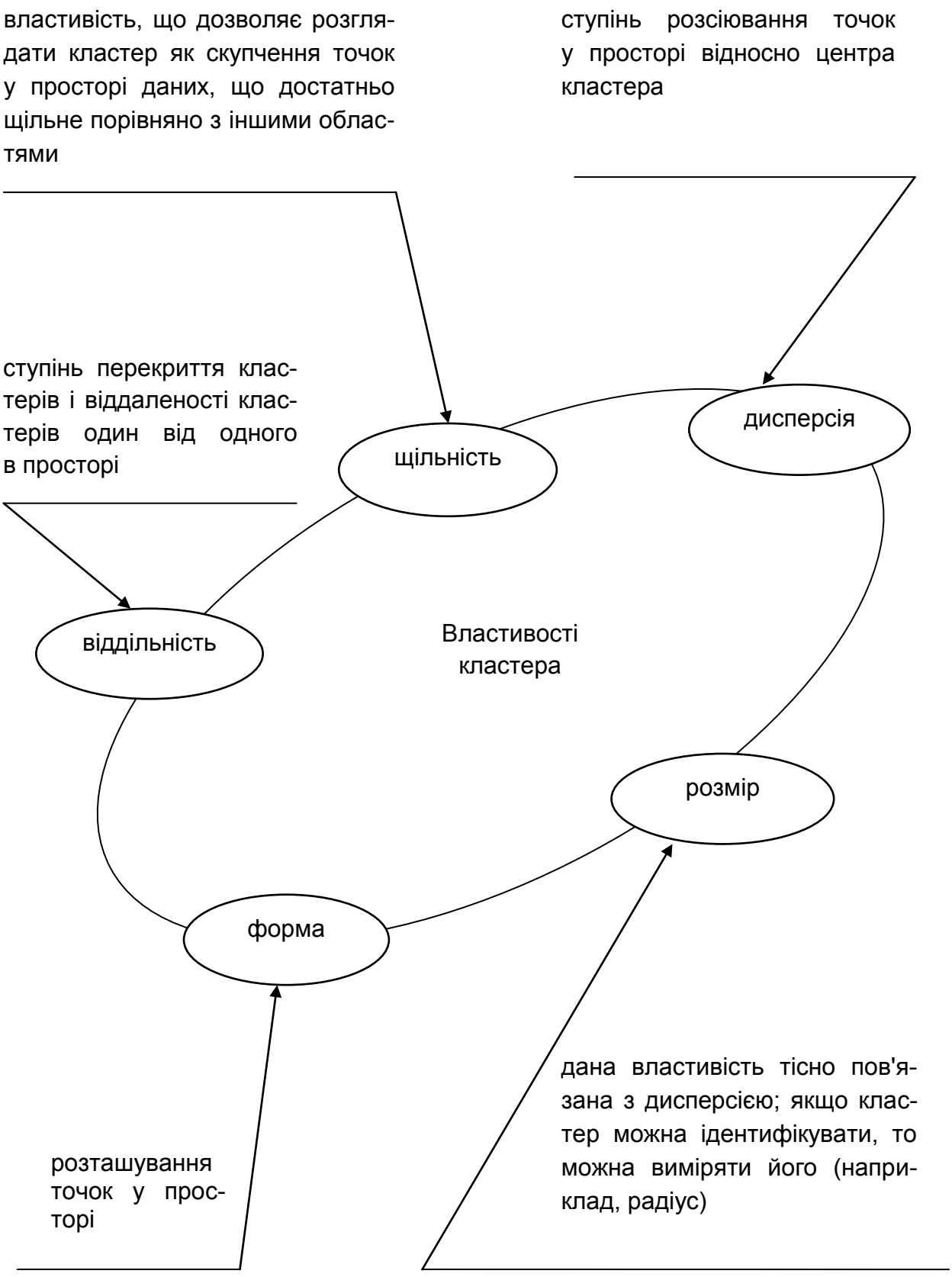
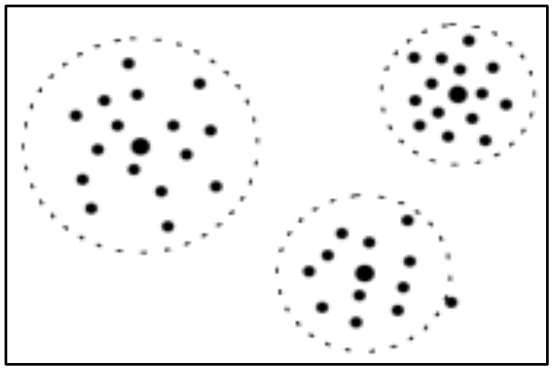
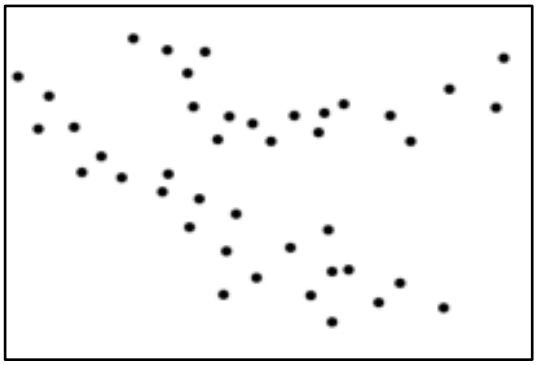


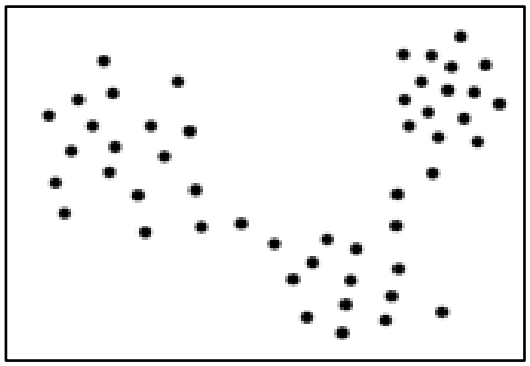
Рис. 3.4. **Властивості кластера**



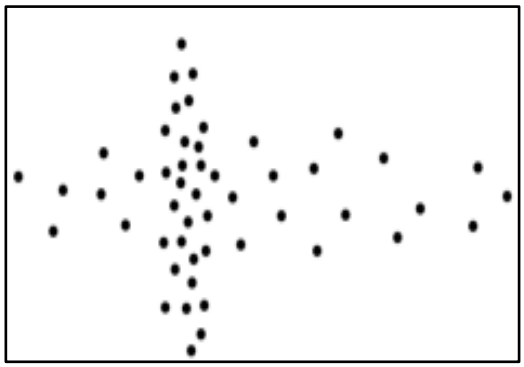
Кластери з центром



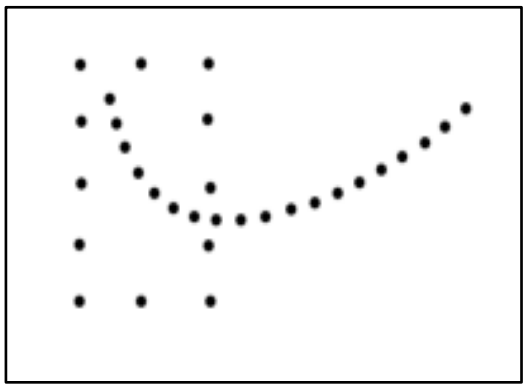
Стрічкові кластери



Кластери, що з'єднані
перемичками



Кластери,
що перекриваються



Кластери, що утворені
не за подібністю, а за іншими
типами регулярності

Відсутність кластерів

Рис. 3.5. **Типи кластерних структур**

Формальна постановка задачі кластеризації. Нехай X – множина об'єктів, Y – множина номерів (імен) кластерів. Задана функція відстані між об'єктами $\rho(x, x')$. Сформована навчальна вибірка об'єктів $X^m = \{x_1, \dots x_m\} \subset X$. Необхідно розбити вибірку на непересічні підмножини, які називають кластерами, так, щоб кожен кластер складався з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів істотно відрізнялися. Кожному об'єкту $x_i \in X^m$ приписується номер кластера Y_i.

У навчальній вибірці об'єкти можуть характеризуватися ознаками, які вимірюються в різних одиницях. Однак для кластерного аналізу ознаки повинні бути однорідними, тобто вимірюватися в порівняльних шкалах. Для цього здійснюється нормування початкових даних (рис. 3.6).

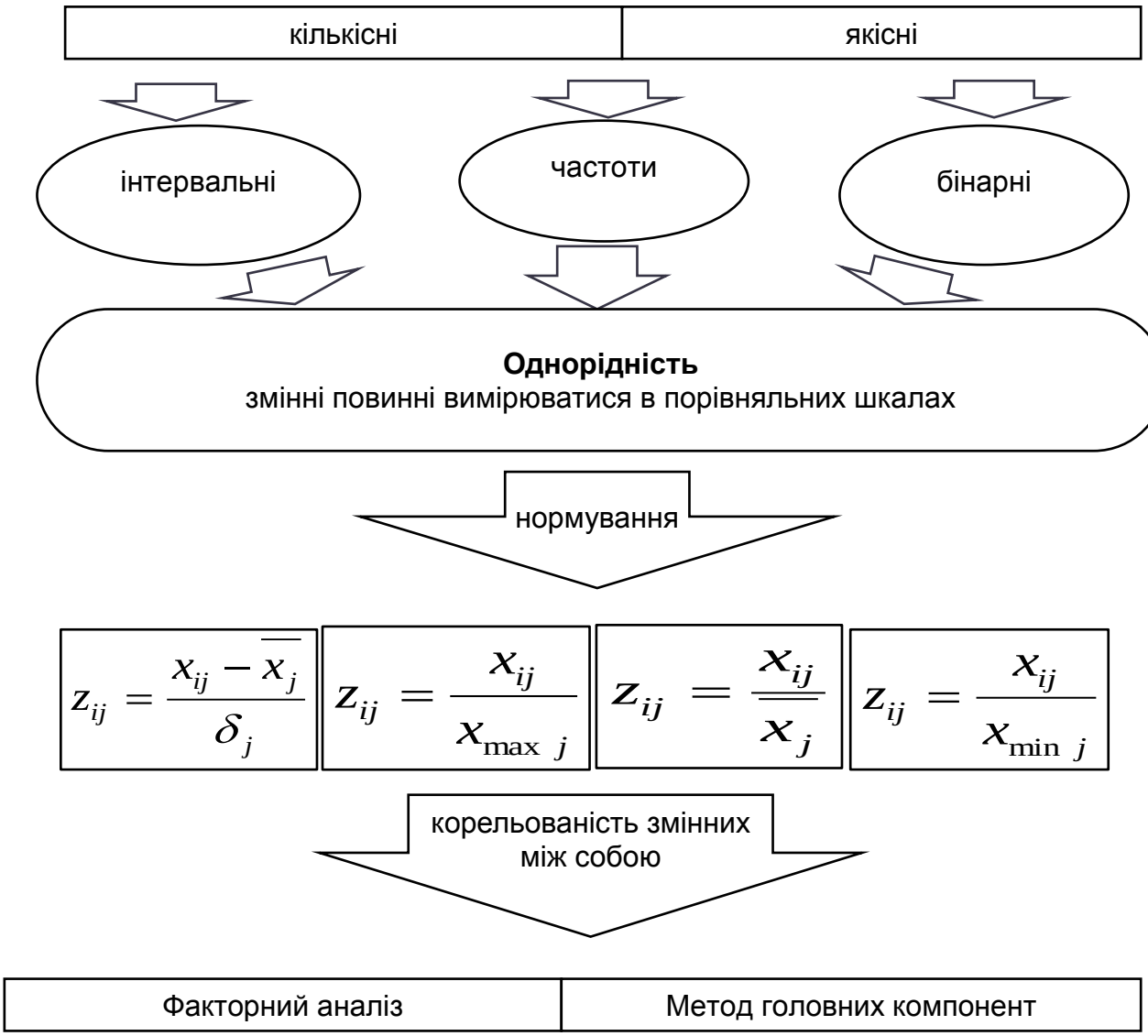


Рис. 3.6. **Вимоги до початкових даних**

Основні **етапи кластерного аналізу** такі.

Етап 1. Відбір вибірки для кластеризації (наявність апріорної інформації).

Етап 2. Визначення множини ознак, за якими будуть оцінюватися об'єкти.

Етап 3. Обчислення міри подібності між об'єктами відповідно до обраної метрики.

Етап 4. Групування об'єктів у кластери за допомогою тієї чи іншої процедури об'єднання.

Етап 5. Перевірка достовірності результатів КА.

3.3. Міри подібності

Важливим етапом кластерного аналізу є обчислення міри подібності між об'єктами, оскільки у ході здійснення кластеризації у кожен кластер повинні потрапити об'єкти з подібними характеристиками. У кожній конкретній задачі цей вибір здійснюється різним чином, з урахуванням головної мети дослідження, фізичної і статистичної природи використовуваної інформації тощо.

У кластерному аналізі можуть використовуватися міри подібності: коефіцієнти кореляції, міри відстані, коефіцієнти асоціативності, ймовірнісні коефіцієнти подібності (рис. 3.7). Кожен з цих показників має свої переваги та недоліки, які попередньо потрібно врахувати.

В 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.

У процесі кластеризації кожен об'єкт розглядається як точка в багатовимірному просторі ознак, що використовуються для його опису. Подібність і відмінність між точками знаходяться у відповідності з метричними відстанями між ними. Для цього однорідність об'єктів задається:

введенням правила обчислень відстаней $d(x_i, x_j)$ між будь-якою парою досліджуваних об'єктів $(x_1, x_2, \dots, x_n)$;

заданням деякої функції $r(x_i, x_j)$, що характеризує ступінь близькості i-го і j-го об'єктів.

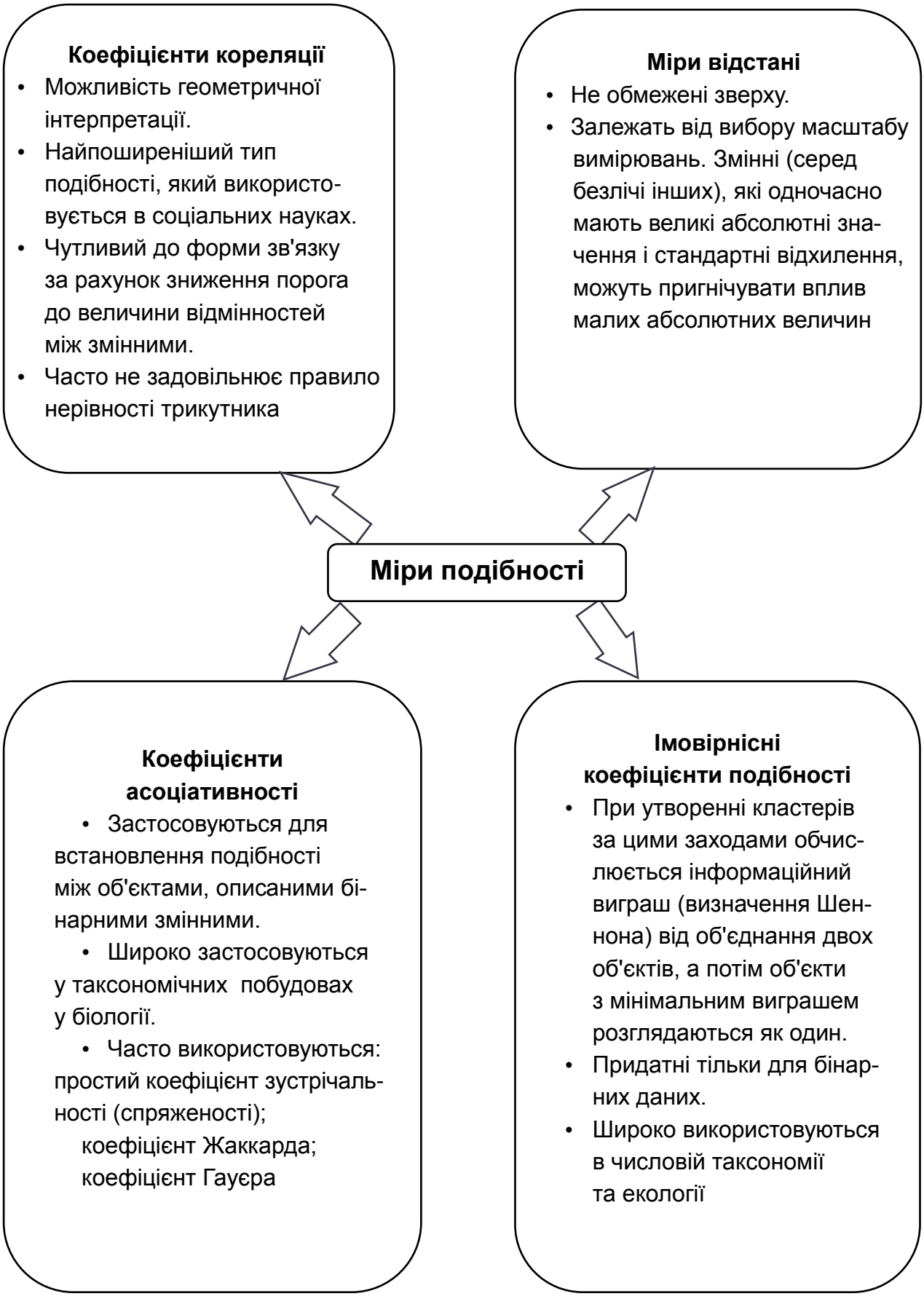


Рис. 3.7. **Міри подібності в кластерному аналізі**

Міра подібності є метрикою, якщо виконуються такі умови:

1. Симетрія. Відстань між об'єктами x і y повинна задовольняти:

$$d(x,y)=d(y,x)\geq 0$$

2. Нерівність трикутника. Відстань між об'єктами x, y і z:

$$d(x,y)\leq d(x,z)+d(y,z)$$

3. Розрізненість нетотожних об'єктів. Дано два об’єкта x і y; якщо

$$d(x,y)\neq 0, \text{ то } x\neq y$$

4. Нерозрізненість ідентичних об'єктів. Якщо x і x‘ ідентичні, то:

$$d(x, x')=0$$

Розглянемо міри подібності, які використовують у кластерному аналізі: коефіцієнт кореляції, ймовірнісний коефіцієнт подібності, міри відстані, коефіцієнти асоціативності.

Коефіцієнт кореляції

Природна міра подібності

$$r_{ij}=\frac{\sum_{h=1}^{N}(x_{hi}-m_i)(x_{hj}-m_j)}{\delta_i\delta_j}$$

$$-1\leq r_{ij}\leq 1$$

де x_{hi}, x_{hj} – значення h-ї ознаки для i-го та j-го об'єктів;

m_i, m_j, δ_i, δ_j – відповідні середні та середньоквадратичні відхилення для характеристик i і j.

$r_{ij} = -1$ – наявність зворотного тісного між об'єктами i і j;

$r_{ij} = 0$ – відсутність зв'язку між об'єктами i і j;

$r_{ij} = 1$ – наявність прямого тісного зв'язку між об'єктами i і j.

Ймовірнісний коефіцієнт подібності

Міра близькості ймовірнісного типу	$I_{ij} = \sum_{x,y} p_{xy} log \frac{p_{xy}}{p_x^i p_y^j}$,

де p_{xy} – ймовірність спільної появи ознак x і y;

p_x^i – ймовірність появи ознаки x в об'єкті i;

p_y^j – ймовірність появи ознаки y в об'єкті j.

Основою для проведення кластеризації є матриця відстаней між об'єктами. Існує кілька способів визначення відстані між кожними двома об'єктами. Вибір міри відстані істотно впливає на результат класифікації. Тому для отримання надійних результатів необхідно враховувати мету дослідження, змістову та статистичну природу вектора спостережень і наявні відомості про характер розподілу досліджуваних ознак. Крім того, після закінчення розрахунків слід перевіряти адекватність отриманої класифікаційної моделі.

У кластерному аналізі використовують такі міри відстані: евклідова відстань, "зважена" евклідова відстань, City-blok (мангетенська), відстань Мінковського.

Міри відстані

Евклідова

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2},$$

де d_{ij} – відстань між об'єктами i і j, і, $j = 1, \dots, n$; $k = 1, \dots, m$;

x_{ik} – значення k-ї змінної для i-го об'єкта;

x_{jk} – значення k-ї змінної для j-го об'єкта;

"Зважена" евклідова

$$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k \cdot \left(x_{ik} - x_{jk}\right)^2},$$

де w_k – вага k-ї ознаки, $0 \le w \le 1$

City-blok (мангетенська)

$$d_{ij} = \sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|,$$

Мінковського

$$d_{ij} = \left(\sum_{k=1}^{m}\left|x_{ik} - x_{jk}\right|^p\right)^{1/r},$$

де p, r – параметри, що визначені користувачем

Махаланобіса

$$d_{ij} = \left(X_i - X_j\right)^T S^{-1}\left(X_i - X_j\right),$$

де X_i, X_j – вектори значень *i*-го та *j*-го об'єктів;

S – загальна коваріаційна матриця

Коефіцієнти асоціативності

Для бінарних даних, змінні, що беруть участь в конструюванні цих заходів, описуються таблицею асоціативності, де "1" вказує на наявність змінної, а "0" – її відсутність

	1	0
1	a	b
0	c	d

Простий коефіцієнт зустрічності

$$S = \frac{(a+d)}{(a+b+c+d)}$$

Коефіцієнт Жаккара

$$J = \frac{a}{(a+b+c)}$$

Коефіцієнт Гауєра

$$s_{ij} = \sum_{k=1}^{p} S_{ijk} \Big/ \sum_{k=1}^{p} W_{ijk},$$

де S_{ijk} – "вклад" у подібність об'єктів, який враховує значущість ознаки k, у порівнянні об'єктів *i* і *j;*

W_{ijk} – ваговий коефіцієнт, який приймає значення 1 – якщо порівняння об'єктів за ознакою *k* варто враховувати і 0 – в іншому випадку.

3.4. Приклади розрахунку мір подібностей

Приклад 3.1. ***Евклідова відстань***

Чотири досліджуваних об'єкти характеризуються двома ознаками x_{i1}, x_{i2}. Вихідні дані подано в табл 3.1. Необхідно розрахувати матрицю Евклідових відстаней між об'єктами.

Таблиця 3.1

Вихідні дані

№ об'єкта / Ознака	1	2	3	4
x_{i1}	8	6	4	5
x_{i2}	13	10	12	14

Розв'язання

Матриця відстаней має такі властивості:

симетричність відносно діагоналі;

властивість тотожності відстані, яка в матриці відстаней проявляється в наявності 0 по діагоналі матриці, оскільки відстань об'єкта з самим собою очевидно дорівнює 0, а також у наявності нульових значень для абсолютно подібних об'єктів;

значення відстаней у матриці завжди невід'ємні.

Розраховуємо матрицю евклідових відстаней за формулою:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

Відстань між першим і другим об'єктом:

$$d_{12} = \sqrt{(8-6)^2 + (13-10)^2} = 3{,}61.$$

Відстань між першим і третім об'єктом:

$$d_{13} = \sqrt{(8-4)^2 + (13-12)^2} = 4{,}12.$$

Аналогічно розраховуємо відстані:

$$d_{14} = \sqrt{(8-5)^2 + (13-14)^2} = 3{,}16;$$

$$d_{23} = \sqrt{(6-4)^2 + (10-12)^2} = 2{,}83;$$

$$d_{24} = \sqrt{(6-5)^2 + (10-14)^2} = 4{,}12;$$

$$d_{34} = \sqrt{(4-5)^2 + (12-14)^2} = 2{,}24.$$

Діагональні елементи матриці відстаней дорівнюють 0, матриця симетрична відносно головної діагоналі, оскільки $d_{ij} = d_{ji}$. Тоді матриця евклідових відстаней між об'єктами матиме вигляд, поданий у табл. 3.2.

Таблиця 3.2

Матриця евклідових відстаней

$D =$

№ об'єкта	1	2	3	4
1	0	3,61	4,12	3,16
2	3,61	0	2,83	4,12
3	4,12	2,83	0	2,24
4	3,16	4,12	2,24	0

Приклад 3.2. ***Зважена евклідова відстань***

Чотири досліджуваних об'єкти характеризуються двома ознаками x_{i1}, x_{i2}. Вихідні дані подано в табл 3.3. Необхідно розрахувати матрицю зважених евклідових відстаней між об'єктами, якщо ознаки мають різну вагу: $w_1 = 0{,}6$; $w_2 = 0{,}4$.

Таблиця 3.3

Вихідні дані

Ознака \ № об'єкта	1	2	3	4
x_{i1}	10	14	9	8
x_{i2}	25	20	17	15

Розв'язання

Розраховуємо матрицю зважених евклідових відстаней за формулою:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k \cdot (x_{ik} - x_{jk})^2}.$$

Відстань між першим і другим об'єктом:

$$d_{12} = \sqrt{0{,}6 \cdot (10-14)^2 + 0{,}4 \cdot (25-20)^2} = 4{,}43.$$

Відстань між першим і третім об'єктом:

$$d_{13} = \sqrt{0{,}6 \cdot (10-9)^2 + 0{,}4 \cdot (25-17)^2} = 5{,}12.$$

Аналогічно розраховуємо відстані між іншими об'єктами:

$$d_{14} = \sqrt{0{,}6 \cdot (10-8)^2 + 0{,}4 \cdot (25-15)^2} = 6{,}51;$$

$$d_{23} = \sqrt{0{,}6 \cdot (14-9)^2 + 0{,}4 \cdot (20-17)^2} = 4{,}31;$$

$$d_{24} = \sqrt{0{,}6 \cdot (14-8)^2 + 0{,}4 \cdot (20-15)^2} = 5{,}62;$$

$$d_{34} = \sqrt{0{,}6 \cdot (9-8)^2 + 0{,}4 \cdot (17-15)^2} = 1{,}48.$$

Діагональні елементи матриці відстаней дорівнюють 0, матриця симетрична відносно головної діагоналі, оскільки $d_{ij} = d_{ji}$. Тоді матриця зважених евклідових відстаней між об'єктами матиме вигляд, поданий у табл. 3.4.

Таблиця 3.4

Матриця зважених евклідових відстаней

$D =$

№ об'єкта	1	2	3	4
1	0	4,43	5,12	6,51
2	4,43	0	4,31	5,62
3	5,12	4,31	0	1,48
4	6,51	5,62	1,48	0

Приклад 3.3. ***Відстань city-blok (мангетенська)***

Чотири досліджуваних об'єкти характеризуються двома ознаками x_{i1}, x_{i2}. Вихідні дані подано в табл. 3.5. Необхідно розрахувати матрицю відстаней city-blok між об'єктами.

Таблиця 3.5

Вихідні дані

№ об'єкта / Ознака	1	2	3	4
x_{i1}	20	22	28	30
x_{i2}	10	12	15	16

Розв'язання

Розраховуємо матрицю відстаней city-blok за формулою:

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|.$$

Відстань між першим і другим об'єктом:

$$d_{12} = |20 - 22| + |10 - 12| = 4.$$

Відстань між першим і третім об'єктом:

$$d_{13} = |20 - 28| + |10 - 15| = 13.$$

Аналогічно розраховуємо відстані між іншими об'єктами:

$$d_{14} = |20 - 30| + |10 - 16| = 16;$$

$$d_{23} = |22 - 28| + |12 - 15| = 9;$$

$$d_{24} = |22 - 30| + |12 - 16| = 12;$$

$$d_{34} = |28 - 30| + |15 - 16| = 3.$$

Діагональні елементи матриці відстаней дорівнюють 0, матриця симетрична відносно головної діагоналі, оскільки $d_{ij} = d_{ji}$. Тоді матриця відстаней city-blok між об'єктами матиме вигляд, поданий у табл. 3.6.

Таблиця 3.6

Матриця відстаней city-blok

$D =$

№ об'єкта	1	2	3	4
1	0	4	13	16
2	4	0	9	12
3	13	9	0	3
4	16	12	3	0

Приклад 3.4. ***Використання коефіцієнта Жаккара***

П'ять досліджуваних об'єктів характеризуються трьома ознаками x_1, x_2, x_3. Вихідні дані подано в табл. 3.7. Необхідно розрахувати матрицю відстаней між об'єктами на основі коефіцієнта Жаккара.

Таблиця 3.7

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5
x_1	0	1	0	0	1
x_2	1	1	1	1	1
x_3	1	1	1	0	1

Розв'язання

Коефіцієнт подібності Жаккара є бінарною мірою подібності. Він обчислюється за формулою:

$$J = \frac{a}{a + b + c}.$$

Таблиця спряженості 2 × 2:

	1	0
1	a	b
0	c	d

Із таблиці спряженості отримуємо значення коефіцієнта Жаккара між відповідними елементами (матрицю подібності).

Для першого та другого об'єктів матриця спряженості має вигляд:

	1	0
1	2	0
0	1	0

$$J_{12} = \frac{a}{a+b+c} = \frac{2}{2+0+1} = 0{,}67.$$

Аналогічно розраховуємо коефіцієнти для інших пар об'єктів (табл. 3.8).

Таблиця 3.8

Коефіцієнти Жаккара

Об'єкти	1	2	3	4	5
1	0	0,67	1,00	0.5	0,67
2	0,67	0	0,67	0,33	1,00
3	1,00	0,67	0	0,5	0,67
4	0,5	0,33	0.5	0	0,33
5	0,67	1,00	0,67	0,33	0

Відстані між об'єктами визначають за формулою:

$$d_{ij} = 1 - J_{ij}.$$

Матриця відстаней між об'єктами подана в табл. 3.9.

Таблиця 3.9

Матриця відстаней між об'єктами на основі коефіцієнтів Жаккара

Об'єкти	1	2	3	4	5
1	0	0,33	0,00	0,50	0,33
2	0,33	0	0,33	0,67	0,00
3	0,00	0,33	0	0,50	0,33
4	0,50	0,67	0,50	0	0,67
5	0,33	0,00	0,33	0,67	0

Отримані матриці відстаней є основою для проведення кластеризації. У кожній конкретній задачі вибір міри відстані здійснюється різним чином, з урахуванням головної мети дослідження, фізичної і статистичної природи використовуваної інформації тощо.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. Для розв'язання яких задач застосовують методи кластерного аналізу?
2. Дайте визначення поняття "кластер".
3. Назвіть основні властивості кластера.
4. Які ви знаєте типи кластерних структур?
5. Назвіть основні етапи кластерного аналізу.
6. З якою метою здійснюється нормування початкових даних у кластерному аналізі?
7. Які міри подібності використовують у кластерному аналізі?
8. Назвіть міри відстані, які використовують найчастіше.
9. У чому полягає відмінність між евклідовою відстанню та зваженою евклідовою відстанню?
10. За виконання яких умов міра подібності є метрикою?

Тестові завдання

1. *Кластерний аналіз – це статистичний метод, який дозволяє:*

а) вивчати відмінності між двома та більше групами об'єктів за декількома змінними одночасно;

б) знаходити групи схожих об'єктів у вибірці даних.

2. Методи кластерного аналізу відносять до групи методів класифікації:

а) з навчанням;

б) без навчання.

3. В якості міри подібності в кластерному аналізі використовують:

а) тільки міру відстані;

б) тільки коефіцієнти кореляції;

в) міру відстані, коефіцієнти кореляції, коефіцієнти асоціативності.

4. Елементи головної діагоналі матриці відстаней між об'єктами дорівнюють:

а) 1;

б) 0;

в) дисперсії;

г) коефіцієнтам кореляції.

5. Ступінь перекриття кластерів і розташування один від одного в просторі характеризує властивість:

а) щільності;

б) віддільності;

в) розміру.

6. Яку властивість має кластер, якщо всі точки знаходяться поблизу його центру ваги:

а) має форму;

б) є повним;

в) має розмір?

7. Міра подібності є метрикою, якщо:

а) $d(x,y) = d(y,x) \geq 0$, де $x, y -$ об′єкти, $d(x,y) -$ відстань;

б) $d(x,y) \neq 0$;

в) $d(x,x) = 0;\ \ d(x,y) = d(y,x) \geq 0;\ \ \ d(x,y) \leq d(x,z) + d(z,y)$.

8. Відстань city-block знаходять за формулою:

а) $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$;

б) $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})}$;

в) $d_{ij} = \sum_{k=1}^{m}|x_{ik} - x_{jk}|$.

9. Евклідову відстань знаходять за формулою:

а) $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$;

б) $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})}$;

в) $d_{ij} = \sum_{k=1}^{m}|x_{ik} - x_{jk}|$.

10. Коефіцієнт кореляції – це:

а) міра відстані;

б) міра подібності;

в) міра відмінності.

Практичні завдання

Завдання 1. У табл. 3.10 наведено вісім об'єктів, які характеризуються двома показниками x_1 і x_2. Необхідно розрахувати матрицю евклідових відстаней між об'єктами.

Таблиця 3.10

Вихідні дані

Показники	1	2	3	4	5	6	7	8
x_1	119,4	121,0	16,6	114,2	115,8	15,2	17,9	117,5
x_2	16,6	18,1	15,5	19,4	23,2	16,7	15,7	15,2

Завдання 2. У табл. 3.11 наведено вісім об'єктів, які характеризуються двома показниками x_1 і x_2. Необхідно розрахувати матрицю евклідових відстаней між об'єктами.

Таблиця 3.11

Вихідні дані

Показники	1	2	3	4	5	6	7	8
x_1	73,2	60,2	63,7	70,6	95,1	75,8	93,4	50,5
x_2	12,2	11,6	1,6	13,7	16,1	11,1	16,5	1,2

Завдання 3. У табл. 3.12 наведено вісім об'єктів, які характеризуються двома показниками x_1 і x_2. Необхідно розрахувати матрицю зважених евклідових відстаней між об'єктами, $w_1 = 0,3$, $w_2 = 0,7$.

Таблиця 3.12

Вихідні дані

Показники	1	2	3	4	5	6	7	8
x_1	114,4	116,0	11,6	19,2	110,8	11,2	12,9	112,5
x_2	12,6	14,1	12,5	15,4	19,2	11,7	12,7	12,2

Завдання 4. У табл. 3.13 наведено вісім об'єктів, які характеризуються двома показниками x_1 і x_2. Необхідно розрахувати матрицю відстаней city-blok (мангетенських) між об'єктами.

Таблиця 3.13

Вихідні дані

Показники	1	2	3	4	5	6	7	8
x_1	133,2	120,2	133,7	120,6	115,1	145,8	153,4	137,5
x_2	24,2	20,6	16,6	36,7	35,1	72,1	56,5	54,2

Завдання 5. У табл. 3.14 наведено п'ять об'єктів, які характеризуються трьома показниками x_1, x_2, x_3.

Таблиця 3.14

Вихідні дані

Номер об'єкта	x_1	x_2	x_3
1	1	1	0
2	1	1	1
3	0	0	1
4	0	1	0
5	1	1	1

Розрахуйте матрицю відстаней між об'єктами за допомогою коефіцієнта Жаккара.

Розділ 4. Методи кластерного аналізу. Класифікація без навчання

4.1. Класифікація кластер-процедур. Ієрархічні агломеративні й ітеративні кластер-процедури.

4.2. Нечіткі методи класифікації.

4.3. Критерії якості класифікації методами кластерного аналізу.

Ключові слова: ієрархічні методи кластерного аналізу; агломеративні методи; дивізимні методи; неієрархічні методи; метод "найближчого сусіда"; метод "далекого сусіда"; метод середнього зв'язку; метод Уорда; дендрограма; метод k-середніх; нечітка кластеризація; метод дендритів; метод куль; метод пошуку згущень; функціонал якості.

Література: [1; 15; 17; 34; 36; 59; 60].

4.1. Класифікація кластер-процедур. Ієрархічні агломеративні й ітеративні кластер-процедури

Об'єднання схожих об'єктів у групи може бути здійснене різними способами. Виділяють певні групи методів кластерного аналізу (рис. 4.1).

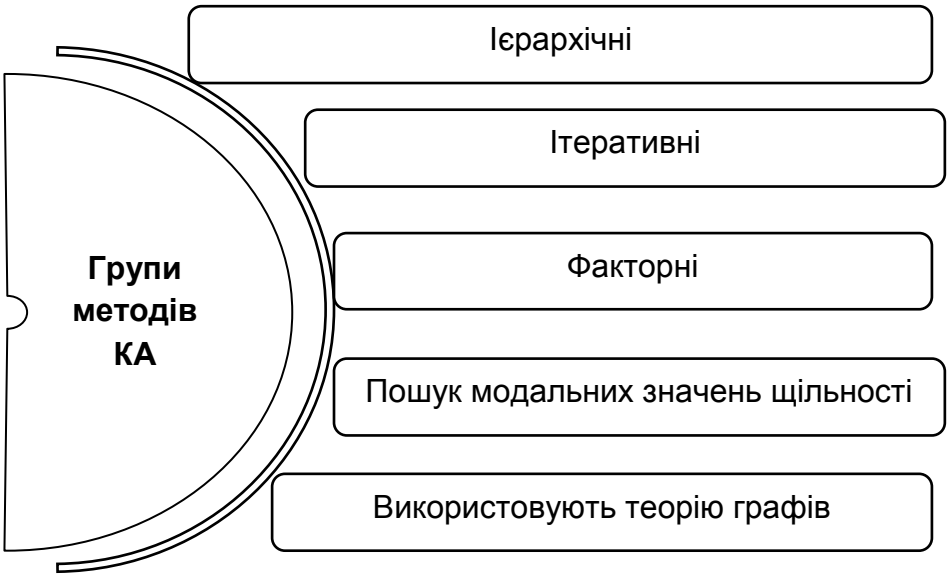


Рис. 4.1. **Групи методів кластерного аналізу**

Класифікація методів кластерного аналізу

1. За способом обробки даних:
ієрархічні (агломеративні, дивізимні);
неієрархічні.
2. За способом аналізу даних:
чіткі;
нечіткі.
3. За кількістю застосування алгоритмів кластеризації:
з одноетапною кластеризацією;
з багатоетапною кластеризацією.
4. За можливістю розширення обсягу оброблюваних даних:
масштабовані;
немасштабовані.
5. За часом виконання кластеризації:
потокові (on-line);
не потокові (off-line).

Розглянемо особливості різних методів кластеризації (рис. 4.2).

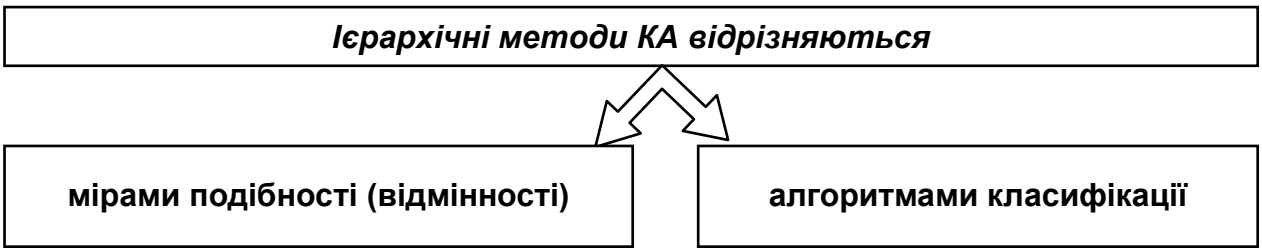


Рис. 4.2. **Особливості ієрархічних методів кластерного аналізу**

Найбільш поширеними є **ієрархічні методи**, серед яких розрізняють агломеративний і дивізимний методи.

Головна ідея агломеративного методу: на першому кроці кожен об'єкт вважається окремим кластером. Два найбільш близьких об'єкта об'єднуються, і утворюється новий кластер. Процедура триває, доки всі об'єкти не будуть об'єднані в один кластер.

Головна ідея дивізимного методу: спочатку всі об'єкти належать одному кластеру. Від цього кластера відокремлюються групи схожих між собою об'єктів. Так, на кожному кроці кількість кластерів зростає, а міра відстані між класами зменшується.

Візуалізація кластерної структури наведена на рис. 4.3.

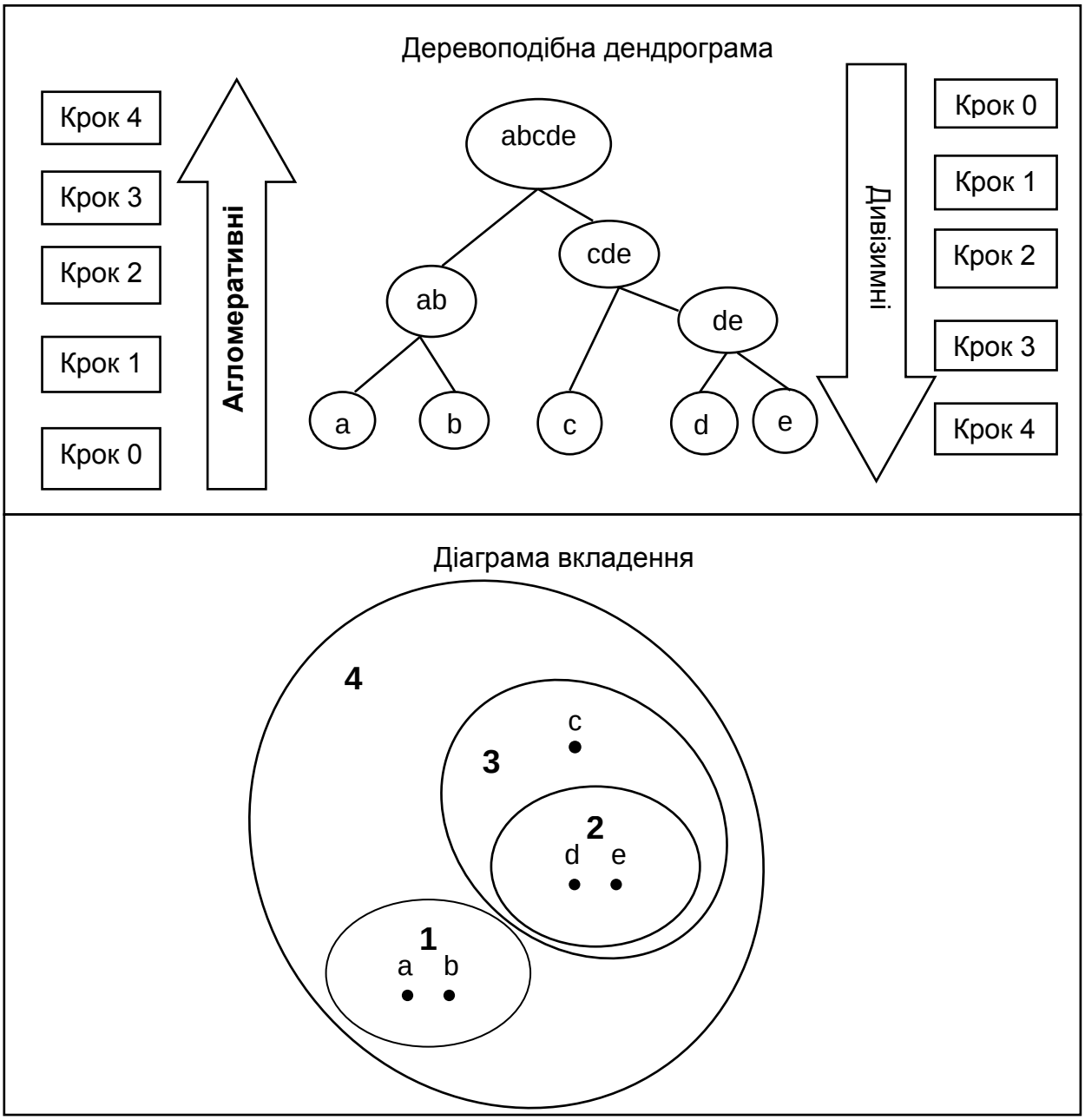


Рис. 4.3. **Візуалізація кластерної структури у застосуванні ієрархічних методів**

Алгоритм агломеративної ієрархічної кластеризації:

1) нормування вихідних даних;

2) розрахунок матриці відстаней або матриці мір подібності;

3) знаходиться пара найближчих кластерів. За обраним алгоритмом об'єднуються ці два кластери. Новому кластеру присвоюється менший з номерів об'єднувальних кластерів;

4) кроки 2, 3 і 4 повторюються, поки всі об'єкти не будуть об'єднані в один кластер або до досягнення заданого "порогу" подібності.

Алгоритм дивізимної ієрархічної кластеризації:

1) нормування вихідних даних;

2) розрахунок матриці відстаней або матриці мір подібності;

3) знаходиться пара найдальніших об'єктів *ni, nj;*

4) оцінюється відстань об'єктів, що залишилися, до виділених об'єктів *ni, nj*. Визначається, до якого з об'єктів *ni* або *nj* вони ближче знаходяться;

5) близькі об'єкти об'єднуються в кластер. Так початковий єдиний кластер розбивається на два;

6) кроки 3, 4 і 5 повторюються, поки всі об'єкти не будуть розділені на кластери.

Зауваження. *Дивізимний алгоритм не вимагає перерахунку матриці відстаней на кожному кроці класифікації на відміну від агломеративного, що сприяє зниженню трудомісткості розрахунків.*

Методи групування

Метод "найближчого сусіда"/одиничного зв'язку

Ступінь подібності оцінюється за ступенем подібності між найбільш схожими (найближчими) об'єктами цих кластерів (рис. 4.4).

d_1 і d_2 – евклідові відстані; якщо $d_1 > d_2$, то S увійде в кластер U по d_2

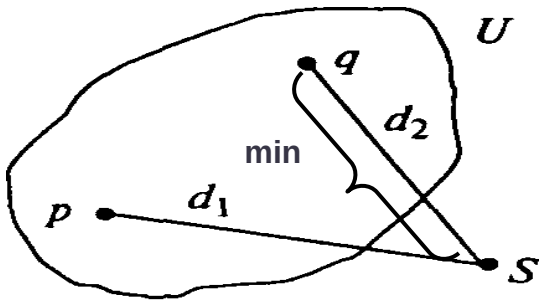


Рис. 4.4. **Графічне зображення методу "найближчого сусіда"**

Метод "дальнього сусіда" / повного зв'язку

Ступінь подібності оцінюється за ступенем подібності між найбільш віддаленими (несхожими) об'єктами цих кластерів (рис. 4.5).

d_1 і d_2 – евклідові відстані;
якщо $d_1 > d_2$, то S увійде
в кластер U по d_1

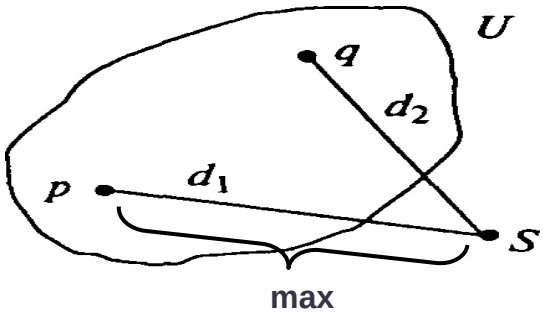


Рис. 4.5. **Графічне зображення методу "дальнього сусіда"**

Метод середнього зв'язку

Ступінь подібності оцінюється як середня величина ступенів подібності між об'єктами кластерів (рис. 4.6).

d_1 і d_2 – евклідові відстані,
S увійде в кластер U по
$d=0,5(d_1+d_2)$.

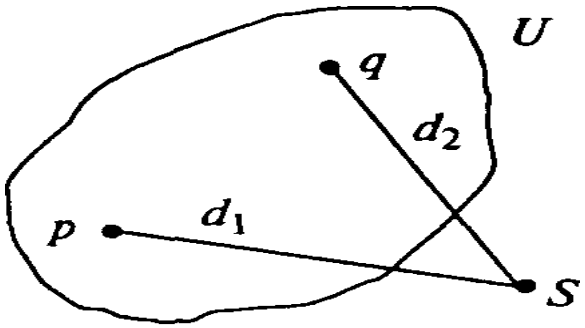


Рис. 4.6. **Графічне зображення методу середнього зв'язку**

Метод медіанного зв'язку

Відстань визначається як відстань від центру будь-якого кластера S до середини відрізка, що з'єднує центри нового кластера, який об'єднав об'єкти p і q (рис. 4.7).

S увійде в кластер U по d,
де d – відстань від центра
S до U(p,q)

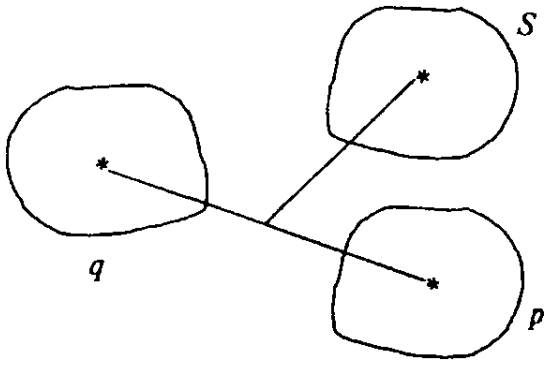


Рис. 4.7. **Графічне зображення методу медіанного зв'язку**

Відстань між кластерами розраховується за формулою Ланса – Уїльямса:

$$d_{US} = \alpha_p d_{ps} + \alpha_q d_{qs} + \beta d_{pq} + \gamma \left| d_{ps} - d_{qs} \right|. \qquad (4.1)$$

Значення параметрів формули для різних методів групування подані в табл. 4.1.

Таблиця 4.1

Значення параметрів для формули Ланса – Уїльямса

Методи	Параметри
Найближчого сусіда	$\alpha_p = \alpha_q = 0{,}5, \quad \beta = 0, \quad \gamma = -0{,}5$
Дальнього сусіда	$\alpha_p = \alpha_q = 0{,}5, \quad \beta = 0, \quad \gamma = 0{,}5$
Середнього зв'язку	$\alpha_p = \frac{n_p}{n_p + n_q}, \quad \alpha_q = \frac{n_q}{n_p + n_q}, \quad \beta = \gamma = 0$
Центроїдний	$\alpha_p = \frac{n_p}{n_p + n_q}, \quad \alpha_q = \frac{n_q}{n_p + n_q}, \quad \beta = -\frac{n_p}{n_p + n_q} \cdot \frac{n_q}{n_p + n_q}$
Медіанного зв'язку	$\alpha_p = \alpha_q = 0{,}5, \quad \beta = -0{,}25, \quad \gamma = 0$

Приклад 4.1. ***Метод "найближчого сусіда" (одиничного зв'язку)***

Вхідні дані: аналізуються чотири об'єкта, які характеризуються двома показниками X_{i1},X_{i2} (табл. 4.2).

Таблиця 4.2

Вхідні дані

№ об'єкта / Ознака	1	2	3	4
x_{i1}	5	6	5	10
x_{i2}	10	12	13	9

Розраховуємо матрицю відстаней між об'єктами за формулою (4.1) (табл. 4.3).

Таблиця 4.3

Матриця відстаней

$D1 =$

№ об'єкта	1	2	3	4
1	0	2,24	3,00	5,10
2	2,24	0	1,41	5,00
3	3,00	1,41	0	6,40
4	5,10	5,00	6,40	0

Знаходимо мінімальну відстань між об'єктами $d_{23} = 1{,}41$. Отже, необхідно об'єднати об'єкти 2 і 3 в один кластер, обираючи в якості нової відстані мінімальну між кластерами, що об'єднуються $\min(2{,}24; 3{,}00) = 2{,}24$ і т. д. Перераховуємо матрицю відстаней (табл. 4.4).

Таблиця 4.4

Матриця відстаней за методом "найближчого сусіда"

$D2 =$

№ об'єкта	1	2,3	4
1	0	2,24	5,10
2,3	2,24	0	5,00
4	5,10	5,00	0

Аналогічно знаходимо мінімальну відстань між об'єктами $d_{1(23)} = 2{,}24$. Отже, необхідно об'єднати об'єкти 1 і (2, 3) в один кластер, обираючи в якості нової відстані мінімальну між кластерами, що об'єднуються: $\min(5{,}10; 5{,}00) = 5{,}00$. Перераховуємо матрицю відстаней (табл. 4.5).

Таблиця 4.5

Матриця відстаней за методом "найближчого сусіда"

$D3 =$

№ об'єкта	1,2,3	4
1,2,3	0	5,00
4	5,00	0

Дендрограма класифікації методом "найближчого сусіда" наведена на рис. 4.8.

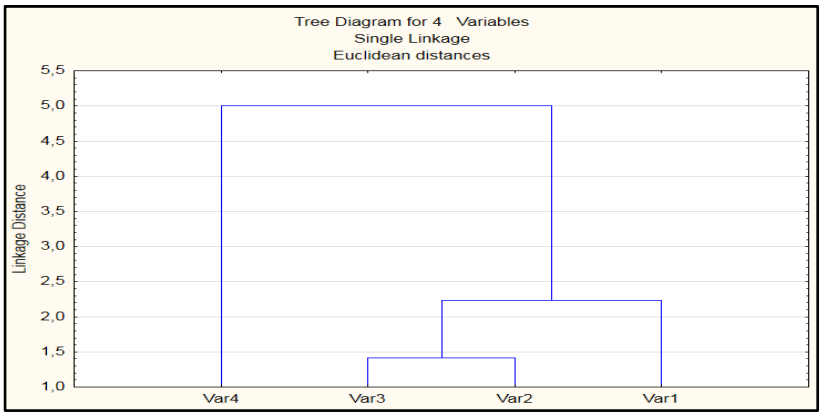


Рис. 4.8. **Дендрограма класифікації методом "найближчого сусіда"**

Приклад 4.2. ***Дивізимна кластеризація***

Вхідні дані: аналізуються вісім об'єктів, які характеризуються двома показниками X_{i1},X_{i2} (табл. 4.6).

Таблиця 4.6

Вхідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_{i1}	0,98	1,24	-1,12	0,12	0,39	-1,35	-0,91	0,66
x_{i2}	-0,35	0,21	-0,76	0,69	2,11	-0,32	-0,69	-0,88

Розраховуємо матрицю відстаней між об'єктами за формулою (4.1) (табл. 4.7).

Знаходимо максимальну відстань між об'єктами, це буде відстань між третім і п'ятим об'єктами $d_{35} = 2{,}54$.

Оцінимо відстані об'єктів, що залишилися до третього та п'ятого об'єктів:

$d_{13} < d_{15}$ – об'єкт n_1 ближче до n_3;

$d_{23} < d_{25}$ – об'єкт n_2 ближче до n_3;

$d_{43} > d_{45}$ – об'єкт n_4 ближче до n_5;

$d_{63} < d_{65}$ – об'єкт n_6 ближче до n_3;

$d_{73} < d_{75}$ – об'єкт n_7 ближче до n_3;

$d_{83} < d_{85}$ – об'єкт n_8 ближче до n_3.

Таблиця 4.7

Матриця відстаней

$D1 =$

№ об'єкта	1	2	3	4	5	6	7	8
1	0	0,49	1,20	0,99	2,08	1,28	1,07	0,48
2	0,49	0	1,53	0,74	1,66	1,49	1,40	0,97
3	1,20	1,53	0	1,39	2,54	0,39	0,13	0,98
4	0,99	0,74	1,39	0	1,2	1,17	1,29	1,35
5	2,08	1,66	2,54	1,2	0	2,25	2,45	2,50
6	1,28	1,49	0,39	1,17	2,25	0	1,50	1,49
7	1,07	1,40	0,13	1,29	2,45	1,50	0	1,57
8	0,48	0,97	0,98	1,35	2,50	1,49	1,57	0

Отримано два кластери: $S_1\{1,2,3,6,7,8\}$ і $S_2\{4,5\}$.

Кластер $S_2\{4,5\}$ розділяємо на два кластери на відстані $d_{45} = 1,2$.

Аналізуємо відстані між об'єктами в кластері $S_1\{1, 2, 3, 6, 7, 8\}$. Для цього треба створити нову матрицю відстаней, яка не містить об'єктів 4 і 5, уже розділених на окремі кластери (табл. 4.8).

Таблиця 4.8

Матриця відстаней

$D2 =$

№ об'єкта	1	2	3	6	7	8
1	0	0,49	1,20	1,28	1,07	0,48
2	0,49	0	1,53	1,49	1,40	0,97
3	1,20	1,53	0	0,39	0,13	0,98
6	1,28	1,49	0,39	0	1,50	1,49
7	1,07	1,40	0,13	1,50	0	1,57
8	0,48	0,97	0,98	1,49	1,57	0

Знаходимо максимальну відстань між об'єктами, це буде відстань між другим та третім об'єктами $d_{35} = 2,54$.

Оцінимо відстані об'єктів, що залишилися до другого та третього об'єктів:

$d_{12} < d_{13}$ – об'єкт n_1 ближче до n_2;

$d_{62} > d_{63}$ – об'єкт n_6 ближче до n_3;

$d_{72} > d_{73}$ – об'єкт n_7 ближче до n_3;

$d_{82} < d_{83}$ – об'єкт n_8 ближче до n_2.

Кластер S_1{1,2,3,6,7,8} розділився на два: S_{11}{1, 2, 8} і S_{12}{3, 6, 7}.

Аналізуємо відстані між об'єктами в отриманих кластерах.

Створимо нові матриці відстаней для кластерів S_{11} ($D3$) и S_{12} ($D4$) (табл. 4.9; 4.10).

Таблиця 4.9

Матриця відстаней

$D3 =$

№ об'єкта	1	2	8
1	0	0,49	0,48
2	0,49	0	0,97
8	0,48	0,97	0

Max $d_{28} = 0,97$.

Оцінимо відстані об'єкта n_1 до n_2 і n_8: $d_{12} < d_{18}$ – об'єкт n_1 ближче до n_2. Тоді n_8 від'єднується від кластера S_{11} {1, 2, 8} на відстані 0,97, а об'єкти n_1 і n_2 розділяються на відстані 0,49.

Таблиця 4.10

Матриця відстаней

$D4 =$

№ об'єкта	3	6	7
3	0	0,39	0,13
6	0,39	0	1,50
7	0,13	1,50	0

Max $d_{67} = 1,5$.

Оцінимо відстані об'єкта n_3 до n_6 і n_7: $d_{36} > d_{37}$ – об'єкт n_3 ближче до n_7. Тоді n_6 від'єднується від кластера S_{12}{3, 6, 7} на відстані 1,5, а об'єкти n_3 і n_7 розділяються на відстані 0,13.

Усі об'єкти розділені на окремі кластери, що свідчить про закінчення процедури розбиття.

Побудуємо дендрограму (рис. 4.9).

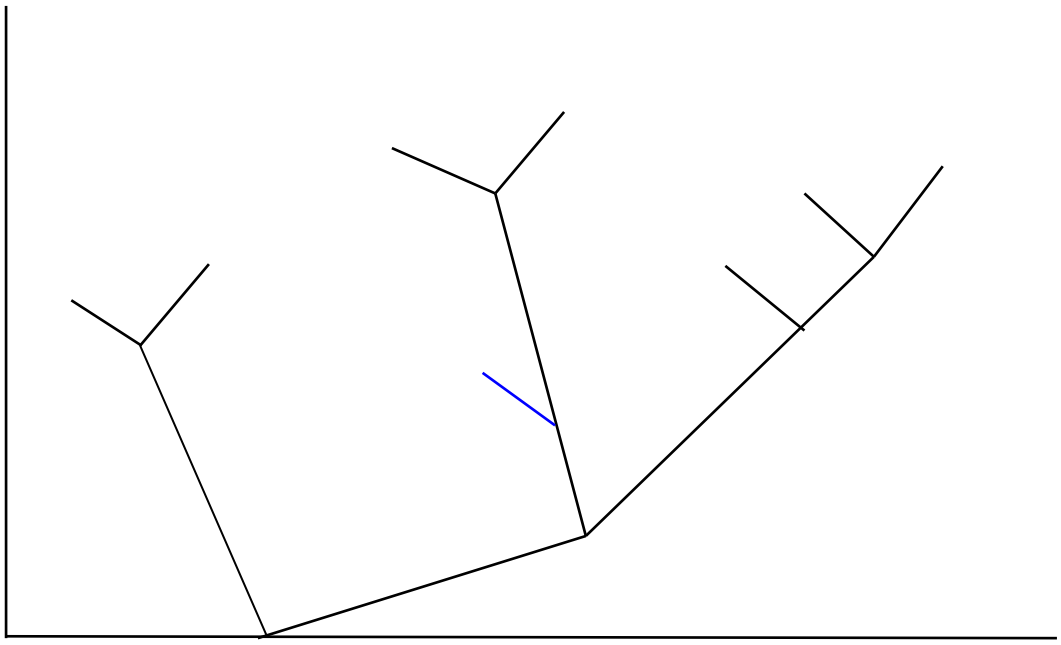


Рис. 4.9. **Дендрограма класифікації дивізимним методом**

Ward's method (Метод Уорда) (1963)

Оптимізація мінімальної дисперсії всередині кластерів:

цільова функція (сума квадратів відхилень (СКВ) → min;

утворюються кластери, приблизно рівних розмірів, мають гіперсферичну форму.

Алгоритм метода Уорда:

1) нормування вихідних даних;

2) розрахунок матриці відстаней або матриці мір подібності;

3) знаходиться пара найближчих кластерів, вони об'єднуються. Новому кластеру присвоюється менший з номерів об'єднувальних кластерів. Для утвореного кластера розраховується СКВ за формулою:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(x_{ij} - \overline{x}_{jk}\right)^2; \qquad (4.2)$$

4) надалі на кожному кроці роботи алгоритму об'єднуються ті об'єкти або кластери, які дають найменший приріст величини V_k.

Процедури 2, 3 і 4 повторюються до тих пір, поки всі об'єкти не будуть об'єднані в один кластер або до досягнення заданого "порогу".

Приклад 4.3 ***Метод Уорда***

1. Вхідні дані: аналізуються п'ять об'єктів, які характеризуються двома показниками X_{i1}, X_{i2} (табл. 4.11).

Таблиця 4.11

Вхідні дані

№ об'єкта / Ознака	1	2	3	4	5
x_{i1}	21	18	15	13	25
x_{i2}	0,3	0,8	0,6	0,7	0,9

2. Розраховуємо матрицю евклідових відстаней між об'єктами (табл. 4.12).

Таблиця 4.12

Матриця евклідових відстаней

$D =$

№ об'єкта	1	2	3	4	5
1	0	3,04	6,01	8,01	4,04
2	3,04	0	3,01	5,00	7,00
3	6,01	3,01	0	2,00	10,00
4	8,01	5,00	2,00	0	12,00
5	4,04	7,00	10,00	12,00	0

Знаходимо мінімальну відстань між об'єктами, це буде відстань між третім і четвертим об'єктами $d_{34} = 2{,}00$.

3. Для кластера $S_{(3,4)}$ визначаємо СКВ за формулою (4.2):

$$V_k = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(x_{ij} - \overline{x}_{j3}\right)^2,$$

де $\overline{x}_{j3}$ – середнє значення *j*-ї ознаки в кластері S_3;

$x_{sr(13)} = 14$; $x_{sr\ 23} = 0{,}65$;

$V_3 = [(15 - 14)^2 + (0{,}6 - 0{,}65)^2] + [(13 - 14)^2 + (0{,}7 - 0{,}65)^2] = 2{,}005$.

4. Вирішуємо питання: який новий об'єкт може бути на наступному кроці приєднаний до третього кластеру або які кластери можна об'єднати?

а) якщо об'єднувати об'єкти, що залишилися (1, 2, 5):

для 1 і 2 об'єктів $V_k = 1{,}625$;

для 1 і 5 об'єктів $V_k = 8{,}0081$;

для 2 і 5 об'єктів $V_k = 12{,}25$;

б) якщо приєднувати до кластеру S_3 почергово кожен з решти об'єктів (1, 2, 5):

для 1 об'єкта і S_3 $V_k = 34{,}75$;

для 2 об'єкта і S_3 $V_k = 12{,}69$;

для 5 об'єкта і S_3 $V_k = 82{,}71$.

5. Кластеризація спостережуваних об'єктів методом Уорда подана на рис. 4.10.

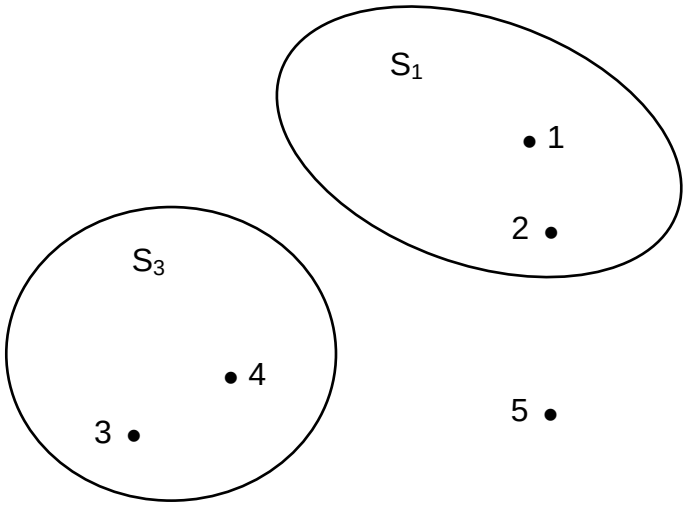


Рис. 4.10. **Кластеризація спостережуваних об'єктів**

6. Кроки 2, 3 і 4 повторюються, поки всі об'єкти не будуть об'єднані в один кластер або до досягнення заданого "порогу" подібності.

Види **ітеративних (ітераційних) методів кластеризації** схематизовані на рис. 4.11.

Характеристика ітеративних методів кластеризації наведена на рис. 4.12.

Рис. 4.11. **Види ітеративних (ітераційних) методів кластеризації**

К-середніх	Заснований на мінімізації функціонала сумарної вибіркової дисперсії розкиду елементів щодо центрів тяжіння кластерів
Fuzzy C-means	Знаходить компактні кластери сферичної форми
PAM (c-means + c-medoids)	Є модифікацією алгоритму k-середніх, алгоритмом k-медіани. Менш чутливий до шумів і викидів даних, ніж алгоритм c-means. Ефективний для невеликих баз даних
CLOPE	Призначення: кластеризація величезних наборів категорійних даних
Метод пошуку згущень	Заснований на ідеї об'єднання в один кластер об'єктів в областях їх найбільшого згущення. Не вимагає задання числа кластерів

Рис. 4.12. **Особливості ітеративних методів кластеризації**

Розглянемо детальніше особливості і основні етапи ітеративних методів кластеризації.

Метод К-середніх

Особливості: метод *k*-середніх зручний для обробки великих статистичних сукупностей. Необхідно попередньо задати кількість кластерів k.

Алгоритм метода К-середніх

1. Виділяємо початкові центри k кластерів: $C^{(0)}_1$, $C^{(0)}_2$, …, $C^{(0)}_k$ (як зважене середнє за кожним показником). Кожному кластеру привласнюють одиничну вагу.

2. Знаходять відстані від точки x_{k+1} до центрів кластерів, побудованих на кроці 1. Точку x_{k+1} відносять до кластеру, відстань до якого мінімальна.

3. Розраховують новий центр ваги $C^{(1)}_i$. Вагу кластера *i* збільшують на одиницю.

4. Повтор кроків 2 і 3, поки всі крапки разом не будуть віднесені до якогось з *k* кластерів.

5. Точки $x_1, x_2, \ldots, x_n$ знову приєднуються до *k* отриманим кластерам, ваги продовжують накопичуватися.

6. Розбиття порівнюється з попереднім, поки $C^{(m+1)}_i <> C^{(m)}_i$.

7. Класифікація закінчена.

Приклад 4.4. ***Метод К-середніх***

1. Вхідні дані: аналізуються шість об'єктів, які характеризуються двома показниками X_i, Y_i (табл. 4.13).

Таблиця 4.13

Вхідні дані

№ об'єкта / Ознака	1	2	3	4	5	6
X_i	5	6	5	10	11	10
Y_i	10	12	13	9	9	7

2. Розбиваємо вихідну сукупність на два кластери : $S^{(0)}_1$ (1, 2, 3, 4) і $S^{(0)}_2$ (5, 6).

3. Знаходимо центри тяжіння (точки еталона) для кожної з сформованих груп $S^{(0)}_1$ і $S^{(0)}_2$:

$$E^{(0)}_1 = (e^{(0)}_{x1};\ e^{(0)}_{y1}),\ E^{(0)}_2 = (e^{(0)}_{x2};\ e^{(0)}_{y2}),$$

де $e^{(0)}_{x1} = (5 + 6 + 5 + 10) / 4 = 6{,}5;$

$e^{(0)}_{y1} = (10 + 12 + 13 + 9) / 4 = 11.$

Таким чином, $E^{(0)}_1 = (6{,}5;\ 11)$ і $E^{(0)}_2 0 = (10{,}5;\ 8)$.

4. Знаходимо відстань від точки еталона до об'єкта (табл. 4.14).

Таблиця 4.14

Відстань від точки еталона до об'єкта

Об'єкт	E_1	E_2	Клас
1	1,803	5,852	S_1
2	1,118	6,021	S_1
3	2,5	7,433	S_1
4	4,031	1,118	S_2
5	4,924	1,118	S_2
6	5,315	1,118	S_2

$E^{(1)}_2 < E^{(1)}_1;$

$4 = > S^{(1)}_2 (4,5,6).$

5. Знаходимо центри тяжіння нового розбиття елементів:

$S^{(1)}_1 (1, 2, 3) : E^{(1)}_1 = (5{,}33;\ 11{,}66);$

$S^{(1)}_2 (5, 6, 4) : E^{(1)}_2 = (10{,}33;\ 8{,}33).$

6. Крок 4 для нового розбиття (табл. 4.15).

Таблиця 4.15

Відстань від точки еталона до об'єкта

Об'єкт	E_1	E_2	Клас
1	1,7	5,588	S_1
2	0,745	5,676	S_1
3	1,374	7,087	S_1
4	5,375	0,745	S_2
5	6,263	0,943	S_2
6	6,6	1,374	S_2

7. Оскільки два останні розбиття збігаються, обчислювальні процедури закінчуються:

$S^{(1)}_1$ (1, 2, 3) і $S^{(1)}_2$ (5, 6, 4).

4.2. Нечіткі методи класифікації

Окремим видом методів кластерного аналізу є нечіткі методи класифікації.

Чітка (непересічна) кластеризація – кластеризація, в якій кожен об'єкт належить тільки до одного кластеру.

Нечітка кластеризація – кластеризація, за якої для кожного об'єкта визначається дійсне значення, що показує ступінь приналежності до кластеру – функція приналежності (ФП). Ступінь приналежності визначається відстанню від об'єкта до відповідних кластерних центрів. Даний алгоритм ітераційно обчислює центри кластерів і нові ступені приналежності об'єктів.

Нечітка кластеризація методом k-середніх здійснюється за допомогою певних алгоритмів (рис. 4.13).

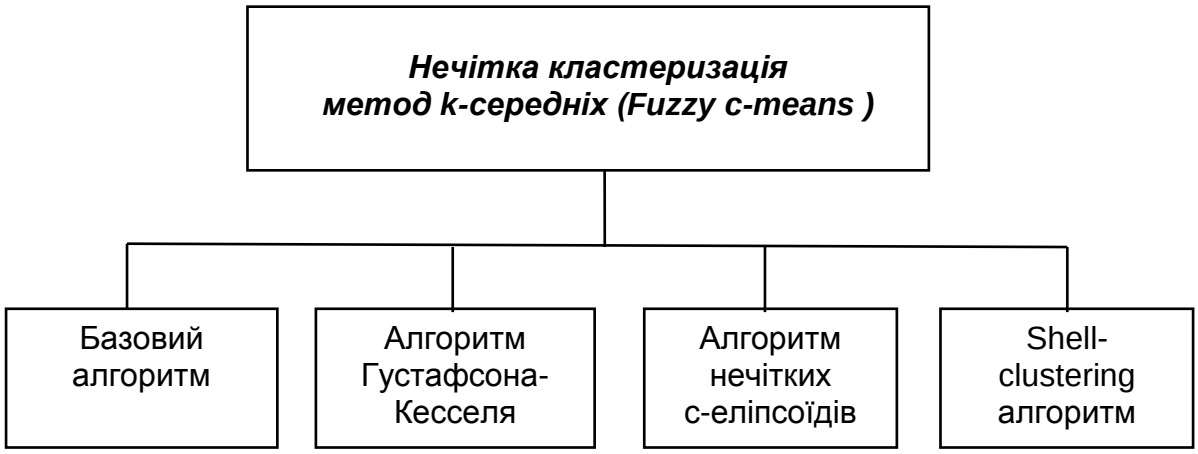


Рис. 4.13. **Алгоритми нечіткої кластеризації методом k-середніх**

Розглянемо детальніше основні етапи реалізації алгоритмів нечітких k-середніх.

Базовий алгоритм нечітких k-середніх. Fuzzy c-means (Бездек)

Алгоритм знаходить компактні кластери сферичної форми.

Дано: множина *K* вхідних векторів x_k, *N* виділяючих кластерів c_j.

Кожний вхідний вектор (x_k) належить будь-якому кластеру (c_j) з ФП μ_{jk}, що приймає будь-яке значення з інтервалу [0;1], де *j* – номер кластера, а *k* – номер вхідного вектора.

Умови нормування для μ_{jk}:

$$\sum_{j=1}^{N} \mu_{jk} = 1, \forall\ k = 1, .., K. \tag{4.3}$$

Мета кластеризації – мінімізація суми всіх зважених евклідових відстаней:

$$\sum_{j=1}^{N} = \sum_{k=1}^{K} \left((\mu_{jk})^q \left\| x_k - c_j \right\|^2 \right) \rightarrow \min, \tag{4.4}$$

де *q* – фіксований параметр, що задається перед ітераціями.

Для цього треба знайти перші похідні по μ_{jk} і по c_j , прирівняти їх до 0 і розв'язати систему рівнянь. Отримуємо:

формула центру кластера (зважений центр гравітації):

$$c_j = \frac{\sum_{j=1}^{N} (\mu_{jk})^q \cdot x_k}{\sum_{j=1}^{N} (\mu_{jk})^q}; \tag{4.5}$$

ФП (елементи матриці нечіткого розбиття):

$$\mu_{jk} = \frac{1}{\left(d_{jk}^2 \cdot \sum_{j=1}^{N} \frac{1}{d_{jk}^2} \right)^{\frac{1}{q-1}}}. \tag{4.6}$$

Базовий алгоритм нечітких k-середніх – FCM

1. Ініціалізація (встановлення параметрів):

кількість кластерів N, $2 < N < K$;

міра відстаней d (евклідова);

фіксований параметр q (~1,5);

ε – параметр зупинки алгоритму;

початкова (на нульовій ітерації) матриця приналежності об'єктів x_k до c_j $U^{(0)} = (\mu_{jk})^{(0)}$ – випадково генерується.

2. Розрахунок центрів кластерів за формулою (4.5).

3. Розрахунок відстані між об'єктами та центрами кластерів:

$$d_{jk} = \sqrt{\left\| x_k - c_j \right\|^2}. \qquad (4.7)$$

4. Перерахунок елементів матриці нечіткого розбиття, якщо

$$d_{jk} > 0 \rightarrow \mu_{jk} = \frac{1}{\left(d_{jk}^2 \cdot \sum_{j=1}^{N} \frac{1}{d_{jk}^2} \right)^{\frac{1}{q-1}}}, \text{ якщо } d_{jk} = 0 \rightarrow \mu_{jk} = \begin{cases} 1, j = k \\ 0, j \neq k \end{cases}. \quad (4.8)$$

Перевіряємо умову $\left\| U^{(t)} - U^{(t-1)} \right\|^2 < \varepsilon$: якщо "так", то алгоритм закінчений; якщо "ні" – перейти до кроку 2.

Метод пошуку згущень "форель"

Робота алгоритму полягає у переміщені гіперсфери заданого радіуса в просторі класифікаційних ознак з метою пошуку локальних згущень точок. Алгоритм методу включає такі етапи.

1. Обчислення матриці відстаней (або матриці мір подібності) між об'єктами.

2. Вибір об'єкта (на основі попереднього аналізу точок та їх околів), який є первинним центром першого кластера – гіперсфери заданого радіуса R : $min\ d_{ij} \leq R \leq max\ d_{ij}$.

3. Визначення сукупності точок, що потрапили всередину цієї сфери та для них обчислюються координати нового центру (вектор середніх значень ознак).

4. Крок 3 повторюється до тих пір, поки черговий перерахунок координат центра сфери приводить до такого ж результату, як і на попередньому кроці.

5. Переміщення сфери припиняється, а точки, що потрапили в неї, утворюють кластер і з подальшого процесу кластеризації виключаються.

6. Для всіх точок, що залишилися, повторюють процедури 2–5.

Метод дендритів

Дендрит – ламана лінія, яка може розгалужуватися; вона з'єднує кожні дві точки досліджуваної сукупності, проте не може містити замкнутих ламаних ліній (контурів).

Критерій оптимальності дендрита – min суми відстаней, це забезпечує більшу схожість сусідніх елементів.

Алгоритм побудови дендритів (нелінійне впорядкування)

1. Нормування вихідних даних.

2. Розрахунок матриці відстаней або матриці мір подібності.

3. Аналіз рядків матриці відстаней і вибір пари елементів з мінімальною відстанню між собою.

4. З'єднання спільних елементів (об'єднання) і отримання груп, які називають об'єднаннями першого роду.

5. Процес об'єднання відбувається до тих пір, поки не будуть об'єднані всі пари елементів досліджуваної сукупності.

Приклад 4.5. ***Метод дендритів***

1. Вхідні дані: аналізуються шість об'єктів, які характеризуються двома показниками x_{i1}, x_{i2} (табл. 4.16).

2. Розраховуємо матрицю евклідових відстаней між об'єктами (табл. 4.17).

Таблиця 4.16

Вхідні дані

№ об'єкта / Ознака	1	2	3	4	5	6
x_{i1}	5	6	5	10	11	10
x_{i2}	10	12	13	9	9	7

Таблиця 4.17

Матриця евклідових відстаней

$D =$

№ об'єкта	1	2	3	4	5	6
1	0,00	2,24	3,00	5,10	6,08	5,83
2	2,24	0,00	1,41	5,00	5,83	6,40
3	3,00	1,41	0,00	6,40	7,21	7,81
4	5,10	5,00	6,40	0,00	1,00	2,00
5	6,08	5,83	7,21	1,00	0,00	2,24
6	5,83	6,40	7,81	2,00	2,24	0,00

3. У матриці D вибираємо елементи з min відстанню між собою (рис. 4.14).

1-2	2,24
2-3	1,41
3-2	1,41
4-5	1,0
5-4	1,0
6-4	2,0

Рис. 4.14. **Елементи з мінімальною відстанню**

4. Отримуємо дві групи із загальними елементами: (1 – 2 – 3) і (4 – 5 – 6).

5. За відповідними зв'язками будуємо дендрит (рис. 4.15).

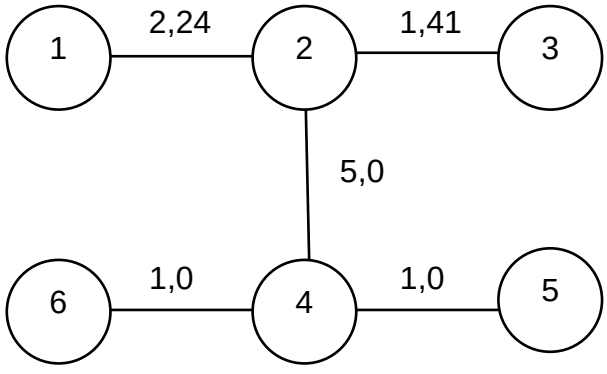


Рис. 4.15. **Дендрит**

Метод куль

Алгоритм методу куль

1. Нормування вихідних даних.

2. Розрахунок матриці відстаней або матриці мір подібності.

3. Визначення радіуса кулі $\rho = \max_S \min_U d_{SU}$, де d_{SU} – відстань між S-м і U-м об'єктами.

4. Для кожного багатовимірного об'єкта утворюється куля з радіусом (R) і визначається кількість точок, що потрапили в кожну кулю.

5. Визначається куля з максимальним числом елементів. Якщо є кілька таких куль, то розглядається куля, що найближче розташована до осі координат.

6. З матриці виключаються елементи, що потрапили в першу виділену групу.

Процедура повторюється, поки не будуть розглянуті всі елементи.

Приклад 4.6. ***Метод куль***

1. Вхідні дані: аналізуються шість об'єктів, які характеризуються двома показниками x_{i1}, x_{i2} (табл. 4.18).

Таблиця 4.18

Вхідні дані

№ об'єкта / Ознака	1	2	3	4	5	6
x_{i1}	5	6	5	10	11	10
x_{i2}	10	12	13	9	9	7

2. Розраховуємо матрицю евклідових відстаней між об'єктами та визначаємо радіус кулі за формулою (рис. 4.16):

$$\rho = \max_S \min_U d_{SU}.$$

$R_1 = \{\rho(X_i, X_j)\} =$

	1	2	3	4	5	6	min	r=2,24	елементи
1	0,00	2,24	3,00	5,10	6,08	5,83	2,24	*	1
2	2,24	0,00	1,41	5,00	5,83	6,40	1,41	1	2-3
3	3,00	1,41	0,00	6,40	7,21	7,81	1,41	1	3-2
4	5,10	5,00	6,40	0,00	1,00	2,00	1,00	2	5-4-6
5	6,08	5,83	7,21	7,21	0,00	2,24	1,00	1	5-4
6	5,83	6,40	7,81	2,00	2,24	0,00	2,00	1	6-4

Рис. 4.16. **Визначення радіуса кулі**

Виходячи з умови ($d_{US} < \rho$) елемент 1 становитиме одноелементну кулю.

3. Знаходимо відстань об'єктів до точки початку координат (рис. 4.17).

d_{10}	11,18034	d_{40}	13,45362
d_{20}	13,41641	d_{50}	14,21267
d_{30}	13,92839	d_{60}	12,20656

Рис. 4.17. **Відстань об'єктів до точки початку координат**

Отримали три кластери: S_1 (п4, п5, п6); S_2 (п2, п3); S_3 (п1).

4.3. Критерії якості класифікації методами кластерного аналізу

Заключним етапом процедури кластеризації є оцінювання якості отриманої класифікації. Використання різних методів кластерного аналізу для тої самої сукупності призводить до різних класифікацій об'єктів (різне число кластерів, різна ступінь близькості об'єктів). Істотний вплив на характеристики кластерної структури надають:

набір ознак кластеризації;

тип алгоритму кластеризації (метод кластерного аналізу);

вибір міри подібності між об'єктами.

Виникає проблема вибору найбільш якісної класифікації об'єктів, яка вирішується за допомогою критеріїв якості класифікації об'єктів. Міру якості класифікації прийнято називати функціоналом, або критерієм якості. Найкращим за обраним функціоналом вважають таку класифікацію об'єктів, в якій досягається екстремальне (максимальне або мінімальне) значення функціоналу якості.

Функціонал (або критерій якості) – деяка міра якості класифікації. Прагне до екстремуму (min / max) (рис. 4.18).

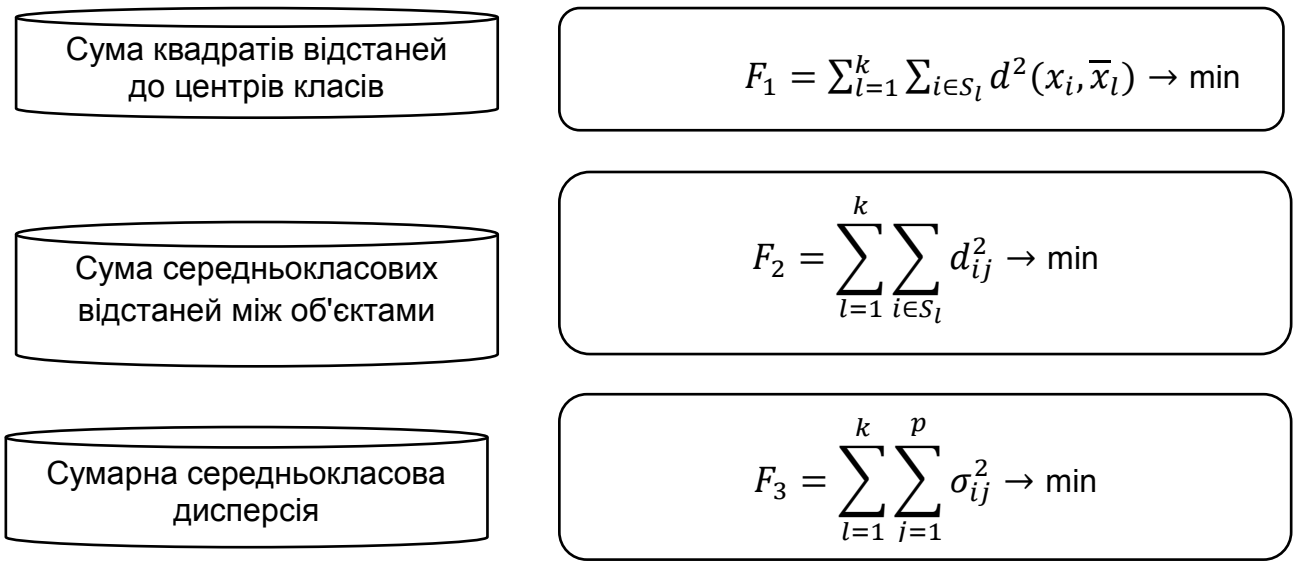


Рис. 4.18. **Критерії якості класифікації**

Формульні символи на рис. 4.18 мають такі значення:

l – номер кластера ($l = 1,2, \dots, k$);

$\overline{x}_l$ – центр l-го кластера;

x_i – вектор значень змінних для i-го об'єкта, що входить в l-й кластер;

$d^2(x_i, \overline{x}_l)$ – квадрат відстані між i-м об'єктом і центром l- го кластера;

d_{ij}^2 – квадрат всередньокласових відстаней між об'єктами;

σ_{ij}^2 – дисперсія j-ї змінної в кластері S_l.

Для оцінювання якості розбиття за мірою віддаленості кластерів один від одного використовують середні міжкласові відстані. Для перевірки гіпотези про рівність векторів середніх для багатовимірних сукупностей використовується критерій Хоттелінга.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. Назвіть основні групи методів кластерного аналізу.
2. У чому полягає відмінність агломеративних і дивізимних методів?
3. Назвіть основні етапи методу куль.

4. Охарактеризуйте метод дендритів.
5. Наведіть критерії оцінювання якості класифікації.
6. У чому полягає особливість ієрархічних кластер-процедур?
7. Наведіть алгоритм методу К-середніх.
8. Назвіть етапи алгоритму побудови дендрита.
9. Назвіть особливості нечітких методів класифікації.
10. Визначте особливості застосування критеріїв якості класифікації.

Тестові запитання

1. Агломеративний метод, в якому відстань між кластерами дорівнює відстані між двома найближчими об'єктами кластера, використовує процедуру:

а) далекого сусіда;

б) середнього зв'язку;

в) найближчого сусіда.

2. При використанні дивізимних методів на першому етапі всі об'єкти:

а) належать до одного кластера;

б) розглядаються як самостійні кластери.

3. Метод Уорда передбачає, що на першому кроці:

а) усі об'єкти входять в один кластер;

б) кожен кластер складається з одного об'єкта.

4. Метод k-середніх належить до групи методів:

а) далекого сусіда;

б) ієрархічних;

в) дивізимних;

г) ітеративних.

5. Радіус ρ у методі куль визначається за формулою:

а) $\rho = \max_S \min_U d_{SU}$, де d_{SU}-відстань між S-тим та U-тим об'єктами;

б) $\rho = \max_S d_{SU}$;

в) $\rho = \min_U d_{SU}$.

6. Елемент включається в кулю із заданим радіусом, якщо:

а) $d_{SU} < \rho$;

б) $d_{SU} > \rho$;

в) $d_{SU} = \rho$.

7. У методі медіанного зв'язку ступінь подібності оцінюється:

а) за ступенем схожості між найбільш схожими об'єктами кластерів;

б) за ступенем схожості між найбільш віддаленими об'єктами кластерів;

в) як відстань від центра кластера до середини відрізка, який з'єднує центри кластерів p і q.

8. У результаті якого варіанту нормування середнє кожного показника дорівнюватиме нулю, а дисперсія – одиниці:

а) $x'_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}$;

б) $x'_{ij} = x_{ij} - \overline{x_j}$;

в) $x'_{ij} = \frac{x_{ij}}{\sigma_j}$?

9. Процедура k-середніх дозволяє:

а) максимізувати внутрішньогрупову дисперсію та мінімізувати міжгрупову дисперсію;

б) мінімізувати внутрішньогрупову дисперсію та максимізувати міжгрупову дисперсію;

в) мінімізувати міжгрупову дисперсію;

г) максимізувати внутрішньогрупову дисперсію.

10. Для оцінювання якості класифікації використовують функціонал:

а) суми квадратів відстаней до центрів класів;

б) суми дисперсій ознак;

в) функціонал Фішера.

Практичні завдання

Завдання 1. У табл. 4.19 наведені вісім країн, які характеризуються двома показниками туристичної привабливості: x_1 – площа, яку займають туристичні ресурси, км^2; x_2 – інвестиції в основний капітал готелів і ресторанів, млн дол. Необхідно провести кластеризацію країн, описаних двома показниками, за методом найближчого сусіда. В обчисленні відстаней використовуйте просту евклідову відстань. Результати кластеризації подайте у вигляді дендрограм. Зробіть висновки щодо наявності природного розбиття сукупності об'єктів на кластери.

Таблиця 4.19

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_1	119,4	121,0	16,6	114,2	115,8	15,2	17,9	117,5
x_2	16,6	18,1	15,5	19,4	23,2	16,7	15,7	15,2

Завдання 2. У табл. 4.20 наведені вісім підприємств, які характеризуються двома показниками фінансового стану: x_1 – виручка від реалізації, млн грн; x_2 – рентабельність підприємства, %. Необхідно провести кластеризацію підприємств, описаних двома показниками, за методом далекого сусіда. В обчисленні відстаней використовуйте відстань city-block. Результати кластеризації подайте у вигляді дендрограм. Зробіть висновки щодо наявності природного розбиття сукупності об'єктів на кластери.

Таблиця 4.20

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_1	121,4	123,0	18,6	116,2	117,8	17,2	19,9	119,5
x_2	18,6	10,1	17,5	11,4	15,2	18,7	17,7	17,2

Завдання 3. У табл. 4.21 наведені вісім регіонів країни, які характеризуються двома показниками туристичної привабливості: x_1 – площа, яку займають туристичні ресурси, км^2; x_2 – інвестиції в основний капітал готелів і ресторанів, млн дол. Необхідно провести кластеризацію країн, описаних двома показниками, за методом середнього зв'язку. В обчисленні відстаней використовуйте просту зважену евклідову відстань w1 = 0.4, w2 = 0.6. Результати кластеризації подайте у вигляді дендрограм. Зробіть висновки щодо наявності природного розбиття сукупності об'єктів на кластери.

Таблиця 4.21

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_1	114,4	116,0	11,6	19,2	110,8	11,2	12,9	112,5
x_2	12,6	14,1	12,5	15,4	19,2	11,7	12,7	12,2

Завдання 4. У табл. 4.22 наведені шість підприємств, які характеризуються двома показниками фінансового стану: x_1 – виручка від реалізації, млн грн; x_2 – рентабельність підприємства, %. Необхідно провести кластеризацію підприємств, описаних двома показниками, за методом дендритів. Зробіть висновки щодо розбиття сукупності об'єктів на кластери.

Таблиця 4.22

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_1	113,2	110,2	113,7	110,6	19,1	125,8	113,2	110,2
x_2	24,2	29,6	26,6	26,7	20,1	12,1	24,2	29,6

Завдання 5. У табл. 4.23 наведені сім країн, які характеризуються двома показниками інвестиційної привабливості: x_1 – обсяг ринку збуту продукції, млрд. дол., x_2 – рентабельність виробництва, %. Необхідно провести кластеризацію підприємств, описаних двома показниками, за методом шарів.

Таблиця 4.23

Вихідні дані

№ об'єкта / Ознака	1	2	3	4	5	6	7	8
x_1	133,2	120,2	133,7	120,6	115,1	145,8	137,5	133,2
x_2	24,2	20,6	16,6	36,7	35,1	30,1	19,2	24,2

Зробіть висновки щодо розбиття сукупності об'єктів на кластери.

Розділ 5. Класифікація з навчанням. Методи дискримінантного аналізу

5.1. Сутність і завдання дискримінантного аналізу. Обмеження та проблеми використання методів дискримінантного аналізу.

5.2. Методи дискримінантного аналізу. Алгоритм лінійного дискримінантного аналізу Фішера для двох класів. Перевірка якості дискримінації.

5.3. Приклади використання дискримінантного аналізу.

Література: [14; 15; 23; 34; 36].

Ключові слова: дискримінантний аналіз; дискримінантні змінні; правило дискримінації; дискримінантні функції; параметричні методи, непараметричні методи; коваріаційна матриця.

5.1. Сутність і завдання дискримінантного аналізу. Обмеження та проблеми використання методів дискримінантного аналізу

Дискримінантний аналіз (ДА) є найважливішим інструментом під час вирішення задач класифікації. На відміну від інших методів, дискримінантний аналіз дозволяє досліднику спрогнозувати, до якого класу належить новий об'єкт. Він містить статистичні методи класифікації багатовимірних об'єктів у ситуації, коли дослідник має так звані навчальні вибірки (класифікація з навчанням). Незважаючи на багато обмежень під час виконання даного методу, дискримінантний аналіз доцільно застосовувати в комплексі з іншими методиками багатовимірного статистичного аналізу.

Дискримінантний аналіз – це розділ математичної статистики, змістом якого є розроблення методу розв’язання задач відмінності, тобто дискримінації об'єктів за певними ознаками.

Як і кластерний і кластерний аналіз, ДА належить до методів багатовимірної класифікації. Основна відмінність між методами полягає в тому,

що в ході ДА нові кластери не утворюються, а формулюється правило, за яким нові одиниці сукупності відносять до одного із уже існуючих класів. ДА дозволяє велику неоднорідну сукупність розбити на однорідні групи, а також віднести певний об'єкт (явище, процес, спостереження) до конкретного класу. Головні завдання дискримінантного аналізу наведено на рис. 5.1.

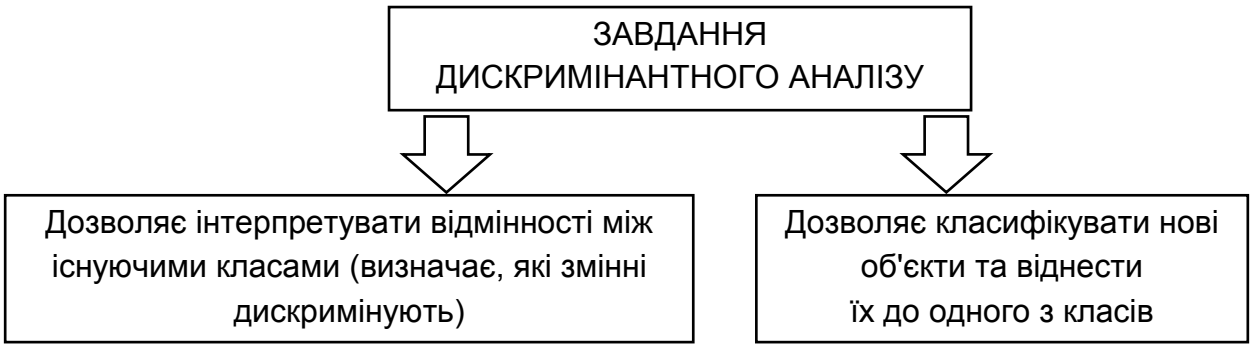


Рис. 5.1. **Завдання дискримінантного аналізу**

Методи дискримінантного аналізу знаходять застосування практично в усіх галузях: економіка, соціологія, медицина, психологія, управління. Обмеження для використання ДА наведено на рис. 5.2.

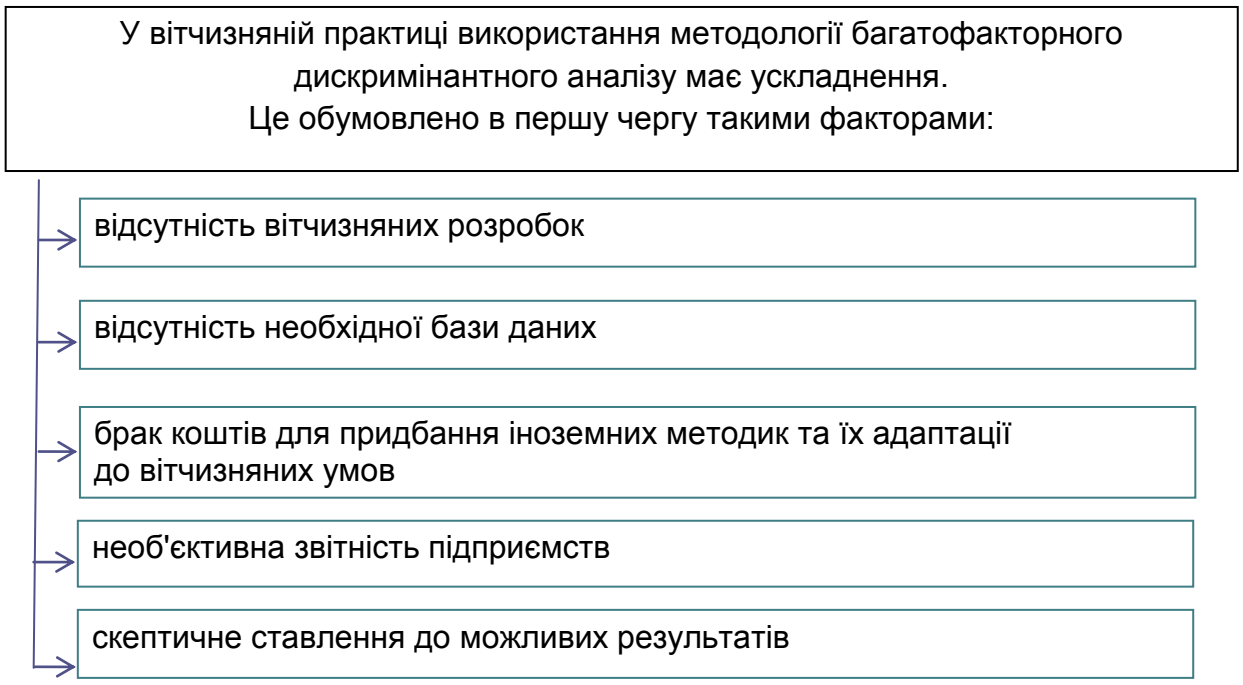


Рис. 5.2. **Обмеження для використання дискримінантного аналізу**

У загальному випадку задача на розрізнення (дискримінації) формулюється так: нехай результатом спостереження над об'єктом є реалізація k-вимірного випадкового вектора $x = (x_1, x_2, \dots, x_k)^T$. Потрібно встановити правило, відповідно до якого за спостереженим значенням вектора x об'єкт відносять до однієї з можливих сукупностей $\varphi_i, i = 1,2, \dots, l$.

Для побудови правила дискримінації весь вибірковий простір R значень вектора x розбивається на області R_i, $i = 1,2, \dots, l$ так, що з влученням x у R_i об'єкт відносять до сукупності φ_i.

Правило дискримінації вибирається відповідно до певного принципу оптимальності на основі апріорної інформації. Остання може бути подана як у вигляді деяких відомостей про функції k-вимірного розподілу ознак у кожній сукупності, так і у вигляді вибірок із цих сукупностей. Апріорні ймовірності можуть бути задані або ні.

Основні проблеми у використанні ДА наведено на рис. 5.3.

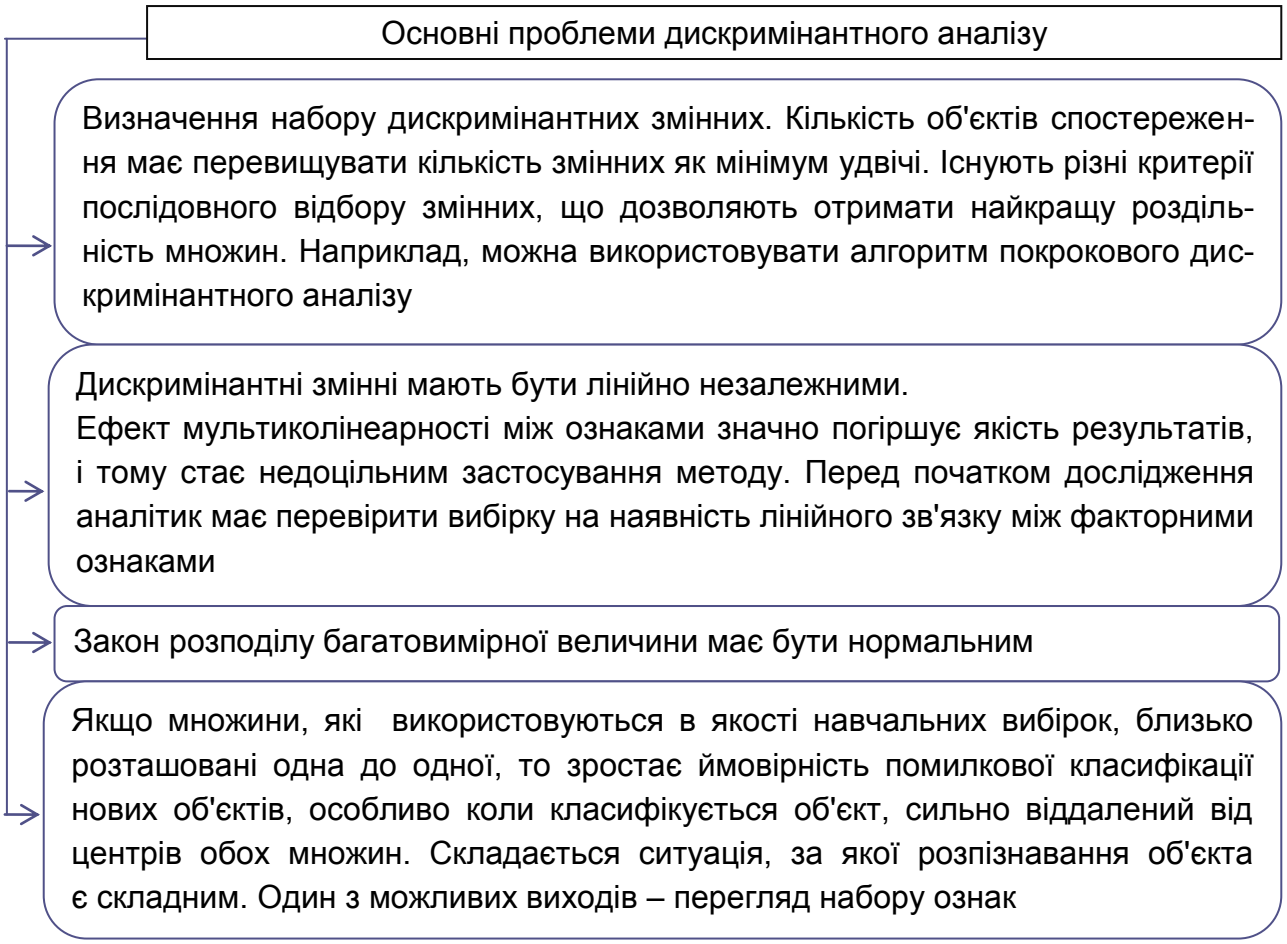


Рис. 5.3. **Проблеми використання дискримінантного аналізу**

Найчастіше вихідна інформація про розподіл представлена вибірками. У цьому випадку задача дискримінації формулюється так: нехай $\chi_1^i, \ldots, \chi_j^i, \ldots, \chi_n^i$ – вибірка із сукупності φ_i, $i = 1,2, \ldots, l$ $i = 1,2, \ldots, l$; причому кожний j-й об'єкт вибірки представлений k-вимірним вектором дискримінантних змінних $\chi_j^i = (x_{jl}^i, \ldots, x_{jq}^i, \ldots, x_{jk}^i)^T$ $x_j^i = x_{jl}^i, \ldots, x_{jq}^i, \ldots, x_{jk}^i$.

Дискримінантні змінні – це ознаки, які використовуються для того, щоб відрізняти один клас (підмножину) від іншого.

Зроблено додаткове спостереження $x = (x_l, \ldots, x_k)$ над об'єктом, що належить сукупності φ_i. Потрібно побудувати правило віднесення спостереження x до однієї із сукупностей. Звичайно в задачі на розпізнавання переходять від вектора ознак, що характеризують об'єкт, до лінійної функції від них. **Дискримінантна функція** – це гіперплощина, яка найкраще розділяє сукупність вибіркових точок. Ці точки використовуються для оцінювання параметрів статистичних функцій розподілу. Як правило, для побудови функції використовують нормальний розподіл.

5.2. Методи дискримінантного аналізу. Алгоритм лінійного дискримінантного аналізу Фішера для двох класів. Перевірка якості дискримінації

Для практичної реалізації дискримінантного аналізу необхідно знати апріорні ймовірності π_j і функції щільності ймовірності $f_j(X)$. Вони можуть бути відомими з теоретичних міркувань або попередніх досліджень. Якщо ж вони невідомі, то їх замінюють статистичними оцінками, отриманими на основі наявних навчальних вибірок [4]. Як оцінки апріорних ймовірностей часто беруть величини:

$$\pi_j = \frac{n_j}{n_{sum}}, \tag{5.1}$$

де n_j – обсяг j-ї вибірки;

$n_{sum} = n_1 + n_2 + \cdots + n_k$ – сумарний обсяг навчальних вибірок.

Для оцінювання функцій щільності ймовірності застосовують два підходи (рис. 5.4). У першому (параметричний дискримінантний аналіз) припускають, що всі класи характеризуються функціями щільності ймовірності, які належать до однієї параметричної сім'ї {fX(Θ)} і розрізняються лише значеннями векторного параметра Θ. У цьому випадку відповідні значення параметра Θ_j оцінюють за спостереженнями, що належать до j-ї вибірки. У другому підході (непараметричний дискримінантний аналіз) загальний вигляд функцій $f_j(X)$ є невідомим. Тому необхідно використовувати спеціальні прийоми їх оцінювання (наприклад, будувати непараметричні оцінки гістограмного або ядерного типу) [4].

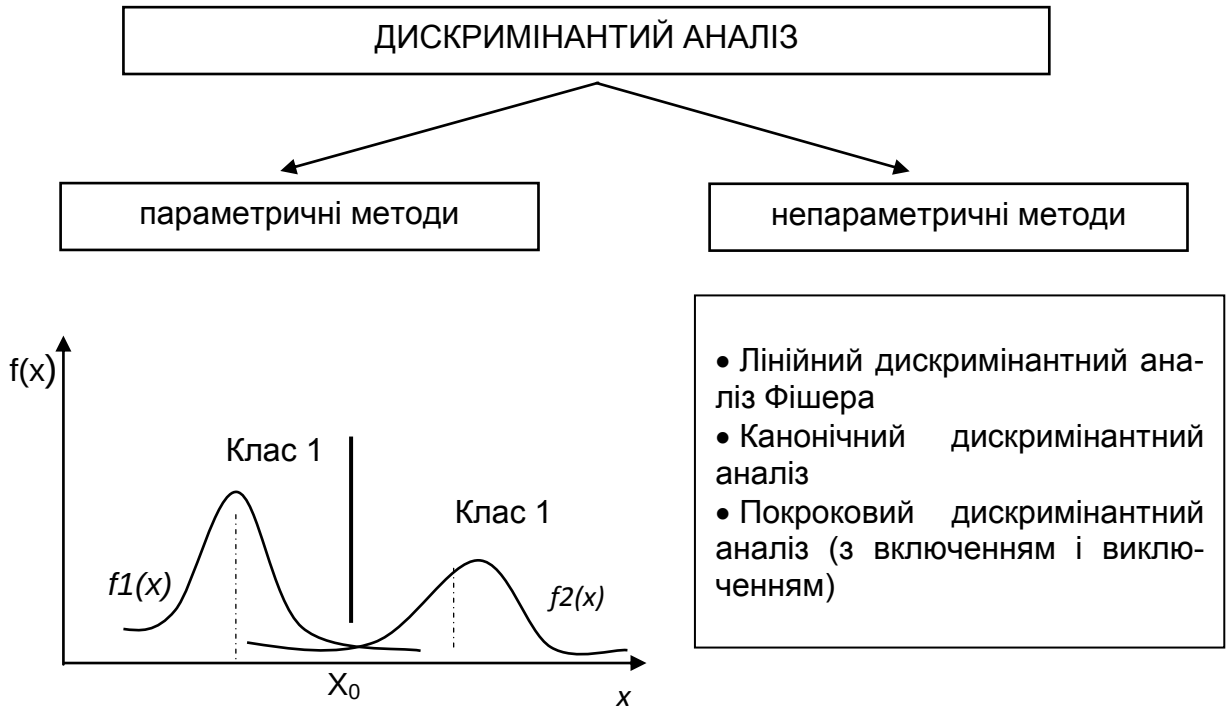


Рис. 5.4. **Методи дискримінантного аналізу**

Розглянемо геометричну інтерпретацію подання дискримінантних змінних. Проаналізуємо об'єкти, що належать двом різним множинам M_1 і M_2 (рис. 5.5).

Як видно з рис. 5.5, кожен об'єкт характеризується в даному випадку двома змінними x_1 і x_2. Якщо розглядати проекції об'єктів (точок) на кожну вісь, то ці множини перетинаються. Тобто деякі об'єкти обох множин мають подібні характеристики за кожною змінною [1].

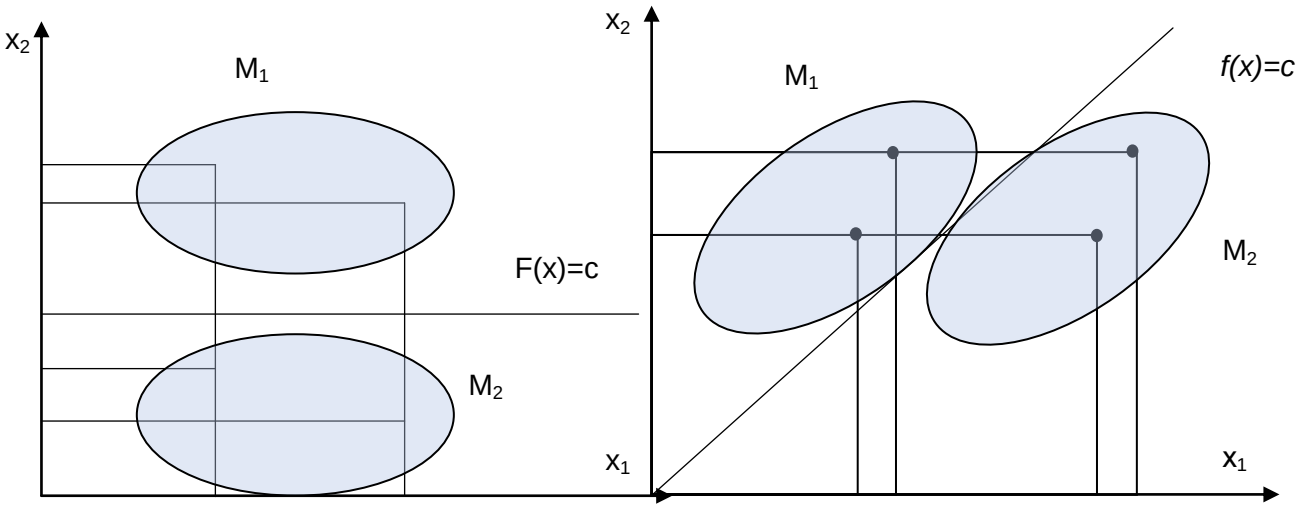


Рис. 5.5. **Геометрична інтерпретація дискримінантних функцій і змінних**

Для того щоб найкращим чином розділити дві подані множини, слід побудувати відповідну лінійну комбінацію змінних x_1 і x_2. Для двовимірного простору це завдання зводиться до визначення нової системи координат. Причому нові осі L і C мають бути розташовані таким чином, щоб на вісь L проекції об'єктів, які належать різним множинам, були максимально розділеними (рис. 5.6).

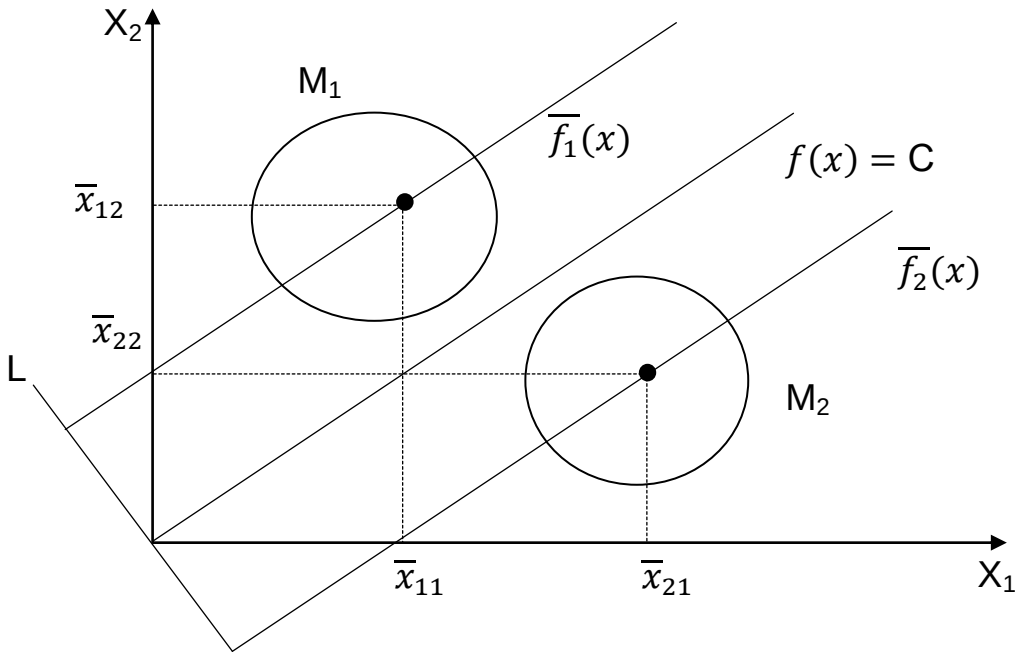


Рис. 5.6. **Центри дискримінантних множин і константа дискримінації**

Вісь С перпендикулярна осі L і розділяє дві "множини" точок таким чином, щоб множини опинилися по різні боки від цієї прямої. Водночас ймовірність помилки класифікації має бути мінімальною [1]. Сформульовані умови слід ураховувати під час обчислення коефіцієнтів a_1 і a_2 дискримінатної функції:

$$D = a_1 x_1 + a_2 x_2 \ . \tag{5.2}$$

Або у загальному випадку:

$$D = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k \ .$$

Дискримінантна функція f(x) – комбінація показників, коефіцієнти якої підбираються за умови найбільшої різниці функції між відомими класами.

Константа дискримінантної функції – межа, що розділяє дві сукупності

Константу обчислюють за формулою:

$$C = \frac{1}{2}\left(\overline{f}_1 + \overline{f}_2\right). \tag{5.3}$$

Позначимо через x_{ij} середнє значення j-ї ознаки у об'єктів і-ї множини (класу). Тоді для множини M_1 середнє значення функції $f_1(x)$ буде таким: $\overline{f_1}(x) = a_1\overline{x_{11}} + a_2\overline{x_{12}}$. Для множини M_2 середнє значення функції $f_2(x)$ дорівнюватиме: $\overline{f_1}(x) = a_1\overline{x_{21}} + a_2\overline{x_{22}}$. Геометрична інтерпретація цих функцій – дві паралельні прямі, що проходять через центри класів (множин) (див. рис. 5.6).

Коефіцієнти дискримінантної функції a_i визначаються таким чином, щоб $f_1(x)$ і $f_2(x)$ якомога більше розрізнялися між собою, тобто щоб для двох множин (класів) був максимальним вираз:

$$\overline{f_1(x)} - \overline{f_2(x)} = \sum_{i=1}^{n} a_1 x_{2i} - \sum_{i=1}^{n} a_2 x_{2i}. \tag{5.4}$$

У цій ситуації має мінімізуватися внутрішньогрупова дисперсія – квадрат суми відхилень від середніх значень:

$$\sum_{i=1}^{2}\sum_{t=1}^{n_k}\left(Y_k - \overline{Y}_k\right) = A'\,(X'_1 X_1 + X'_2 X_2). \qquad (5.5)$$

Разом з тим міжгрупова варіація має бути максимальною, тобто:

$$(\overline{Y}_1 - \overline{Y}_2)^2 = A\left(\overline{X}_1 - \overline{X}_2\right)\overline{(X}_1 - \overline{X})_2{}'\ A. \qquad (5.6)$$

За необхідності можна проводити розбиття множини об'єктів на k класів, потрібно розрахувати k дискримінантних функцій, оскільки класи будуть відокремлюватися один від одного індивідуальними роздільними поверхнями (рис. 5.7). Наприклад, коли є три сукупності, можна оцінити: функцію для дискримінації між сукупністю M_1 і множинами M_2 і M_3, взятими разом, та іншу функцію для дискримінації між сукупністю М2 і сукупністю M_3 [2].

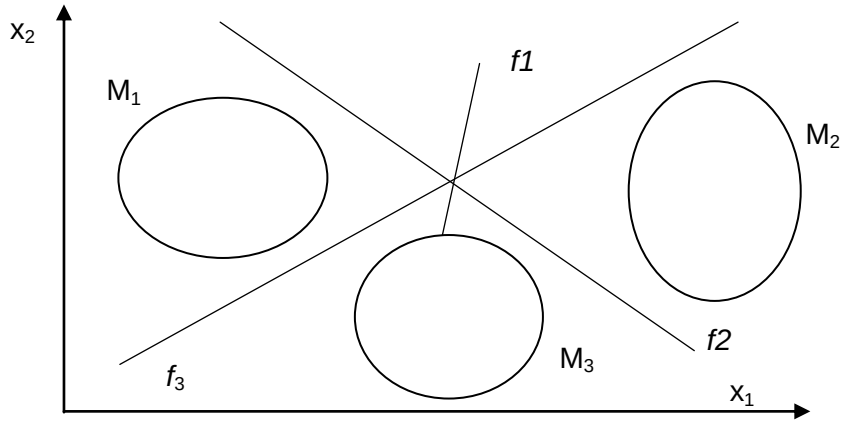


Рис. 5.7. **Дискримінантні функції для трьох вибірок**

Для оцінювання якості дискримінації слід використовувати критерій Фішера (F-критерій). Найкращий розподіл груп відбувається за максимального його значення:

$$F = \frac{A\left(\overline{X}_1 - \overline{X}_2\right)\overline{(X}_1 - \overline{X})_2{}'\ A}{A'[(n_1 + n_2 - 2)S_*]A} \rightarrow max; \qquad (5.7)$$

$$S_* = \frac{1}{n_1 - n_2 - 2}(X'_1 X_1 + X'_2 X_2). \qquad (5.8)$$

Вектор коефіцієнтів обчислюється таким чином:

$$A' = S_*^{-1}\left(\overline{X}_1 + \overline{X}_2\right). \quad (5.9)$$

Розглянемо *алгоритм лінійного дискримінантного аналізу Фішера* для двох класів за нормального закону розподілу показників. Для використання методу необхідно виконання таких умов:

обсяг вибірки має бути більшим, ніж кількість змінних;

кластери, серед яких здійснюють дискримінацію, підпорядковані багатовимірному нормальному розподілу;

класи можуть перетинатися, але їх центри мають бути достатньо віддаленими один від одного;

різниця між коваріаційними матрицями цих кластерів є статистично незначущою; кількість навчальних вибірок у кластері є меншою, ніж кількість дискримінантних функцій [4].

В основу методу лінійного дискримінантного аналізу, запропонованого Р. Фішером у 1936 р., закладене припущення, що класифікацію можна здійснити за допомогою лінійної комбінації дискримінантних (пояснювальних) змінних [4].

Нехай є дві генеральні сукупності *X* і *Y*, що мають тривимірний закон розподілу з невідомими, але рівними коваріаційними матрицями. З них узяті навчальні вибірки з обсягами n_1 в *X* и n_2 в *Y*. Розглянемо алгоритм методу.

1. Задають вхідні матриці X і У з об'єктами n1 і n2, відповідно.

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}; \qquad Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{pmatrix}.$$

2. Записують нове спостереження (z), яке слід класифікувати:

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & Z_{13} \\ Z_{21} & Z_{22} & Z_{23} \\ Z_{31} & Z_{32} & Z_{33} \end{pmatrix}.$$

3. Обчислюють вектори середніх значень кожної з підмножин:

$$\bar{x} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_2} \end{pmatrix} \text{ і } \bar{y} = \begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ \overline{y_3} \end{pmatrix} \overline{x}_j = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{ij}.$$

4. Обчислюють оцінки коваріаційних матриць Sx і Sy:

$$S_x = (S_{ki})_x \;\; i \; S_y = \; (S_{ki})_y.$$

Знаходимо елемент матриці $\overline{S_x}$:

$$S_{ki} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(x_{ij} - \overline{x_j}\right) (x_{ik} - \overline{x_k}) = \; \overline{x_j x_k} - \overline{x_j} \; \overline{x_k}; \quad j,k = 1,2,3,$$

де x_j і x_k – середні значення.

5. Визначають незміщену оцінку сумарної коваріаційної матриці:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2} \left(n_1 S_x + n_2 S_y\right).$$

6. Визначають матрицю, $\hat{S}^{-1}$, зворотну до $\hat{S}$.

7. Обчислюють вектор оцінок дискримінантної функції:

$$a = \bar{S}^{-1}(\bar{x} - \bar{y}).$$

8. Обчислюють оцінки векторів значень дискримінантної функції для матриць вихідних даних $\widehat{U_x} = Xa$, $\widehat{U_y} = Ya$.

9. Обчислюють середні значення оцінок дискримінантної функції:

$$\overline{\widehat{u_x}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{u_{xi}}, \qquad \overline{\widehat{u_y}} = \frac{1}{n_1} \sum_{i=1}^{n_2} \widehat{u_{yi}}.$$

10. Обчислюють константи:

$$\hat{C} = \frac{1}{2} \left(\overline{\widehat{u_x}} + \overline{\widehat{u_y}}\right).$$

11. Записують дискримінантну функцію і її економічну інтерпретацію.

12. Для перевірки умови максимально чіткого поділу груп доцільно використовувати критерій "лямбда Уїлкса":

$$L_W = \frac{1}{1+\lambda}, \tag{5.10}$$

де λ знаходять відповідно до формули:

$$\lambda = F = \frac{a^T S_{xy} a}{a^T S * a} \rightarrow max. \tag{5.11}$$

За $L_W \rightarrow 1$ групи поділені чітко.
За $L_W \rightarrow 0$ результати аналізу не можна використовувати.

Вплив окремих показників на результати дискримінантного аналізу розраховують так:

$$R_{x_i} = \frac{\left|a_j^*\right|}{\sum_{j=1}^{m}\left|a_j^*\right|}, \tag{5.12}$$

де $\left|a_j^*\right|$ – модуль стандартизованої оцінки показника Xj.

13. Обчислюють значення дискримінантної функції для v-го спостереження, що підлягає дискримінації, вирішивши рівняння:

$$\widehat{U_v} = z_{v1}a_1 + z_{v2}a_2 + z_{v3}a_3.$$

Для віднесення нового об'єкта Z до матриці X або У отримане значення дискримінантної функції для v-го спостереження $\widehat{U_v}$ порівнюють з константою С або використовують шкалу інтерпретації нових об'єктів. Правила віднесення об'єктів до класів наведені на рис. 5.8.

Рис. 5.8. **Правила віднесення об'єктів до класів**

Описаний алгоритм можна використовувати у випадку розподілу навчальної вибірки на два класи спостережень.

5.3. Приклади використання дискримінантного аналізу

Приклад 1. Використання дискримінантного аналізу для двох класів. Задано вибірку з підприємств двох класів. Нехай у галузі виділено дві групи підприємств: передова, що складається із чотирьох підприємств, та інша, що містить п'ять підприємств. Оцінювання ефективності діяльності кожного підприємства галузі здійснювалася за трьома показниками: середньорічна вартість основних виробничих фондів (ОПФ), середня чисельність промислово-виробничого персоналу (ПВП), балансовий прибуток. Вихідні дані подано в табл. 5.1:

x_1 – продуктивність праці;

x_2 – коефіцієнт змінності обладнання;

x_3 – фондовіддача активної частини ОПФ.

Таблиця 5.1

Вхідні дані

Класи	Підприємства	x_1	x_2	x_3
А	1	9,26	1,37	1,45
	2	9,38	1,49	1,3
	3	12,11	1,44	1,37
	4	10,81	1,42	1,65
Б	5	5,49	1,1	1,02
	6	6,61	1,23	0,88
	7	4,32	1,39	0,62
	8	7,37	1,38	1,09

Необхідно визначити можливість віднесення підприємства до передової групи підприємств галузі. z_1 = (9; 6,7; 0,79; 1,24); z_2= (10; 9,42; 0,7; 2,03).

Розв'язання

1. Знайдемо середні значення за кожною з ознак групи, де $n_1 = n_x = 4$; $n_2 = n_y = 5$.

Xcp	Ycp
10,39	5,95
1,43	1,28
1,44	0,9

2. Визначимо оцінки коваріаційних матриць за формулою:

$$S_{kj}(x) = \frac{1}{n_i}\sum_{i=1}^{n_1}\left(x_{ij} - \overline{x_j}\right)\left(x_{ik} - \overline{x_k}\right) \quad (14)$$

Отримаємо:

S_x			S_y		
1,35795	0,00505	0,024475	1,330119	0,003112	0,164456
0,00505	0,00185	-0,00295	0,003112	0,014225	-0,00809
0,024475	-0,00295	0,017169	0,164456	-0,00809	0,032319

$$S_{12} = \frac{1}{4} \times \sum_{i=1}^{4} (x_{i1} - \overline{x_1}) \times \left(x_{i2} - \overline{x_{iJ}}\right) =$$

$$= \frac{1}{4} \times ((9{,}26 - 10{,}39) \times (1{,}37 - 4{,}42) + \cdots +$$

$$+(10{,}81 - (10{,}39) \times (1{,}37 - 4{,}42).$$

3. Знайдемо незміщену оцінку сумарної коваріаційної матриці за формулою:

$$S = \frac{1}{n_1+n_2-2}(n_1 S_x + n_2 S_x)\hat{S} = \frac{1}{4+5-2}\left(4S_x + 5S_y\right).$$

Отримуємо:

S =

1,792	0,0054	0,126
0,0054	0,0107	-0,0074
0,126	-0,0074	0,033

4. Знаходимо зворотну коваріаційну матрицю до $\hat{S}$.

S^{-1}=

0,8434	-3,1398	-3,9245
-3,1398	122,3002	39,4131
-3,9245	39,4131	54,1254

5. Знаходимо вектор коефіцієнтів дискримінації за формулою:

$$A' = \hat{S}^{-1}\left(\overline{X} - \overline{Y}\right).$$

Отримуємо: A = (1,1481; 26,1074; 17,8073).

$$S_{12} = \frac{1}{4} \cdot \sum_{i=1}^{4} (x_{i1} - \overline{x_1}) \cdot \left(x_{i2} - \overline{x_{ij}}\right) =$$

$$= \frac{1}{4} \cdot ((9{,}26 - 10{,}39) \cdot (1{,}37 - 4{,}42) + \cdots +$$

$$+(10{,}81 - 10{,}39) \cdot (1{,}42 - 4{,}42).$$

6. Обчислимо оцінки дискримінантних функцій $\widehat{U_x} = Xa$.

Ux	Uy
72,2196	53,1850
72,8192	55,3718
75,8947	52,2898
78,8661	63,9000

7. Визначаємо середні значення отриманих оцінок: $\widehat{u_x} = 74{,}9499$; $\widehat{u_y} = 56{,}18666$.

8. Обчислимо константу дискримінації за формулою:

$$\hat{C} = \frac{1}{2}(\overline{f_1} + \overline{f_2}).$$

Отримуємо: С = 65,568.

9. Визначимо можливість включення об'єднання z у групу передових. Оскільки матриця Z подана одним рядком, то $\widehat{U_y}$ позначимо $\widehat{U_z}$:

$$\widehat{u_z} = z_1 a_1 + z_2 a_2 + z_3 a_3.$$

$\widehat{u_z}$ =50,399 z2=65,240.

Середнє значення дискримінантної функції $\hat{u}_z$ менше ніж константа $\hat{C}$ ($\widehat{u_z} < \hat{C}$). Отже, об'єднання z з характеристиками Z^T слід віднести до 2-ї групи підприємств галузі.

10. Розрахуємо критерій "лямбда Уїлкса":

10.1) внутрішньогрупова варіація дорівнює:

$$Q_{SW} = \sum_{i=1}^{n_1} (Y_{1i} - Y_{1G})^2 + (Y_{2i} - Y_{2G})^2 = 19{,}94 + 6{,}25 = 26{,}19;$$

10.2) міжгрупова варіація дорівнює:

$$Q_{sb} = n_1 \times (Y_{1G} - Y_g)^2 + n_2 (Y_{2G} - Y_g)^2 = 4 \times 5{,}01 + 4 \times 5{,}01 = 40{,}1;$$

10.3) критерій "лямбда Уїлкса":

$$\lambda = \frac{40{,}1}{26{,}19} = 1{,}53;\ L_w = \frac{1}{1+1{,}53} = 0{,}395 => \text{дискримінація груп.}$$

11. Розрахуємо вплив окремих параметрів:

$$R_{x_i} = \frac{\left|a_j{}^*\right|}{\sum_{j=1}^{m}\left|a_j{}^*\right|}.$$

R_{x1} = 2,55; R_{x2} = 57,94; R_{x3} = 39,52. Отже, на 57,94 % дискримінація на класи пояснюється другою ознакою, на 39,52 % – третьою, на 2,55 % – першою.

12. За шкалою інтерпретацій: $Z < 56{,}19/2$, $Z < 28{,}1$ – 1-й клас, $28{,}1 \le Z \le 74{,}95$ – невизначеність, $Z > 37{,}48$ – 2-й клас.

Приклад 2. Розв'язання прикладу для випадку з чотирма класами. Дано 32 підприємства, розділені на чотири групи за чотирма ознаками (табл. 5.2).

Таблиця 5.2

Вхідні дані

№ п/п	X_1	X_2	X_3	X_4	S
1	1,75	5,25	8	16	1
2	2,65	5,5	11	15	1
3	1,8	4,47	10	6	1
4	2,5	4,75	7	9	1
5	3	5	12	7	1
6	3,54	4,71	8	6	1
…	…	…	…	…	…
28	4	1,75	7	2	4
29	-3,03	1	8	2	4
30	3,25	0,88	6	3	4
31	3,69	0,49	5	2	4
32	4,2	1	7	2	4

Визначте, до якого класу слід віднести два нових підприємства.

33	3,85	4,47	10,5	8
34	4,6	6,2	8,7	9

Розв'язання

1. Знайдемо середні значення для класів, квадрантів (x_1,x_2 (1-18), x_3,x_4 (1-18), x_1,x_2 (19-32), x_3,x_4 (19-32)) і для всіх 32-х об'єктів узятих сукупно для класів:

для першого квадранта: $\overline{x_{1(1-18)}} = \frac{56,48}{18} = 3,14$; $\overline{x_{2(1-18)}} = \frac{71,89}{18} = 3,99$;

для другого: $\overline{x_{3(1-18)}} = \frac{193}{18} = 10,72$; $\overline{x_{4(1-18)}} = \frac{150}{18} = 8,83$;

для третього $\overline{x_{1(19-32)}} = \frac{61,46}{14} = 4,39$; $\overline{x_{1(19-32)}} = \frac{37,12}{14} = 2,65$;

для четвертого: $\overline{x_{1(19-32)}} = \frac{127}{14} = 9,07$; $\overline{x_{1(19-32)}} = \frac{69}{14} = 4,93$;

для всіх об'єктів: (3,69; 3,41; 10,00; 6,84).

A	B	C	D
2,57	4,27	5,58	3,49
4,47	3,04	4,51	1,26
9,50	13,17	13,17	6,00
9,00	7,00	7,33	3,13

2. Обчислимо у квадрантах значення лінійних відхилень L_1, L_2, L_3, L_4 ($L_i = x_I - x_{icp}$) для кожного об'єкта, що входить у відповідний квадрант.

3. За отриманими відхиленнями знайдемо дисперсії для кожного квадранта:

$$G_{ij} = \sum_i (x_i - \overline{x_j})^2.$$

Отримуємо:

σ_1	σ_2	σ_3	σ_4
32,48	51,3	312,54	320,93

4. Обчислимо значення коваріацій V_{12}, V_{13}, V_{14}, V_{23}, V_{24}, V_{34}. Для цього необхідно перемножити відповідні відхилення. Наприклад, $V_{12} = L_1 \cdot L_2$.

	V_{12}	V_{13}	V_{14}	V_{23}	V_{24}	V_{34}
1	-1,74	3,778	-10,6	-3,42	9,63	-20,9
2	-0,74	-0,14	-3,25	0,418	10,04	1,852
3	-0,64	0,966	3,121	-0,34	-1,11	1,685
4	-0,48	2,374	-0,43	-2,81	0,504	-2,48
5	-0,14	-0,18	0,184	1,286	-1,34	-1,7
6	0,288	-1,1	-0,94	-1,95	-1,67	6,352
7	0,072	0,23	1,059	-0,16	-0,75	2,407
…	…	…	…	…	…	…
29	2,245	1,456	3,981	1,769	4,836	3,138
30	2,018	3,499	2,197	5,441	3,416	5,923
31	1,511	2,847	2,048	8,8	6,33	11,92
32	0,313	0,392	0,554	3,421	4,836	6,066

5. Знайдемо параметри a_1 і a_2 для першого квадранта за такою формулою матричних визначників:

$$\begin{cases} a_1 \cdot \sigma_{11} + a_{12} \cdot \sigma_{12} = \overline{x_{11}} - \overline{x_{12}} \\ a_1 \cdot \sigma_{21} + a_{12} \cdot \sigma_{22} = \overline{x_{21}} - \overline{x_{22}}. \end{cases}$$

Звідси:

$$\sigma_{11} = \frac{\sum(x_1 - \overline{x_1})^2(1-18)}{18} = 0{,}8435;$$

$$\sigma_{12} = \sigma_{21} = \frac{\sum(x_1 - \overline{x_1})^* \times (x_2 - \overline{x_2})(1-18)}{18} = -0{,}5687;$$

$$\sigma_{22} = \frac{\sum(x_2 - \overline{x_2})^2(1-18)}{18} = 0{,}6859.$$

Підставляючи вже знайдені середні значення x і σ, запишемо систему рівнянь:

$$\begin{cases} a_1 \cdot 0{,}8435 - a_2 \cdot 0{,}5687 = 2{,}5741 - 4{,}2650 \\ -a_1 \cdot 0{,}5687 - a_2 \cdot 0{,}6859 = 3{,}0416. \end{cases}$$

Розв'язавши систему, a_1 отримуємо: $a_1 = -1{,}3617;\ a_2 = 0{,}9530.$

Аналогічно розв'язуємо систему для всіх параметрів усіх квадрантів. Отримуємо такі оцінки параметрів:

Квадранти	I	II	III	IV
Параметри	a_1 =-1,3617	a_3 =-0,6840	a_1 = 0,4164	a_3 = 0,3056
	a_2 = 0,9530	a_4 = 0,0864	a_2 = 0,9146	a_4 = 0,341

6. На основі отриманих параметрів обчислюємо значення дискримінантних функцій: $f(x)_{12}$ – за параметрами x_1 і x_2; $f(x)_{34}$ – за параметрами x_3 і x_4. Причому значення функцій для конкретного об'єкту розраховуються відповідно до конкретного значення оцінок параметрів для певного квадранта.

№ п\п	$f(x)_{12}$	$f(x)_{34}$
1	2,62	-4,09
2	1,633	-6,228
3	1,809	-6,322
4	1,123	-4,01
5	0,68	-7,603
6	-0,332	-4,954
7	0,182	-6,408
…	…	…
29	2,176	3,127
30	2,158	2,857
31	1,985	2,21
32	2,663	2,821

Обчислюємо суму частинних функцій в межах кожного класу.

Класи	1	2	3	4
$f(x)_{12}$	9,056	-17,454	38,698	20,893
$f(x)_{34}$	-68,645	-50,407	39,146	23,194

7. Знайдемо значення відповідних субфункцій і субконстант:

7.1) для $f(x)_{12}$:

$$C^I = a_1\overline{x_{11(1-12)}} + a_2\overline{x_{21(1-12)}} = 1{,}36 \times 2{,}57 + 0.953 \times 4{,}47 = 0.755;$$

$$C^{II} = a_1\overline{x_{12(13-18)}} + a_2\overline{x_{22(13-18)}} = 1{,}36 \times 4{,}265 + 0{,}953 \times 3{,}042 = -2{,}909;$$

$$C^{III} = a_1\overline{x_{13(19-24)}} + a_2\overline{x_{23(19-24)}} = 0{,}416 \times 5{,}585 + 0{,}915 \times 4{,}51 = 6{,}45;$$

$$C^{IV} = a_1\overline{x_{14(25-32)}} + a_2\overline{x_{24(25-32)}} = 0{,}416 \times 3{,}494 + 0{,}915 \times 1{,}258 = 2{,}605;$$

$$\frac{C^I + C^{II}}{2} = \frac{0{,}755 - 2{,}909}{2} = 1{,}077; \qquad \frac{C^{III} + C^{IV}}{2} = \frac{6{,}45 - 2{,}605}{2} = 4{,}53;$$

$$C_{(x1x2)} = \frac{C^I + C^{II} + C^{III} + C^{IV}}{4} = 1{,}725;$$

7.2) для f(x)34:

$$C^I = a_3\overline{x_{31(1-12)}} + a_3\overline{x_{41(1-12)}} = -0{,}684 \times 9.5 + 0{,}086 \times 9 = -5{,}72;$$
$$C^{II} = a_3\overline{x_{32(13-18)}} + a_3\overline{x_{42(13-18)}} = -0{,}684 \times 13{,}167 + 0{,}086 \times 7 = -8{,}41;$$
$$C^{III} = a_3\overline{x_{33(19-24)}} + a_3\overline{x_{43(19-24)}} = 0{,}306 \times 13{,}167 + 0{,}341 \times 7{,}33 = 6{,}52;$$
$$C^{IV} = a_3\overline{x_{34(25-32)}} + a_3\overline{x_{4(1-12)}} = -0{,}606 \times 6 + 0{,}341 \times 3{,}125 = 2{,}899;$$
$$\frac{C^I + C^{II}}{2} = \frac{-5{,}72 - 8{,}41}{2} = -7{,}061; \frac{C^{III} + C^{IV}}{2} = \frac{6{,}52 - 2{,}89}{2} = 4{,}712;$$
$$C_{x3x4} = \frac{C^I + C^{II} + C^{III} + C^{IV}}{4} = -1{,}175.$$

8. Записуємо загальні функції F (32) і загальні константи за 32-ма об'єктами. Для цього будуємо систему рівнянь за чотирма класами об'єктів з чотирма невідомими:

$$\begin{cases} a_1 \cdot f_{11} + a_2 \cdot f_{12} + a_3 \cdot f_{13} + a_4 \cdot f_{14} = \overline{x_{11}} + \overline{x_{12}} + \overline{x_{13}} + \overline{x_{14}} \\ a_1 \cdot f_{21} + a_2 \cdot f_{22} + a_3 \cdot f_{23} + a_4 \cdot f_{24} = \overline{x_{21}} + \overline{x_{22}} + \overline{x_{23}} + \overline{x_{24}} \\ a_1 \cdot f_{31} + a_2 \cdot f_{32} + a_3 \cdot f_{33} + a_4 \cdot f_{34} = \overline{x_{31}} + \overline{x_{32}} + \overline{x_{33}} + \overline{x_{34}} \\ a_1 \cdot f_{41} + a_2 \cdot f_{42} + a_3 \cdot f_{43} + a_4 \cdot f_{44} = \overline{x_{41}} + \overline{x_{42}} + \overline{x_{43}} + \overline{x_{44}} \end{cases}.$$

Розв'язавши систему рівнянь, отримуємо:

$$a_1 = -13.0507, a_2 = 1{,}1251, a_3 = 1{,}0337, a_4 = -1{,}0445.$$

9. Знайдемо шукане значення загальної функції для 32-х об'єктів:

$$C_{\text{общ}} = a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + a_4 \cdot x_4$$
$$= -13{,}05 \cdot 3{,}69 + 1{,}14 \cdot 3{,}401 + 10 \cdot 1{,}034 - 6{,}84 \cdot 1{,}045 = -41{,}0442.$$

10. Обчислимо значення загальних функцій для кожного класу:

$$f_1 = a_1 \cdot \overline{x_{11}} + a_2 \cdot \overline{x_{21}} + a_3 \cdot \overline{x_{31}} + a_4 \cdot \overline{x_{41}} = -28{,}1;$$
$$f_2 = -45{,}91; f_3 = -61{,}82; f_4 = 41{,}23.$$

11. Обчислюємо значення за новими об'єктами (33 – 34):

$f_{33} = -13{,}051 \cdot 2.3 + 6{,}11 \cdot 1{,}135 + 1{,}034 \cdot 10 - 1{,}045 \cdot 14 = -27{,}35;$
$f_{34} = -13{,}051 \cdot 5{,}21 + 1{,}135 \cdot 1{,}18 + 1{,}034 \cdot 13 - 1{,}045 \cdot 4 = -57{,}395.$

12. Визначимо приналежність об'єктів 33 – 34 до одного з класів. Об'єкт належить саме до того класу, дискримінант якого більше його функції. Якщо функція об'єкта більше дискримінанти, то об'єкт відносять до наступного класу: $33: -27{,}3671 > -28{,}1002$ обєкт належить до 1 класу.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. У чому сутність дискримінантного аналізу та його відмінність від інших методів БСА?
2. Назвіть переваги дерев класифікацій як методу класифікації.
3. Назвіть основні етапи процесу побудови дерева класифікацій.
4. Яким чином здійснюють розпізнавання та прогнозування за правилами та закономірностями?
5. Опишіть процес побудови дерева, назвіть особливості розпізнавання ситуацій.
6. Яким чином виконують оцінювання якості моделей?
7. Як визначається кількість дискримінантних функцій?
8. Сформулюйте правило дискримінації.
9. Як визначити константу дискримінації?

Тестові завдання

1. Дискримінантний аналіз – це статистичний метод, який дозволяє:

а) вивчати відмінності між двома та більше групами об'єктів за декількома змінними одночасно;

б) знаходити групи схожих об'єктів у вибірці даних.

2. Який закон розподілу повинні мати вихідні дані для дискримінантного аналізу:

а) нормальний;

б) показниковий;

в) експоненційний;

г) закон розподілу не має значення?

3. Кількість об'єктів у вихідній вибірці має відрізнятися від кількості класів більше ніж:

а) удвічі;

б) утричі;

в) має бути менше кількості класів на ½.

4. Дисперсію показника обчислюють за формулою:

а) $\mathrm{D} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$;

б) $\mathrm{D} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$;

в) $\mathrm{D} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)$.

5. Коефіцієнт коваріації обчислюють за формулою:

а) $S_{kj}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)(x_{ik} - \bar{x}_k)$;

б) $S_{kj}(x) = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)(x_{ik} - \bar{x}_k)$;

в) $S_{kj}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)$.

6. Елементи коваріаційної матриці симетричні щодо:

а) головної діагоналі;

б) середньої рядка;

в) середнього стовпця.

7. Методи дискримінантного аналізу відносять до групи методів класифікації:

а) з навчанням;

б) без навчання.

8. Лінійна дискримінантна функція має такий вигляд:

а) $Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$;

б) $Y = a_1z_1 + z_2x_2 + \cdots + z_nx_n$;

в) $Y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$.

9. Нове спостереження Z відносять до сукупності, якщо:

а) $\hat{u}_v \geq \hat{C}$;

б) $\hat{u}_v < \hat{C}$;

в) $\hat{u}_v \geq \frac{1}{2}\hat{C}$.

10. Для оцінювання якості дискримінації використовують критерій:

а) Пірсона;

б) лямбда Уїлкса;

в) Стьюдента.

11. Дисперсія помилки – це:

а) частка дисперсії, обумовлена варіабельною специфікою ознаки;

б) частка дисперсії, обумовлена недосконалістю вимірювань;

в) частка дисперсії характерного фактора без урахування помилки.

12. Чи може об'єкт, який класифікується, бути віднесений відразу до декількох груп (в дискримінантному аналізі):

а) так;

б) не може;

в) не завжди?

13. Від початкових даних у дискримінантному аналізі переходять до:

а) кореляційної матриці;

б) коваріаційної матриці;

в) матриці евклідових відстаней;

г) матриці суміжності.

14. Ступінь дискримінації моделі визначають за допомогою критерію:

а) Фішера;

б) лямбди Уїлкса;

в) максимальної правдоподібності.

15. Елементи головної діагоналі коваріаційної матриці дорівнюють:

а) 1;

б) 0;

в) дисперсії.

16. Елементи головної діагоналі матриці відстаней між об'єктами дорівнюють:

а) 1;

б) 0;

в) дисперсії;

г) коефіцієнтам кореляції.

17. Дискримінантна модель є адекватною, якщо лямбда Уїлкса є наближеною до:

а) 0;

б) 1;

в) середнього значення.

18. Вектор оцінок коефіцієнтів дискримінантної функції розраховують за формулою:

а) $a = \hat{S}^{-1}(\bar{X} - \bar{Y})$;

б) $a = \hat{S}(\bar{X} - \bar{Y})$;

в) $a = \frac{1}{2}(\bar{\widehat{U}}_x - \bar{\widehat{U}}_y)$.

19. Незміщена оцінка сумарної коваріаційної матриці обчислюється за формулою:

а) $S = \frac{1}{2}(\bar{\widehat{U}}_x + \bar{\widehat{U}}_y)$;

б) $\hat{S} = \frac{1}{n_1+n_2-2}\left(n_1 S_x + n_2 S_y\right)$;

в) $S_{kj}(x) = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)(x_{ik} - \bar{x}_k)$;

г) $\hat{S} = \frac{1}{n_1+n_2}\left(nS_x + nS_y\right)$.

20. Константа дискримінантної функції розраховується за формулою:

а) $C = \hat{S}(\bar{X} - \bar{Y})$;

б) $C = \frac{1}{2}(\bar{\widehat{U}}_x + \bar{\widehat{U}}_y)$;

в) $C = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)$.

Практичні завдання

Завдання 1. Діяльність підприємства характеризується такими показниками: x_1 – продуктивність праці; x_2 – коефіцієнт змінності обладнання; x_3 – питома вага втрат від браку. За допомогою методів дискримінантного аналізу за даними табл. 5.3, використовуючи класи (у кожній таблиці подано дві групи машинобудівних підприємств) як навчальні вибірки, слід провести дискримінацію зазначених підприємств, кожне з яких характеризується трьома показниками, та віднести їх до відповідного класу. Надайте економічну інтерпретацію та сформулюйте висновки.

Таблиця 5.3

Початкові дані

Класи	№ підприємства	x_1	x_2	x_3
A	1	9,26	1,37	0,23
	2	9,38	1,49	0,39
	3	12,11	1,44	0,43
	4	10,81	1,42	0,18
B	5	5,49	1,10	0,05
	6	6,61	1,23	0,48
	7	4,32	1,39	0,41
	8	7,37	1,38	0,62
Необхідно розпізнати	9	6,7	0,79	0,39
	10	9,42	0,70	0,72

Завдання 2. Діяльність підприємства характеризується такими показниками: x_1 – продуктивність праці; x_2 – коефіцієнт змінності обладнання; x_3 – фондовіддача активної частини ОПФ.

За допомогою методів дискримінантного аналізу за даними табл. 5.4, використовуючи класи (у кожній таблиці подано дві групи машинобудівних підприємств) як навчальні вибірки, слід провести дискримінацію зазначених підприємств, кожне з яких характеризується трьома показниками, та віднести їх до відповідного класу. Дайте економічну інтерпретацію і зробіть висновки.

Таблиця 5.4

Початкові дані

Класи	№ підприємства	x_1	x_2	x_3
A	1	9,26	1,37	1,45
	2	9,38	1,49	1,3
	3	12,11	1,44	1,37
	4	10,81	1,42	1,65
B	5	5,49	1,10	1,02
	6	6,61	1,23	0,88
	7	4,32	1,39	0,62
	8	7,37	1,38	1,09
Необхідно розпізнати	9	6,7	0,79	1,24
	10	9,42	0,70	2,03

Завдання 3. Діяльність підприємства характеризується такими показниками (табл. 5.5).

Таблиця 5.5

Початкові дані

Класи	№ підприємства	x_1	x_2	x_3
А	1	9,26	0,78	1,37
	2	9,37	0,79	1,24
	3	10,02	0,76	1,22
	4	10,81	0,7	1,42
В	5	5,49	0,74	1,10
	6	6,64	0,77	1,35
	7	5,22	0,79	1,33
	8	7,37	0,77	1,38
Необхідно розпізнати	9	12,11	0,68	1,44
	10	4,32	0,68	1,39

x_1 – продуктивність праці; x_2 – питома вага робітників у складі ППП; x_3 – коефіцієнт змінності обладнання За допомогою методів дискримінантного аналізу, використовуючи класи (у кожній таблиці подано дві групи машинобудівних підприємств) як навчальні вибірки, слід провести дискримінацію зазначених підприємств, кожне з яких характеризується трьома показниками, та віднести їх до відповідного класу. Дайте економічну інтерпретацію і зробіть висновки.

Завдання 4. Діяльність підприємства характеризується такими показниками (табл. 5.6): x_1 – продуктивність праці; x_2 – коефіцієнт використання сировини та матеріалів; x_3 – рентабельність продукції.

Таблиця 5.6

Початкові дані

Класи	№ підприємства	x_1	x_2	x_3
А	1	4,0	80	6,0
	2	4,9	78,6	6,3
	3	6,1	75,9	7,0
	4	5,3	74,0	7,1
В	5	8,7	90,7	9,0
	6	10,3	94,6	10,5
	7	11,6	94,0	10,9
	8	10,8	92,5	11,0
Необхідно розпізнати	9	5,5	74,0	6,1
	10	9,7	92,5	11,1

За допомогою методів дискримінантного аналізу за даними табл. 5.6, використовуючи класи (у кожній таблиці подано дві групи машинобудівних підприємств) як навчальні вибірки, слід провести дискримінацію зазначених підприємств, кожне з яких характеризується трьома показниками, та віднести їх до відповідного класу.

Дайте економічну інтерпретацію і зробіть висновки.

Завдання 5. Діяльність підприємства характеризується такими показниками: x_1 – середньорічна вартість ОПФ; x_2 – середньооблікова чисельність працюючих; x_3 – обсяг виробленої продукції. За допомогою методів дискримінантного аналізу за даними табл. 5.7, використовуючи класи (у кожній таблиці подано дві групи машинобудівних підприємств) як навчальні вибірки, слід провести дискримінацію зазначених підприємств, кожне з яких характеризується трьома показниками, та віднести їх до відповідного класу.

Таблиця 5.7

Початкові дані

Класи	№ підприємства	x_1	x_2	x_3
A	1	170,5	10,0	250,95
	2	200,0	18,2	380,6
	3	186,4	15,8	300,2
	4	154,2	10,3	280,36
B	5	60,6	9,0	100,5
	6	90,8	9,7	147,6
	7	100,4	8,3	194,3
	8	85,2	7,9	170,2
Необхідно розпізнати	9	165,3	17,2	290,5
	10	170,5	10,0	250,95

Дайте економічну інтерпретацію і зробіть висновки.

Розділ 6. Методи повної редукції. Таксономічний показник рівня розвитку

6.1. Поняття редукції ознак. Класифікація методів редукції ознак.

6.2. Таксономічний показник рівня розвитку.

6.3. Приклад застосування таксономічного показника рівня розвитку в економічних дослідженнях.

Ключові слова: редукція ознак; методи неповної редукції; методи повної редукції; таксономічний показник рівня розвитку; статична характеристика безлічі елементів; динамічна характеристика одного елемента; стандартизація; еталон розвитку; відстань від об'єкта до точки еталона; лінійне впорядкування багатовимірних об'єктів.

Література: [7; 22; 33; 43; 44; 46].

6.1. Поняття редукції ознак. Класифікація методів редукції ознак

В економічних дослідженнях досить часто виникає завдання скорочення розмірності вихідного простору ознак. Це пов'язане з тим, що економічні системи (регіон, банк, корпорація, холдинг, страхова компанія, інвестиційний фонд і т.д.) мають складну багаторівневу структуру. Для таких систем характерні наявність безлічі елементів, великої кількості різноманітних зв'язків, циркуляція значних потоків інформації, що визначають їх внутрішню динаміку. Таким чином, економічні системи характеризуються високим ступенем складності. Як наслідок, їх інформаційна модель повинна включати велике число кількісних і якісних показників. Разом із цим урахування великої кількості показників призводить до інформаційної перевантаженості процесів ухвалення рішень. Саме тому постає необхідність формування системи найбільш інформативних, діагностичних показників, що дозволяють знизити розмірність вихідного інформаційного простору ознак без втрати значущої інформації.

Поняття "редукція признакового простору" розробив і запровадив доктор наук, професор З. Хельвіг. Надалі цей напрям багатомірного статистичного аналізу розвивали такі вчені, як В. Плюта, С. Бартосевич

та ін. Сутність завдання скорочення (редукції) розмірності ознакового простору полягає в тому, щоб подати вихідну інформацію, яка задана у вигляді досить великої кількості ознакових описів, у просторі меншої розмірності, за можливості, з мінімізацією втрати інформації.

Класифікація методів редукції ознак наведена на рис. 6.1.

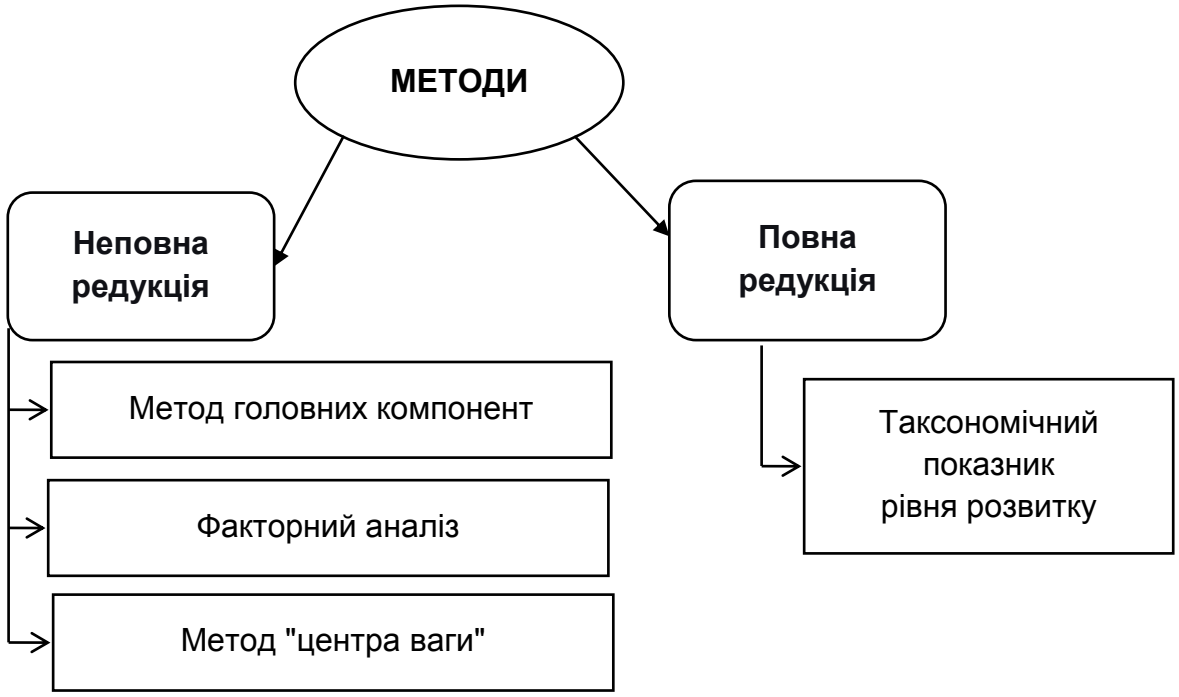


Рис. 6.1. **Класифікація методів редукції ознак**

Методи редукції ознак підрозділяють на методи повної та неповної редукції.

Перша група методів спрямована на побудову синтетичних величин, що мають багатоознакову природу, у вигляді деякої функції від вихідної системи ознак $f(y_1, y_2, \dots, y_n)$, яка відображає вплив усіх ознак, і в такий спосіб дозволяє впорядкувати досліджувані об'єкти. До цієї групи методів належить, зокрема, таксономічний показник рівня розвитку.

Методи неповної редукції ознак припускають формування так званих діагностичних ознак, якими є деякі з вихідних ознак. Для цього первинний набір q ознак $y = (y_1, y_2, \dots, y_q)$ заміняється набором s діагностичних ознак $x = (x_1, x_2, \dots, x_s),\ (s < q)$.. Представниками другої групи є метод головних компонент, факторний аналіз, метод "центру ваги".

6.2. Таксономічний показник рівня розвитку

Таксономічний показник рівня розвитку запропонований З. Хельвігом. Таксономічний показник рівня розвитку є синтетичною величиною, "рівнодіючою" всіх ознак, які характеризують об'єкти. Це дозволяє за його допомогою лінійно впорядкувати елементи досліджуваної сукупності. Завдання, які вирішуються за допомогою цього методу, наведені на рис. 6.2.

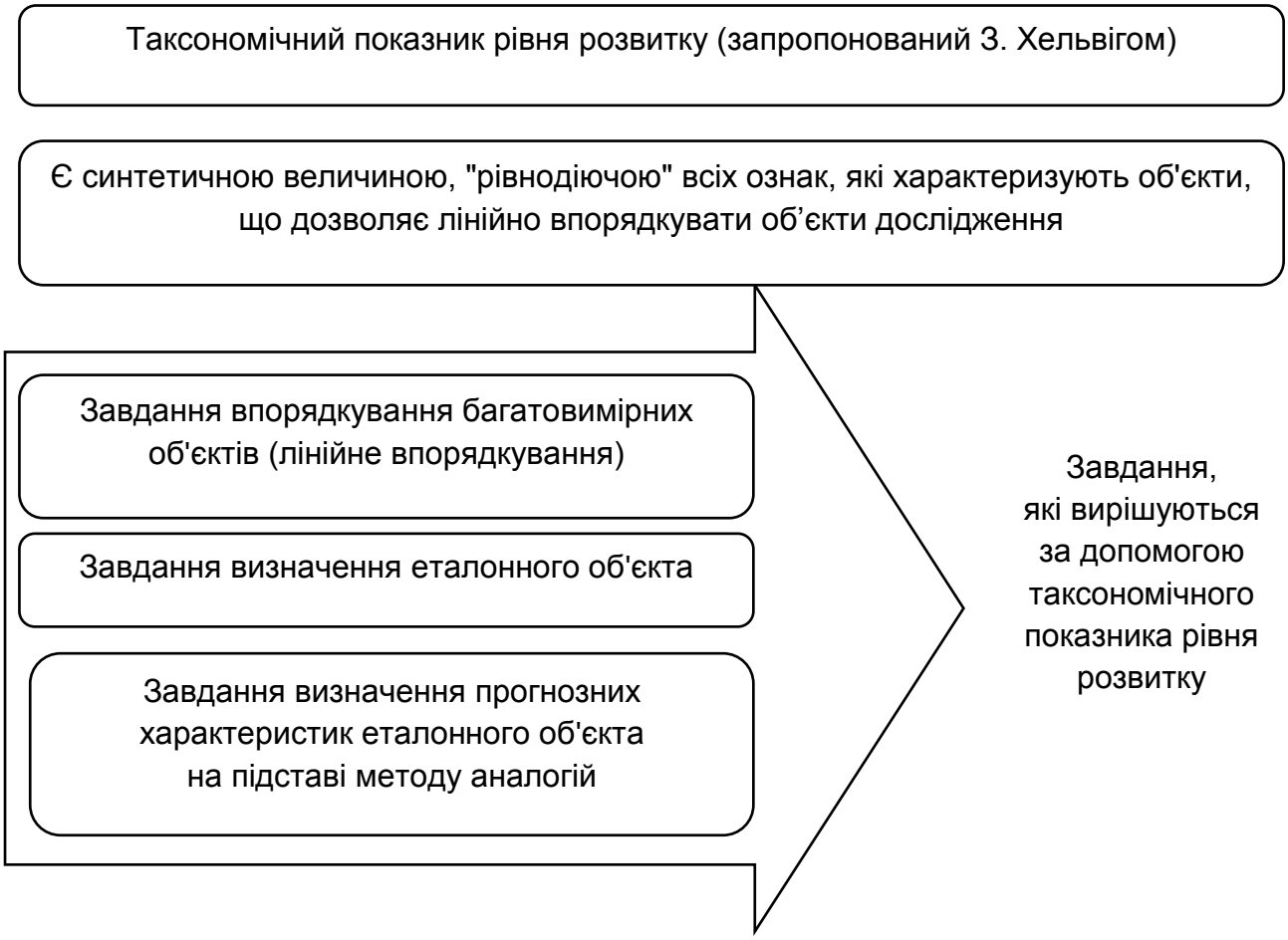


Рис. 6.2. **Завдання, які вирішуються за допомогою таксономічного показника рівня розвитку**

Алгоритм побудови таксономічного показника рівня розвитку наведений на рис. 6.3. Першим кроком процесу побудови таксономічного показника рівня розвитку (див. рис. 6.3) є визначення елементів матриці спостережень, що можна подати таким чином:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{\omega 1} & x_{\omega 2} & \dots & x_{\omega j} & \dots & x_{\omega m} \end{bmatrix},$$

де ω – кількість досліджуваних об'єктів;

m – кількість ознак;

x_{ij} – значення j-ї ознаки для і-го об'єкта.

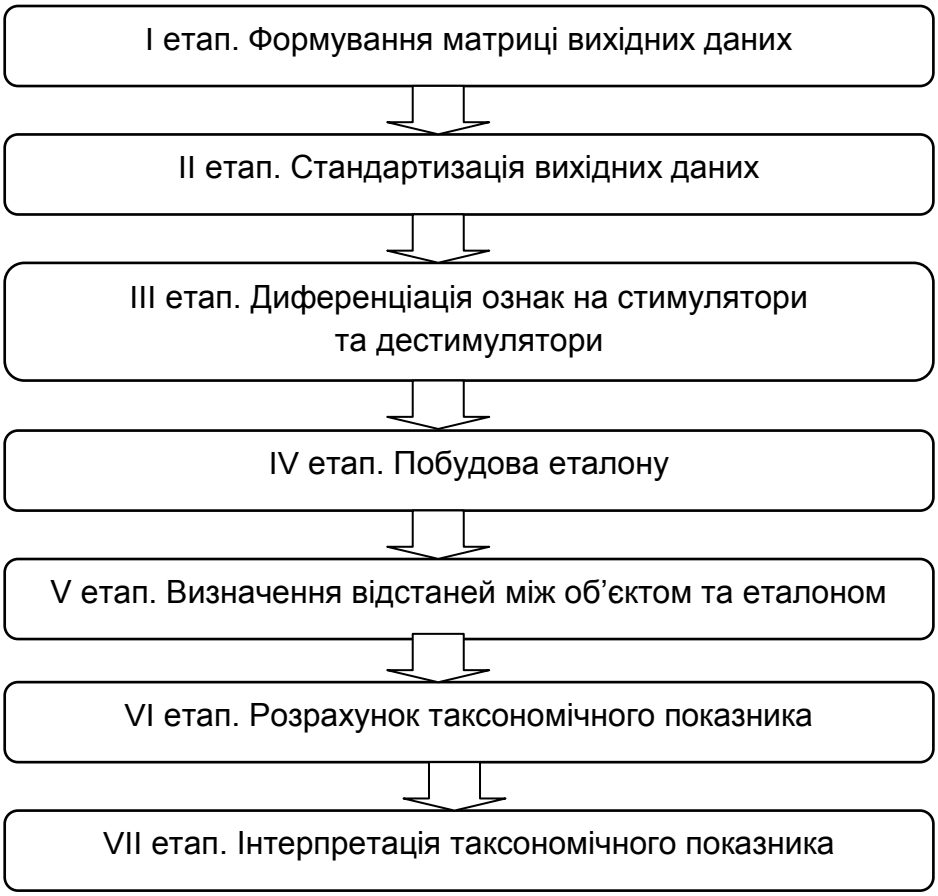


Рис. 6.3. **Алгоритм побудови таксономічного показника рівня розвитку**

Оскільки ознаки, включені в матрицю спостережень, неоднорідні, проводиться стандартизація їх значень за формулою:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad j = 1, 2, \dots, m, \qquad (6.1)$$

де $\bar{x}_j$ – середнє арифметичне значення j-ї ознаки;

S_j – стандартне відхилення j-ї ознаки;

x_{ij} – стандартизоване значення j-ї ознаки для і-го об'єкта.

Наступний крок у розглянутій процедурі (див. рис. 6.3) полягає в диференціації ознак матриці спостережень. Усі змінні розподіляють на стимулятори та дестимулятори. Підставою розподілу ознак на дві групи слугує характер впливу кожного з них на рівень розвитку досліджуваних об'єктів. Ознаки, що позитивно, стимулююче впливають на рівень розвитку об'єктів, називаються стимуляторами, на відміну від ознак-дестимуляторів. Розрізнення ознак на стимулятори та дестимулятори є основою для побудови так званого еталона розвитку, що є точкою з координатами:

$$P_0(z_{01}, z_{02}, \dots, z_{0m}),$$

де $z_{0s} = \max_r z_{rs}$, якщо $s \in I$;

$z_{0s} = \min_r z_{rs}$, якщо $s \notin I$, $(s = 1, \dots, m)$;

I – множина;

z_s – стандартизоване значення ознаки *s* для об'єкта r.

Відстань між окремими точками-одиницями та точкою P_0, що є еталоном розвитку, позначають c_{i0} і розраховують таким чином:

$$c_{io} = \sqrt{\sum_{j=1}^{m}(Z_{ij} - Z_{oj})^2}. \quad (6.2)$$

Обчислені відстані є вхідними величинами, що використовують для розрахунку показника рівня розвитку:

$$d_i^* = 1 - \frac{c_{i0}}{c_0}, \quad (6.3)$$

де $c_0 = \bar{c}_0 + 2 \cdot S_0$;

$\bar{c}_0 = \frac{1}{w}\sum_{i=1}^{w} c_{io}$;

$S_0 = \sqrt[2]{\frac{1}{w}\sum_{i=1}^{w}(c_{io} - \bar{c}_o)}$.

Інтерпретація показника рівня розвитку така: чим ближче значення показника рівня розвитку до одиниці, тим на більш високому рівні розвитку перебуває об'єкт. Таким чином, можна виділити переваги застосування таксономічного показника рівня розвитку в економічних дослідження (рис. 6.4).

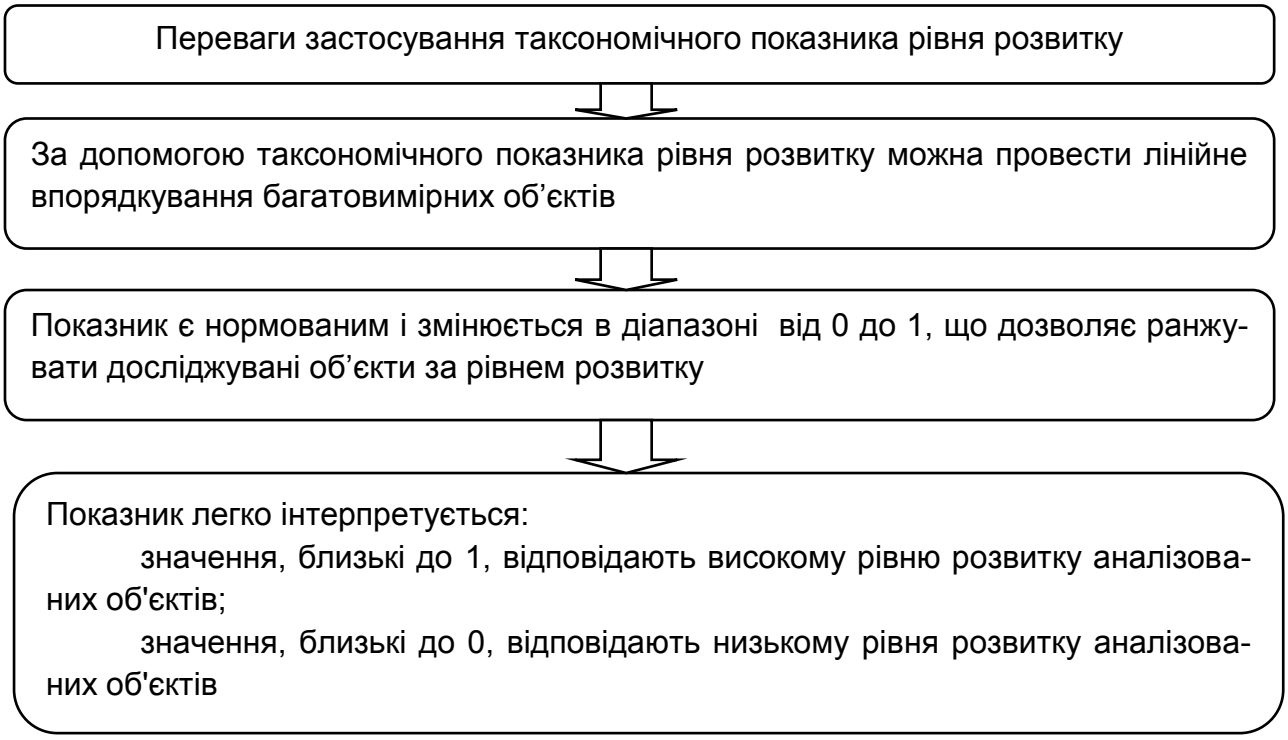


Рис. 6.4. **Переваги застосування таксономічного показника рівня розвитку в економічних дослідженнях**

Слід зазначити, що таксономічний показник рівня розвитку універсальний. Він може використовуватися не тільки в дослідженнях статистичних одиниць, що належать до сукупності, але й для аналізу властивостей однієї одиниці. В останньому випадку властивості одиниці характеризуються значеннями ознак, заданими у вигляді часових рядів:

$$= \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{\omega 1} & x_{\omega 2} & \dots & x_{\omega j} & \dots & x_{\omega m} \end{bmatrix},$$

де ω – кількість досліджуваних періодів часу;

m – кількість ознак;

x_{ij} – значення j-ї ознаки для і-го об'єкта.

Значення інтегрального показника, знайдені на основі наведеної матриці вихідних даних відповідно до алгоритму на рис. 6.3 описують динаміку змін досліджуваних наборів ознак. Вони в узагальненій формі представляють зміни, що відбуваються в аналізованому явищі.

6.3. Приклад застосування таксономічного показника рівня розвитку в економічних дослідженнях

Розглянемо реалізацію алгоритму таксономічного показника рівня розвитку на прикладі. У табл. 6.1 наведені показники ліквідності та фінансової стійкості промислових підприємств. Необхідно впорядкувати дані підприємства на основі методу рівня розвитку та дати економічну інтерпретацію отриманим результатам.

Таблиця 6.1

Вихідні дані

№ підприємства	Показники ліквідності			Показники фінансової стійкості		
	x_1	x_2	x_3	x_4	x_5	x_6
1	2,713	0,984	0,001	0,855	0,024	0,930
2	3,079	1,006	0,002	0,979	0,026	0,926
3	3,553	1,091	0,020	1,174	0,024	0,634
4	2,304	1,244	0,015	0,885	0,002	0,885
5	1,572	0,877	0,002	0,460	0,022	0,910
6	1,659	0,751	0,001	0,410	0,032	0,912

Умовні позначення: x_1 – коефіцієнт поточної ліквідності; x_2 – коефіцієнт швидкої ліквідності; x_3 – коефіцієнт абсолютної ліквідності; x_4– коефіцієнт забезпеченості власними обіговими коштами; x_5 – коефіцієнт маневреності власного капіталу; x_6 – коефіцієнт автономії.

Розв'язання

Оскільки показники мають різні одиниці вимірювання, здійснюємо їх стандартизацію за формулою (6.1). Стандартизовані значення ознак наведені в табл. 6.2.

Таблиця 6.2

Стандартизовані значення ознак

№ підприємства	Показники ліквідності			Показники фінансової стабільності		
	x_1	x_2	x_3	x_4	x_5	x_6
1	0,296	-0,048	-0,692	0,204	0,228	0,556
2	0,761	0,081	-0,574	0,617	0,424	0,521
3	1,364	0,581	1,563	1,268	0,228	-2,022
4	-0,224	1,479	0,969	0,304	-1,922	0,164
5	-1,154	-0,677	-0,574	-1,113	0,033	0,382
6	-1,044	-1,417	-0,692	-1,280	1,010	0,399

Відповідно до розглянутого алгоритму (див. рис. 6.3) наступним кроком є класифікація ознак на стимулятори та дестимулятори. Підставою для розподілу є характер впливу ознак на рівень стійкості фінансового стану підприємства. Ознаки, зростання значень яких свідчить про підвищення рівня стійкості фінансового стану підприємства, належать до ознак-стимуляторів. Показники, зростання значень яких свідчить про зниження рівня фінансової стійкості, належать до ознак-дестимуляторів. Результати класифікації ознак наведені в табл. 6.3.

Таблиця 6.3

Класифікація ознак

Показники	Характер впливу	Група	Показники	Характер впливу	Група
Коефіцієнт поточної ліквідності	Позитивний	Стимулятор	Коефіцієнт забезпеченості власними обіговими коштами	Позитивний	Стимулятор
Коефіцієнт швидкої ліквідності	Позитивний	Стимулятор	Коефіцієнт маневреності власного капіталу	Позитивний	Стимулятор
Коефіцієнт абсолютної ліквідності	Позитивний	Стимулятор	Коефіцієнт автономії	Позитивний	Стимулятор

Розподіл ознак на стимулятори та дестимулятори є основою для визначення еталона розвитку (табл. 6.4).

Таблиця 6.4

Визначення еталона розвитку

№ підприємства	Показники ліквідності			Показники фінансової стабільності		
	x_1	x_2	x_3	x_4	x_5	x_6
1	0,296	-0,048	-0,692	0,204	0,228	0,556
2	0,761	0,081	-0,574	0,617	0,424	0,521
3	1,364	0,581	1,563	1,268	0,228	-2,022
4	-0,224	1,479	0,969	0,304	-1,922	0,164
5	-1,154	-0,677	-0,574	-1,113	0,033	0,382
6	-1,044	-1,417	-0,692	-1,280	1,010	0,399
Об'єкт – еталон (P_0)	max= 1,364	max= 1,479	max= 1,563	max= 1,268	max= 1,010	max= 0,556

На наступному кроці розраховуються відстані між окремими підприємствами й об'єктом – еталоном за формулою (6.2). Значення відстаней, їх середнє арифметичне значення та стандартне відхилення наведені в табл. 6.5.

Таблиця 6.5

Відстані між об'єктами і об'єктом-еталоном

№ підприємства	Відстань (c_{io})	Характеристики	Відстань (c_{io})
1	3,210	$\bar{c}_0$	3,692
2	2,766	S_0	0,895
3	2,840	c_0	5,481
4	3,543		
5	4,713		
6	5,078		

Отримані відстані слугують вихідними величинами, що використовую для розрахунку показника рівня розвитку за формулою (6.3). Так, значення показника рівня розвитку для першого підприємства визначається таким чином:

$$d_i^* = 1 - \frac{3{,}21}{5{,}481} = 0{,}414.$$

Значення показника рівня розвитку для інших підприємств подані в табл. 6.6.

Таблиця 6.6

Значення показника рівня розвитку

№ підприємства	d_i^*	Рейтинг підприємства	№ підприємства	d_i^*	Рейтинг підприємства
1	0,414	3	4	0,354	4
2	0,495	1	5	0,140	5
3	0,482	2	6	0,073	6

Інтерпретація показника рівня розвитку така: чим ближче значення показника рівня розвитку до одиниці, тим вище рівень стійкості фінансового стану підприємства. Таким чином, найбільш стійким є друге підприємство, а найгірше фінансове становище спостерігається у шостого підприємства.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. Дайте визначення редукції.
2. Наведіть класифікацію методів редукції.
3. У чому сутність методів повної редукції?
4. У чому сутність методів неповної редукції?
5. Наведіть приклади застосування методу рівня розвитку в економічних дослідженнях.
6. Які підходи використовують для формування еталонної точки?
7. Яким чином здійснюється розподіл ознак на стимулятори та дестимулятори?
8. Яку міру відстані застосовують під час розрахунку показника рівня розвитку?
9. Як інтерпретувати значення показника рівня розвитку?

Тестові завдання

1. Ознаку, більшим значенням якої відповідають більші значення таксономічного показника, називають:

а) дестимулятором;

б) номінатором;

в) стимулятором.

2. Ознаку, меншим значенням якої відповідають більші значення таксономічного показника, називають:

а) дестимулятором;

б) номінатором;

в) стимулятором.

3. Таксономічний показник – це:

а) синтетична ознака;

б) індивідуальна діагностична ознака;

в) факторна ознака.

4. Таксономічний показник змінюється в межах:

а) [0,1];

б) [-1;1];

в) [-1;0];

г)[-1;+1].

5. Які з методів належать до методів повної редукції:

а) метод головних компонент;

б) факторний аналіз;

в) багатовимірне шкалювання;

г) таксономічний показник рівня розвитку?

6. Які з методів можна віднести до методів неповної редукції:

а) метод головних компонент;

б) факторний аналіз;

в) багатовимірне шкалювання;

г) таксономічний показник рівня розвитку?

7. Редукція – це:

а) процес зменшення аналізованої безлічі даних до розміру, оптимального з погляду розв'язуваного завдання та використовуваної аналітичної моделі;

б) процес збільшення аналізованої безлічі даних до розміру, оптимального з погляду розв'язуваного завдання та використовуваної аналітичної моделі.

8. Зниження розмірності може знадобитися, якщо (кілька правильних відповідей):

а) завдання можна розв'язати так само ефективно, але використовуючи для цього меншу кількість даних;

б) завдання неможливо розв'язати без проведення зниження розмірності даних;

в) завдання можна розв'язати ефективно, не використовуючи для цього меншу кількість даних.

9. Значення інтегрального показника, рівне 0,9, говорить про:

а) низький рівень розвитку об'єкта;

б) високий рівень розвитку об'єкта;

в) середній рівень розвитку об'єкта.

10. Відстань між об'єктом і точкою еталона розраховують за формулою:

а) $d_{i0} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{ok})^2}$;

б) $d_{i0} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{0k})}$;

в) $d_{i0} = \sqrt{(x_{0k} - x_{ik})^2}$.

11. Діагностичні ознаки мають такі властивості:

а) ознаки що корельовані або слабо корельовані між собою;

б) сильно корельовані з ознаками, що не входять в діагностичний набір;

в) усі відповіді правильні.

12. Матриця, елементи якої відбивають ступінь близькості об'єктів, показників, ознак у дослідній сукупності, є:

а) матрицею відстаней;

б) матрицею коваріацій;

в) матрицею дисперсій;

г) матрицею контингеції.

13. Формат відстаней, розрахованих для матриці відстаней, – це:

а) міра відстані;

б) міра редукції;

в) вектор відстаней;

г) вектор еталона.

14. Процес зведення показників, які мають різні одиниці вимірювання, до загального типу вимірювань – це:

а) стандартизація;

б) неповна редукція;

в) повна редукція;

г) формалізація.

15. Процесом побудови деяких синтетичних величин, що мають багатоознакову природу, у вигляді деякої функції $y = (y_1, y_2, \dots, y_q)$*, яка відбиває вплив усіх ознак, і дозволяє впорядкувати досліджувані об'єкти, є:*

а) неповна редукція;

б) повна редукція;

в) стандартизація;

г) формалізація.

16. Еталоном є:

а) об'єкт, що має найбільш якісний набір ознак вихідної вибірки;

б) об'єкт, що має максимальні значення ознак вихідної вибірки;

в) об'єкт, який має набір мінімальних значень ознак вихідної вибірки.

17. Рівень розвитку об'єкта вище, чим ближче значення інтегрального показника до:

а) 1;

б) 0;

в) -1.

Практичні завдання

Завдання 1. У табл. 6.7 наведено показники ефективності діяльності промислових підприємств за такими групами: ліквідності, ділової активності, фінансової стійкості, рентабельності, майнового стану. Необхідно здійснити впорядкування підприємств за рівнем стійкості фінансового стану на основі методу рівня розвитку, дати економічну інтерпретацію отриманим результатам.

Умовні позначення: x_1 – коефіцієнт поточної ліквідності; x_2 – коефіцієнт швидкої ліквідності; x_3 – коефіцієнт абсолютної ліквідності; x_4 – коефіцієнт забезпеченості власними обіговими коштами; x_5 – коефіцієнт маневреності власного капіталу; x_6 – коефіцієнт автономії; x_7 – коефіцієнт оборотності активів; x_8 – коефіцієнт оборотності оборотних коштів; x_9 – коефіцієнт оборотності запасів; x_{10} – коефіцієнт рентабельності активів; x_{11} – коефіцієнт рентабельності власного капіталу.

Таблиця 6.7

Вихідні дані

№ п/п	Показники ліквідності			Показники фінансової стійкості			Показники ділової активності			Показники рентабельності	
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	2,71	0,98	0,001	0,85	0,024	0,93	0,05	0,28	0,99	0,05	0,06
2	3,08	1,06	0,002	0,97	0,026	0,92	0,03	0,18	0,83	0,02	0,02
3	3,55	1,09	0,020	1,17	0,024	0,63	0,02	0,15	0,73	0,06	0,06
4	2,30	1,24	0,015	0,88	0,002	0,88	0,16	0,89	1,77	0,04	0,05
5	1,57	0,87	0,002	0,46	0,022	0,91	0,08	0,44	1,18	0,08	0,09
6	1,65	0,75	0,001	0,41	0,032	0,91	0,04	0,20	0,91	0,07	0,08
7	2,24	0,78	0,000	0,59	0,035	0,93	0,07	0,93	0,02	0,03	0,03
8	3,04	0,91	0,000	0,90	0,032	0,90	0,10	0,90	0,01	0,01	0,02
9	3,85	1,09	0,000	1,25	0,026	0,95	0,05	0,95	0,03	0,07	0,08
10	4,50	1,26	0,000	1,54	0,020	0,94	0,06	0,94	0,02	0,12	0,13

Завдання 2. На підставі наведених даних (табл. 6.8) слід провести впорядкування підприємств, кожне з яких характеризується трьома ознаками: x_1 – інвестиції в основний капітал (млн грн), x_2 – чистий прибуток (млн грн), і x_3 – коефіцієнт заборгованості (%).

Упорядкування здійсніть за допомогою таксономічного показника рівня розвитку. Побудуйте рейтинг підприємств за рівнем конкурентоспроможності, надайте економічну інтерпретацію.

Таблиця 6.8

Вихідні дані

№ об'єкта	Ознаки		
	x_1	x_2	x_3
1	15,5	12,2	2,5
2	16,5	8,3	1,8
3	5,4	7,9	4,4
4	10,2	9,2	2,5
5	11,3	8,4	3,2
6	9,4	7,1	2,25
7	20,8	10,1	1,3
8	15,1	6,6	3,8
9	12,4	8,7	5,4
10	8,7	7,5	2,3

Завдання 3. У табл. 6.9. наведені значення евклідових відстаней між об'єктом (банком) і точкою-еталоном. Побудуйте рейтинг банків за рівнем фінансової стійкості та надайте інтерпретацію.

Таблиця 6.9

Вхідні дані

№	1	2	3	4	5	6	7	8	9	10
D_{i0}	15,5	16,5	5,4	10,2	11,3	9,4	20,8	15,1	12,4	8,7

Завдання 4. У табл. 6.10 наведені показники привабливості регіональних сегментів ринку для збуту сільськогосподарчої техніки. Необхідно провести впорядкування регіонів на основі таксономічного показника рівня розвитку. Побудувати рейтинг регіональних ринків збуту за рівнем привабливості, привести інтерпретацію.

Таблиця 6.10

Вхідні дані

Області	Показники				
	x_1	x_2	x_3	x_4	x_5
1	2	3	4	5	6
Вінницька	1744,5	51,8	34,2	66,02317	39,827
Волинська	503,1	55,94	4,8	8,580622	47,244
Дніпропетровська	1571,4	96,78	121,4	125,4391	28,056
Донецька	1011	45,41	41,7	91,82999	34,127
Житомирська	721,4	46,44	22	47,37295	41,108
Закарпатська	166,8	109,82	6,2	5,645602	50,98
Запорізька	1150,1	131,57	61,7	46,89519	28,73
Івано–Франківська	253,1	45,38	9,5	20,93433	57,443
Київська	1430,5	83,06	40,2	48,39875	50,515
Кіровоградська	1378,4	232,68	83,5	35,8862	29,401
Луганська	41,7	62,86	48,9	77,79192	35,319
Львівська	481,4	45,19	19,8	43,815	55,065
Миколаївська	921,3	359,89	50,8	14,11542	36,443
Одеська	1581,3	238,13	60,3	25,3223	44,502
Полтавська	1407,8	95,13	37,7	39,62998	31,738
Рівненська	494,7	47,16	21,8	46,22561	49,169
Сумська	967,6	63,13	22,4	35,48234	26,386

Закінчення табл. 6.10

1	2	3	4	5	6
Тернопільська	740,6	65,33	14,7	22,50115	63,307
Харківська	1267,6	45,98	38,9	84,602	22,747
Херсонська	1166,3	266,92	48,4	18,13277	28,772
Хмельницька	1241,7	78,95	27,1	34,32552	65,835
Черкаська	1594,1	72,91	18,1	24,82513	37,312
Чернівецька	325,9	70,11	4	5,70532	48,48
Чернігівська	832,3	41,3	11,5	27,84504	30,951

Умовні позначки: x_1 – виробництво основних видів продукції зернових культур, тис. т; x_2 – загальні посівні площі, тис. га; x_3 – посівні площі зернових культур, тис. га; x_4 – відсоток посівних площ зернових культур, % до загальної посівної площі; x_5 – забезпеченість регіонів справними комбайнами, %.

Завдання 5. У табл. 6.11 наведені вхідні дані для оцінювання рівня інвестиційної привабливості підприємств. Проведіть упорядкування підприємств за допомогою таксономічного показника рівня розвитку та надайте економічну інтерпретацію.

Таблиця 6.11

Вихідні дані

№ п/п	Підприємства	Показники					
		x_1	x_2	x_3	x_4	x_5	x_6
1	2	3	4	5	6	7	8
1	ДП Електроважмаш	0,6965	0,0000	0,6405	0,0032	0,1521	8,8297
2	ВАТ Укрелектромаш	0,9565	0,0023	0,7204	0,0027	0,0403	28,0526
3	ЗАТ Південкабель	4,2908	0,1912	0,8853	0,0145	0,0198	171,6810
4	ВАТ Електромашина	2,2175	0,0000	0,7564	0,0010	0,0101	15,9000
5	ВАТ Гідропривід	5,4143	0,0000	0,8941	0,0014	2,1099	16,7550
6	ВАТ Харківський завод агрегатних верстатів	1,7012	0,0303	0,6829	0,0004	1,2427	18,6070
7	Харківський верстатобудівний завод	2,6233	0,0000	0,8379	0,0001	0,2040	13,9913
8	ВАТ Харківський завод технологічного обладнання	5,2549	0,0051	0,9095	0,0039	3,2961	16,5300
9	ВАТ Харківський завод штампів і пресформ	2,5875	0,0000	0,7330	0,0000	2,3687	7,7700
10	ЗАТ ХВЗ ім. Петровського	0,3582	0,0646	0,5175	0,0002	0,4547	8,8349

Закінчення табл. 6.11

1	2	3	4	5	6	7	8
11	ВАТ Чугуївська паливна апаратура	0,5900	0,0000	0,1652	0,0003	1,5550	5,8378
12	ВАТ ХЗТСШ	0,3154	0,0000	0,1007	0,0002	1,4049	4,2503
13	ЗАТ Куп'янський ливарний завод	0,1545	0,0000	0,6139	0,0001	1,5717	5,4940
14	ВАТ Дергачівський завод турбокомпресорів	0,7324	0,0000	0,6096	0,0001	1,7507	14,2259
15	ВАТ Серп і Молот	0,4126	0,0000	0,7977	0,0024	1,9425	6,8861
16	ВАТ Завод ім. Фрунзе	8,3631	0,0973	0,9210	0,0068	1,9026	82,2930
17	ВАТ Автрамат	1,2104	0,0000	0,7633	0,0008	0,7369	24,5190
18	ЗАТ Лозівський завод Трактородеталь	12,0745	0,0361	0,9410	0,0128	0,6198	32,1249
19	ВАТ ЛЗМК	2,6799	0,0048	0,9594	0,0001	0,7220	27,9055
20	АТЗТ ХЗЕМВ	3,5309	0,0069	0,8744	0,0000	1,5958	13,9965

Умовні позначення: x_1 – коефіцієнт абсолютної ліквідності; x_2 – коефіцієнт рентабельності капіталу; x_3 – коефіцієнт автономії; x_4 – коефіцієнт відновлення основних фондів; x_5 – коефіцієнт фондовіддачі; x_6 – середній виробіток.

Розділ 7. Методи неповної редукції. Метод центра ваги

7.1. Поняття системи діагностичних ознак.

7.2. Метод "центра ваги".

7.3. Приклад застосування методу "центра ваги" в економічних дослідженнях.

7.4. Оцінювання якості діагностичного простору ознак.

Ключові слова: методи неповної редукції ознак; діагностична ознака; властивості діагностичних ознак; методи вибору репрезентантів груп; метод "центра ваги"; процедури стандартизації; міри відстані; правила вибору репрезентантів груп; оцінка якості простору діагностичних ознак.

Література: [7; 22; 33; 43; 44; 46].

7.1. Поняття системи діагностичних ознак

Як зазначалося в розділі 6, *методи неповної редукції припускають формування системи так званих діагностичних ознак, якими є деякі з вихідних ознак.* Для цього первинний набір q ознак $y = (y_1, y_2, \dots, y_q)$ заміняється набором s діагностичних ознак $x = (x_1, x_2, \dots, x_s)$, $(s < q)$. Слід зазначити, що зі скороченням кількості змінних треба дотримуватись деяких вимог для того, щоб створюваний опис не спотворював дійсності. Система діагностичних ознак складається з елементів, що найбільш повно характеризують об'єкти, формуючи якнайменш численний набір. Наведені вимоги виконуються тоді, коли діагностичні ознаки мають такі властивості:

некорельовані або слабкокорельовані між собою;

сильно корельовані з ознаками, що не входять у діагностичний набір;

дозволяють розділяти досліджувані одиниці, тобто характеризуються високою варіацією за всіма одиницями безлічі та достатньо низькою варіацією за одиницями всередині виділених груп;

не мають зовнішніх впливів.

Основними є перші дві властивості, оскільки вони виключають взаємно дубльовані ознаки, а також забезпечують вибір таких, що якнайкраще представляють усі ті елементи, які не входять в отриманий перелік. Для того щоб виконати першу вимогу, використовують методи розбивки на групи, які докладно описано в розділі 4. Використання методів вибору репрезентантів груп дозволяє створити систему індикаторів, що відповідає другій вимозі. Таким чином, якщо в сукупності ознак виділити однорідні групи, а потім вибрати репрезентанти в кожній з них, то отримана система показників відповідає першим двом умовам.

7.2. Метод "центра ваги"

Одним з найпоширеніших методів вибору репрезентантів груп є метод "центра ваги". Цільове призначення методу полягає в виборі репрезентантів, тобто ознак, які передають найбільш суттєві особливості багатовимірного набору вхідних ознак. Алгоритм методу наведений на рис. 7.1.

Рис. 7.1. **Алгоритм методу центра ваги**

На першому кроці алгоритму формуються матриці вихідних даних за кожною групою показників стану об'єкта дослідження $Y_1, Y_2, \dots, Y_q$, де q – кількість груп показників. Для k-ї групи показників структура цієї матриці може бути визначена таким чином:

$$Y_k = (y_{ij})_k, i = [1; m], j = [1; n],$$

де y_{ij} – значення i-го показника в j-му досліджуваному періоді (або для j-го досліджуваного об'єкта);

m – кількість показників, що входять у k-ту групу;

n – кількість досліджуваних періодів (або об'єктів).

Можливі типи матриць вихідних даних наведені на рис. 7.2.

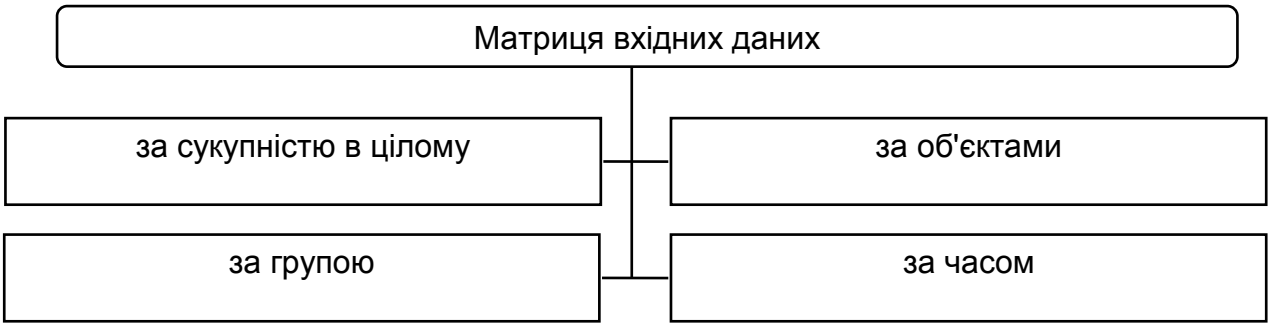


Рис. 7.2. **Типи матриць вхідних даних**

На другому кроці здійснюється процедура стандартизації за формулою (6.1). Альтеративними засобами стандартизації ознак є:

$$z_{ij} = \frac{x_{ij}}{x_j^{max} - x_j^{min}};$$

$$z_{ij} = \frac{x_{ij}}{\overline{x_j}}.$$

Результатом цього кроку є набір матриць стандартизованих значень показників кожної групи $Z_1, Z_2, \dots, Z_q$.

Описані обчислювальні процедури є основою для розрахунку матриць відстаней $P_1, P_2, \dots, P_q$, елементи яких відображають ступінь близькості показників усередині кожної групи. Розглядаються такі міри відстані (рис. 7.3).

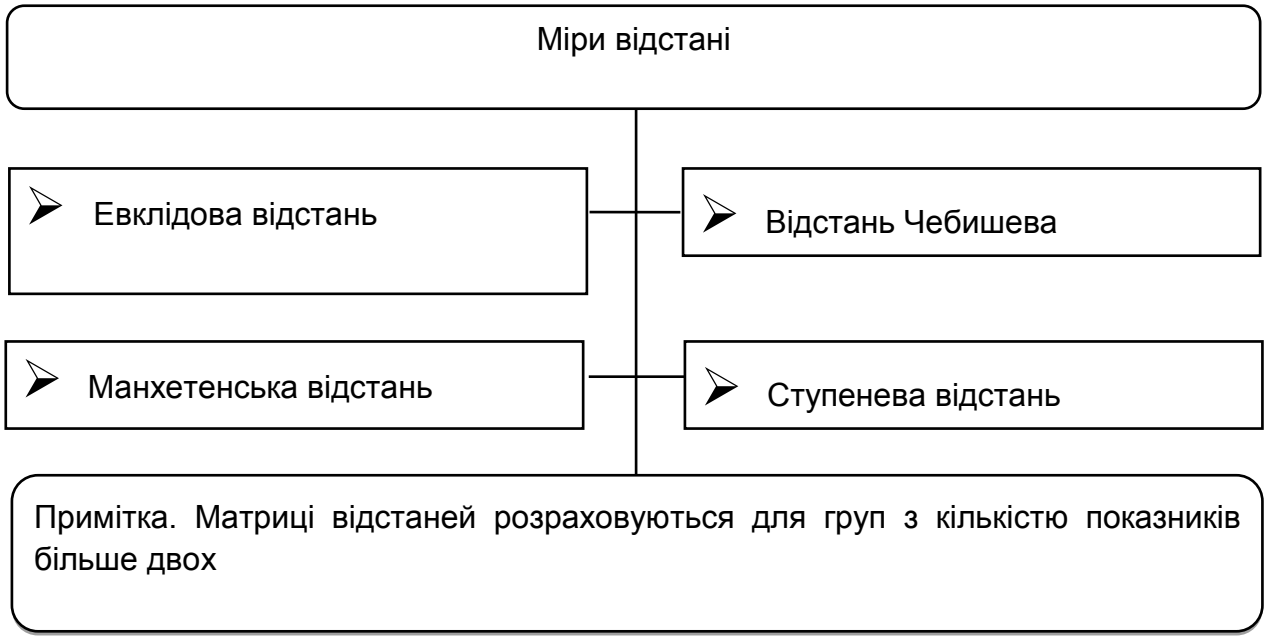


Рис. 7.3. **Міри відстаней**

Формули розрахунку різних мір відстаней наведені в табл. 7.1.

На четвертому кроці (див. рис. 7.1) здійснюється вибір так званих показників-репрезентантів груп з найбільш значущою інформацією, властивою групі, за певними правилами (рис. 7.4 – 7.6).

Таблиця 7.1

Міри відстані між об'єктами

Міра відстані між об'єктами	Формула розрахунку	Умови застосування
Евклідова відстань	$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$, де x_{jk} – значення k-го показника, відповідно, у i-го та j-го об'єктів	Застосовується у випадку, якщо компоненти вектора спостережень X однорідні за фізичним змістом і однаково важливі для класифікації. Є найбільш використовуваною
Зважена евклідова відстань	$d_{ij} = \sqrt{\sum_{k=1}^{m} w_k(x_{ik} - x_{jk})^2}$, де w_k – вага, що приписується k-му показнику	Застосовується у випадку, коли кожному компоненту вектора X приписується вага, пропорційна ступеню важливості ознаки $0 \le w_k \le 1$
Манхетенська відстань	$d_{ij} = \frac{1}{m}\sum_{k=1}^{m}\lvert x_{ik} - x_{jk}\rvert$	У більшості випадків ця міра відстані приводить до таких же результатів, що й евклідова відстань. Однак для цієї міри вплив окремих великих різниць (викидів) зменшується, оскільки вони не зводяться у квадрат
Відстань Чебишева	$d(z_i, z_j) = max\lvert z_{il} - z_{jl}\rvert$	Ця відстань використовується, коли дослідник прагне визначити два об'єкти як "різні", якщо вони відрізняться за якою-небудь однією координатою (яким-небудь одним виміром)
Ступенева відстань	$d(z_i, z_j) = (\sum_i \lvert z_{il} - z_{jl}\rvert^p)^{1/r}$, де p і r – параметри, які визначаються дослідником	Параметр *p* відповідає за поступове зважування різниць за окремими координатами. Параметр *r* відповідає за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметра – *r* і *p* дорівнюють 2, то ця відстань збігається з відстанню Евкліда

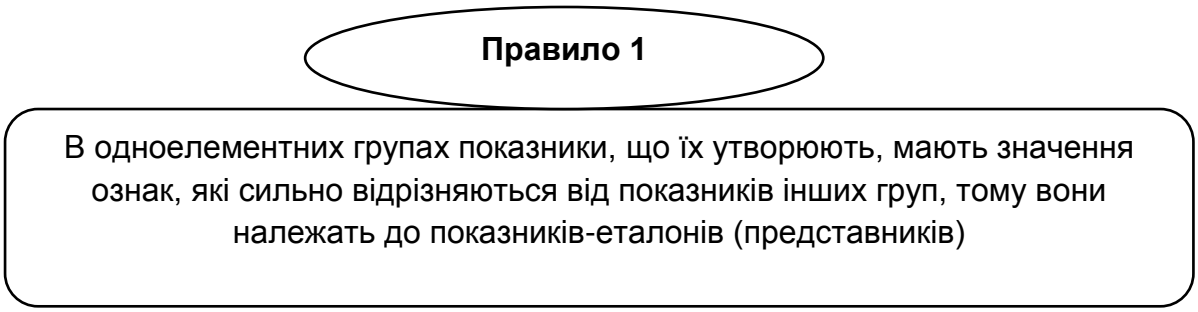


Рис. 7.4. **Правило вибору показників-репрезентантів для групи з одного елемента**

У групах, де кількість показників більше двох, розраховується сума відстаней кожного показника до інших показників групи: $\rho_i = \sum_{j=1}^{m} \rho(z_i, z_j)$, де m – число показників групи. До складу показників-репрезентантів входить показник з найменшою сумою відстаней: $\rho_s = \min \rho_i$ (рис. 7.5).

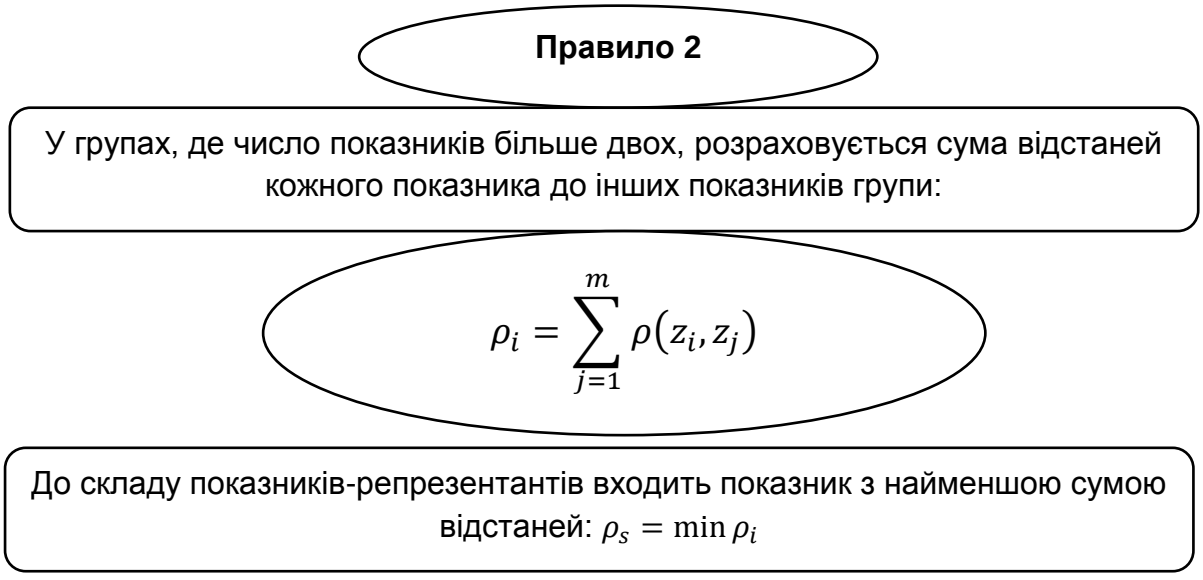


Рис. 7.5. **Правило вибору репрезентантів для групи із кількістю елементів більше двох**

У групах, де кількість показників дорівнює двом, визначається сума відстаней показників, що входять у групу, від показників-репрезентантів: $\sum_{j=1}^{k} \rho(z_i, z_j)$, де k – кількість показників-репрезентантів. До репрезентантів групи, де кількість показників дорівнює двом, належить той показник,

у якого сума відстаней від відособлених елементів і елементів-репрезентантів, виділених із груп елементів із числом більше двох, максимальна: $\rho_s = \max_i p_i$ (рис. 7.6).

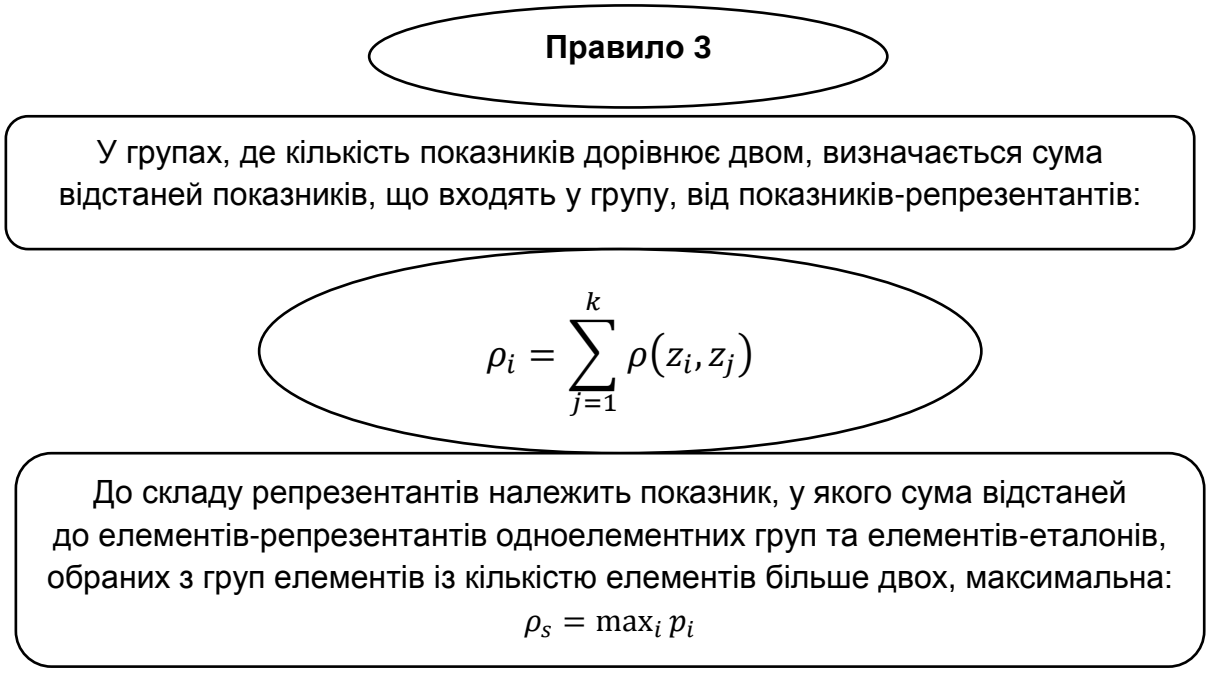


Рис. 7.6. **Правило вибору показників-репрезентантів для групи з кількістю елементів, яке дорівнює двом**

Таким чином, результатом четвертого кроку є набір показників-репрезентантів $x = (x_1, x_2, \ldots, x_k)$, що описують найбільш важливі аспекти стану об'єкта дослідження.

7.3. Приклад застосування методу "центра ваги" в економічних дослідженнях

Розглянемо реалізацію алгоритму методу на такому прикладі. У табл. 7.2, 7.3 наведені показники ліквідності, ділової активності, фінансової стійкості, рентабельності промислових підприємств. Необхідно вибрати найбільш значущі для оцінювання фінансового стану цих підприємств показники (показники-репрезентанти) за допомогою методу "центра ваги".

Таблиця 7.2

Вхідні дані

№ підприємства	Показники ліквідності			Показники фінансової стійкості		
	x_1	x_2	x_3	x_4	x_5	x_6
1	2,713	0,984	0,001	0,855	0,024	0,930
2	3,079	1,006	0,002	0,979	0,026	0,926
3	3,553	1,091	0,020	1,174	0,024	0,634
4	2,304	1,244	0,015	0,885	0,002	0,885
5	1,572	0,877	0,002	0,460	0,022	0,910
6	1,659	0,751	0,001	0,410	0,032	0,912

Таблиця 7.3

Вхідні дані

№ підприємства	Показники ділової активності			Показники рентабельності	
	x_7	x_8	x_9	x_{10}	x_{11}
1	0,052	0,287	0,998	0,005	0,006
2	0,032	0,189	0,837	0,023	0,025
3	0,024	0,158	0,735	0,058	0,064
4	0,164	0,897	1,775	0,044	0,050
5	0,087	0,447	1,184	0,084	0,093
6	0,043	0,204	0,910	0,075	0,082

Умовні позначення: x_1 – коефіцієнт поточної ліквідності; x_2 – коефіцієнт швидкої ліквідності; x_3– коефіцієнт абсолютної ліквідності; x_4 – коефіцієнт забезпеченості власними обіговими коштами; x_5 – коефіцієнт маневреності власного капіталу; x_6 – коефіцієнт автономії; x_7 – коефіцієнт оборотності активів; x_8 – коефіцієнт оборотності обігових коштів; x_9 – коефіцієнт оборотності запасів; x_{10} – коефіцієнт рентабельності активів; x_{11} – коефіцієнт рентабельності власного капіталу.

Розв'язання

Оскільки показники мають різні одиниці вимірювання, здійснюємо їх стандартизацію за формулою (6.1). Стандартизовані значення ознак наведені в табл. 7.4.

Таблиця 7.4

Стандартизовані значення ознак

№ підприємства	Показники ліквідності			Показники фінансової стійкості		
	x_1	x_2	x_3	x_4	x_5	x_6
1	0,32	-0,05	-0,76	0,22	0,25	0,61
2	0,83	0,09	-0,63	0,68	0,46	0,57
3	1,49	0,64	1,71	1,39	0,25	-2,21
4	-0,25	1,62	1,06	0,33	-2,11	0,18
5	-1,26	-0,74	-0,63	-1,22	0,04	0,42
6	-1,14	-1,55	-0,76	-1,40	1,11	0,44

№ підприємства	Показники ділової активності			Показники рентабельності	
	x_7	x_8	x_9	x_{10}	x_{11}
1	-0,31	-0,30	-0,22	-1,56	-1,55
2	-0,73	-0,68	-0,69	-0,91	-0,93
3	-0,90	-0,80	-0,98	0,35	0,35
4	2,03	2,08	2,04	-0,15	-0,11
5	0,42	0,32	0,32	1,29	1,30
6	-0,50	-0,62	-0,48	0,97	0,94

Відповідно до розглянутого алгоритму наступним кроком є розрахунок елементів матриці відстаней. Як міра відстані використовується евклідова відстань (див. табл. 7.1). Так, відстань між 1-м і 2-м показниками групи показників ліквідності визначається таким чином:

$$\rho(z_1, z_2) = \sqrt{(0{,}32 + 0{,}05)^2 + (0{,}83 - 0{,}09)^2 + (1{,}49 - 0{,}64)^2} + \\ + \sqrt{(-0{,}25 - 1{,}62)^2 + (-1{,}26 - 0{,}74)^2 + (-1{,}14 - 1{,}55)^2} = 2{,}31.$$

Аналогічно розраховують інші елементи матриці відстаней. Матриця відстаней для групи показників ліквідності має вигляд:

$$R_1 = \begin{pmatrix} 0{,}00 & 2{,}31 & 2{,}37 \\ 2{,}31 & 0{,}00 & 1{,}77 \\ 2{,}37 & 1{,}77 & 0{,}00 \end{pmatrix}.$$

Матриці відстаней для груп показників фінансової стійкості та ділової активності виглядають так:

$$R_2 = \begin{pmatrix} 0{,}00 & 3{,}89 & 4{,}39 \\ 3{,}89 & 0{,}00 & 3{,}47 \\ 4{,}39 & 3{,}47 & 0{,}00 \end{pmatrix},$$

$$R_3 = \begin{pmatrix} 0{,}00 & 0{,}2 & 0{,}17 \\ 0{,}2 & 0{,}00 & 0{,}25 \\ 0{,}17 & 0{,}25 & 0{,}00 \end{pmatrix}.$$

Для вибору показників-репрезентантів груп знаходимо суму відстаней кожного показника групи до інших (табл. 7.5 – 7.7).

Таблиця 7.5

Вибір показника-репрезентанта групи показників ліквідності

Умовна позначка показників	x_1	x_2	x_3	ρ_i
x_1	0,00	2,31	2,37	4,69
x_2	**2,31**	**0,00**	**1,77**	**4,08**
x_3	2,37	1,77	0,00	4,14

Таблиця 7.6

Вибір показника-репрезентанта групи показників фінансової стійкості

Умовна позначка показників	x_4	x_5	x_6	ρ_i
x_4	0,00	3,89	4,39	8,28
x_5	**3,89**	**0,00**	**3,47**	**7,36**
x_6	4,39	3,47	0,00	7,85

Таблиця 7.7

Вибір показника-репрезентанта групи показників ділової активності

Умовна позначка показників	x_7	x_8	x_9	ρ_i
x_7	**0,00**	**0,20**	**0,17**	**0,36**
x_8	0,20	0,00	0,25	0,45
x_9	0,17	0,25	0,00	0,42

До складу показників-репрезентантів входять показники з найменшою сумою відстаней $\rho_s = \min \rho_i$.

Таким чином, як показник-репрезентант у групі показників ліквідності був виділений показник абсолютної ліквідності, у групі показників фінансової стійкості – коефіцієнт забезпеченості власними обіговими коштами, у групі показників ділової активності – коефіцієнт оборотності активів. Оскільки група показників рентабельності включає тільки два показники, то для вибору показника-репрезентанта цієї групи знайдемо відстані кожного показника до раніше виділених показників-репрезентантів (табл. 7.8). Як показник-репрезентант вибирається той, у якого сума відстаней від показників-репрезентантів, виділених із груп елементів із числом більше двох, максимальна $\rho_s = \max_i p_i$.

Таблиця 7.8

Вибір показника-репрезентанта групи показників рентабельності

Умовна позначка показників	x_2	x_5	x_7	ρ_i
x_{10}	**4,12**	**3,25**	**3,29**	**10,66**
x_{11}	4,09	3,29	3,25	10,64

Таким чином, до показників-репрезентантів належать: x_2 – коефіцієнт швидкої ліквідності; x_5 – коефіцієнт маневреності власного капіталу; x_7 – коефіцієнт оборотності активів; x_{10} – коефіцієнт рентабельності.

7.4. Оцінювання якості діагностичного простору ознак

Слід зазначити, що розглянуті в розділах 6 і 7 способи побудови системи діагностичних показників мають певні недоліки. Щодо застосування методу "центра ваги" в принципі немає впевненості щодо правильності вибору саме цієї, а не іншої ознаки, оскільки значення показників, серед яких вибираються істотні ознаки, не завжди достатньо відрізняються. Таким чином, роль істотної ознаки однаково добре можуть виконувати кілька ознак. Метод таксономічного показника рівня розвитку вільний від згаданого недоліку. Однак вадою цього методу є суб'єктивна розбивка ознак на стимулятори та дестимулятори. Тому для вибору найбільш доцільного методу скорочення інформаційного простору показників проводиться аналіз одного із критеріїв якості діагностичного простору – відсотка поясненої дисперсії. За допомогою цього критерію визначається, яку частину всього обсягу інформації, що втримується в усіх вихідних показниках, удалося пояснити введеними діагностичними показниками. У табл. 7.9 подані формули розрахунку даного критерію для різних методів скорочення простору ознак.

Таблиця 7.9

Способи розрахунків відсотка поясненої дисперсії для різних методів скорочення інформаційного простору ознак

Методи	Формула розрахунку відсотка поясненої дисперсії	Умовні позначення
Таксономічний показник рівня розвитку	$e = \frac{1}{N}\sum_{n=1}^{N} r^{2_n}$	r_n– коефіцієнт кореляції n- го вихідного показника з таксономічним показником $(n = 1, N)$
Метод "центра ваги"	$e_k = \frac{1}{p_k}\sum_{l=1}^{p_k} r^{2_{lk}}$; $k = 1,2, \dots, K$; $e = \frac{1}{K}\sum_{k=1}^{K} e_k$	k – номер групи показників; p_k – число показників у підмножині k; r_{lk} - коефіцієнт кореляції між діагностичним показником групи k і l-м показником, що входить до складу групи k показників

Серед альтернативних систем діагностичних ознак, які отримані за допомогою різних методів редукції ознак, обирається система, для якої відсоток поясненої дисперсії є найбільшим.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. Дайте визначення поняттю "система діагностичних показників".
2. Які методи належать до методів неповної редукції ознак? У чому їх особливості?
3. Наведіть алгоритм методу "центра ваги".
4. Які процедури використовують для стандартизації ознак?
5. Які міри відстані застосовують для побудови матриці відстаней? У чому їх особливості?
6. У чому особливості побудови матриць вихідних даних за статичними та динамічними даними?
7. Яким чином здійснюють вибір об'єкта-репрезентанта в групах з числом елементів більше двох?
8. Яким чином здійснюють вибір репрезентантів у одноелементних і двохелементних групах?
9. Який критерій застосовується для оцінювання якості простору діагностичних ознак?

Тестові запитання

1. Вибір репрезентантів груп здійснюється на основі:

а) методу куль;

б) методу дендритів;

в) методу "центра ваги".

2. Репрезентант – це:

а) синтетичний показник;

б) ваговий коефіцієнт ознаки;

в) типопредставник групи.

3. В алгоритмі "центра ваги" використовується така метрика:

а) евклідова відстань;

б) відстань Махалонобіса.

4. Методи вибору репрезентантів груп призначені для:

а) упорядкування одиниць сукупності;

б) розбивки безлічі на групи однорідних елементів;

в) виділення діагностичних змінних.

5. Репрезентантом є елемент, який задовольняє умові (для m>2):

а) $d_m = \min d_i$;

б) $d_m = \max d_i$;

в) $d_m = \min \max d_i$.

6. Діагностичні ознаки мають такі властивості:

а) ознаки не корельовані або слабко корельовані між собою;

б) сильно корельовані з ознаками, що не входять у діагностичний набір;

в) усі відповіді правильні.

7. Елементи головної діагоналі матриці відстаней між об'єктами дорівнюють:

а) 1;

б) 0;

в) дисперсіям;

г) коефіцієнтам кореляції.

8. Метод "центра ваги" належить до групи методів:

а) вибору репрезентантів груп;

б) дискримінантного аналізу;

в) факторного аналізу.

9. Евклідова відстань розраховується за формулою:

а) $d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})}$;

б) $d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$;

в) $d_{ij} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$.

10. У результаті якого варіанта нормування середнє кожного показника буде дорівнювати нулю, а дисперсія – одиниці:

а) $x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$;

б) $x'_{ij} = x_{ij} - \bar{x}_j$;

в) $x'_{ij} = \frac{x_{ij}}{\sigma_j}$?

11. Репрезентантом групи з більше ніж двома об'єктами є:

а) об'єкт з мінімальною сумою відстаней від усіх об'єктів;

б) об'єкт з максимальною сумою відстаней від усіх об'єктів;

в) об'єкт з максимальною сумою відстаней від репрезентантів виділених груп.

12. Репрезентантом групи з двох об'єктів є:

а) об'єкт з мінімальною сумою відстаней від усіх об'єктів;

б) об'єкт з максимальною сумою відстаней від репрезентантів виділених груп;

в) об'єкт з максимальною сумою відстаней від усіх об'єктів.

13. Результатом яких методів є індивідуальні діагностичні ознаки, де первинний набір q ознак $y = (y_1, y_2, \dots, y_q)$ замінюється набором s діагностичних ознак $x = (x_1, x_2, \dots, x_s)$, $(s < q)$:

а) неповної редукції;

б) повної редукції;

в) стандартизації;

г) діагностики?

Практичні завдання

Завдання 1. У табл. 7.10 наведені показники ліквідності, ділової активності, фінансової стійкості, рентабельності, майнового стану промислових підприємств. Необхідно вибрати найбільш значущі для оцінювання стійкості фінансового стану цих підприємств показники (показники-репрезентанти) за допомогою методу "центра ваги".

Таблиця 7.10

Вхідні дані

№ підприємства	Показники ліквідності			Показники фінансової стійкості		
	x_1	x_2	x_3	x_4	x_5	x_6
1	2,713	0,984	0,001	0,855	0,024	0,930
2	3,079	1,006	0,002	0,979	0,026	0,926
3	3,553	1,091	0,020	1,174	0,024	0,634
4	2,304	1,244	0,015	0,885	0,002	0,885
5	1,572	0,877	0,002	0,460	0,022	0,910
6	1,659	0,751	0,001	0,410	0,032	0,912
7	2,241	0,787	0,000	0,598	0,035	0,935
8	3,046	0,919	0,000	0,909	0,032	0,909
9	3,855	1,091	0,000	1,250	0,026	0,950
10	4,500	1,260	0,000	1,548	0,020	0,940

Умовні позначення: x_1 – коефіцієнт поточної ліквідності; x_2 – коефіцієнт швидкої ліквідності; x_3 – коефіцієнт абсолютної ліквідності; x_4 – коефіцієнт забезпеченості власними оборотними коштами; x_5 – коефіцієнт маневреності власного капіталу; x_6 – коефіцієнт автономії.

Наступні вхідні показники наведено у табл. 7.11.

Таблиця 7.11

Вхідні дані

№ підприємства	Показники ділової активності			Показники рентабельності	
	x_7	x_8	x_9	x_{10}	x_{11}
1	0,052	0,287	0,998	0,005	0,006
2	0,032	0,189	0,837	0,023	0,025
3	0,024	0,158	0,735	0,058	0,064
4	0,164	0,897	1,775	0,044	0,050
5	0,087	0,447	1,184	0,084	0,093
6	0,043	0,204	0,910	0,075	0,082
7	0,070	0,935	0,021	0,035	0,038
8	0,100	0,909	0,015	0,019	0,021
9	0,053	0,950	0,017	0,075	0,082
10	0,064	0,940	0,022	0,121	0,133

Умовні позначення: x_7 – коефіцієнт оборотності активів; x_8 – коефіцієнт оборотності оборотних коштів; x_9 – коефіцієнт оборотності запасів; x_{10} – коефіцієнт рентабельності активів; x_{11} – коефіцієнт рентабельності власного капіталу.

Завдання 2. Попередній перелік показників ефективності управління капіталом підприємства, сформований за допомогою методів експертного аналізу, наведений у табл. 7.12.

Матриці відстаней для кожної з наведених груп показників, знайдені на основі даних фінансової звітності одного з промислових підприємств, наведені в табл. 7.13 – 7.15.

Необхідно сформувати систему діагностичних ознак ефективності управління капіталом підприємства за допомогою методу "центра ваги".

Таблиця 7.12

Попередній список показників ефективності управління капіталом підприємства

Назва групи	Показники	Умовна позначка
Показники структури джерел формування капіталу	Коефіцієнт забезпеченості власними обіговими коштами	U_1
	Коефіцієнт маневреності власного капіталу	U_2
	Коефіцієнт автономії	U_3
	Показник фінансового левериджу	U_4
	Коефіцієнт фінансової стійкості	U_5
Показники оборотності капіталу	Коефіцієнт оборотності капіталу	A_1
	Коефіцієнт оборотності основних засобів	A_2
	Коефіцієнт оборотності обігових коштів	A_3
	Коефіцієнт оборотності запасів	A_4
	Коефіцієнт оборотності дебіторської заборгованості	A_5
	Коефіцієнт оборотності кредиторської заборгованості	A_6
	Коефіцієнт оборотності власного капіталу	A_7
Показники рентабельності капіталу	Рентабельність активів	R_1
	Рентабельність власного капіталу	R_2
	Рентабельність капіталу	R_3

Таблиця 7.13

Матриця відстаней для показників структури джерел формування капіталу

Умовна позначка показників	U_1	U_2	U_3	U_4	U_5	ρ_i
U_1	0,00	0,38	1,65	0,64	3,17	5,84
U_2	0,38	0,00	1,40	0,42	2,95	5,15
U_3	1,65	1,40	0,00	1,23	1,58	5,85
U_4	0,64	0,42	1,23	0,00	2,76	5,04
U_5	3,17	2,95	1,58	2,76	0,00	10,45

Таблиця 7.14

Матриця відстаней для показників рентабельності капіталу

Умовна позначка показників	R_1	R_2	R_3	ρ_i
R_1	0,00	0,41	0,52	0,93
R_2	0,41	0,00	0,64	1,05
R_3	0,52	0,64	0,00	1,16

Таблиця 7.15

Матриця відстаней для показників оборотності капіталу

Умовна позначка показників	A_1	A_2	A_3	A_4	A_5	A_6	A_7	ρ_i
A_1	0,00	0,51	0,88	0,82	0,61	0,49	0,51	3,82
A_2	0,51	0,00	0,81	0,84	0,62	0,49	0,50	3,77
A_3	0,88	0,81	0,00	0,62	0,77	0,77	0,87	4,72
A_4	0,82	0,84	0,62	0,00	0,63	0,62	0,62	4,16
A_5	0,61	0,62	0,77	0,63	0,00	0,57	0,49	3,68
A_6	0,49	0,49	0,77	0,62	0,57	0,00	0,44	3,38
A_7	0,51	0,50	0,87	0,62	0,49	0,44	0,00	3,43

Завдання 3. У табл. 7.16 наведені показники соціально-економічного розвитку регіонів. Необхідно вибрати регіони-репрезентанти кластеру 1 (регіони з високим рівнем соціально-економічного розвитку) та кластеру 2 (регіони з низьким рівнем соціально-економічного розвитку).

Таблиця 7.16

Вхідні дані

Області	Показники				Кластер, до якого належить регіон
	x_1	x_2	x_3	x_4	
1	2	3	4	5	6
Вінницька	-5	59,6	110,4	268	кластер 2
Волинська	1	59,7	102,9	275	кластер 2
Дніпропетровська	-4,6	62,1	98,5	434	кластер 1
Донецька	-6,5	60,3	93,6	317	кластер 1

Закінчення табл. 7.16

1	2	3	4	5	6
Житомирська	-4,6	59,9	113,4	285	кластер 2
Закарпатська	2,9	58,6	96,9	191	кластер 2
Запорізька	-4,7	61,3	97,1	460	кластер 1
Івано-Франківська	-0,5	55,4	95,3	316	кластер 2
Київська	-3,9	59,5	99,1	199	кластер 1
Кіровоградська	-6	59,5	106,6	199	кластер 2
Луганська	-6,8	59,4	91,1	365	кластер 1
Львівська	-0,9	58,8	101,2	548	кластер 1
Миколаївська	-3,7	60,6	96,5	336	кластер 2
Одеська	-1,9	59,6	100,6	517	кластер 1
Полтавська	-6,8	59,4	94,7	374	кластер 1
Рівненська	2,5	59,6	91,4	356	кластер 2
Сумська	-7,8	60,1	107	378	кластер 2
Тернопільська	-2,7	56,2	99,5	404	кластер 2
Харківська	-4,7	61,5	94,5	765	кластер 1
Херсонська	-3,5	59,6	92,4	276	кластер 2
Хмельницька	-4,6	59,7	97,6	307	кластер 2
Черкаська	-6,6	59,9	95,2	353	кластер 2
Чернівецька	-0,1	58,7	103,7	361	кластер 2
Чернігівська	-9,4	60,6	89,6	235	кластер 2

Умовні позначення: x_1 – коефіцієнт природнього приросту, %; x_2 – рівень зайнятості, %; x_3 – індекси промислової продукції, %; x_4 – кількість студентів у ВНЗ, осіб на 10 000 населення.

Завдання 4. У табл. 7.17 наведені показники демографічної ситуації регіонів і результати групування регіонів за рівнем соціально-економічного розвитку. Необхідно вибрати показники-репрезентанти демографічної ситуації для кластеру 1 (регіони з високим рівнем соціально-економічного розвитку) та кластеру 2 (регіони з низьким рівнем соціально-економічного розвитку). Позначення: x_1 – коефіцієнт народжуваності, осіб на 1 000 нас.; x_2 – коефіцієнт смертності, осіб на 1 000 нас.; x_3 – коефіцієнт природного приросту (скорочення) населення, осіб на 1 000 нас.; x_4 – коефіцієнт міграційного приросту (скорочення) населення, осіб на 1 000 нас.; x_5 – коефіцієнт одруження, осіб на 1 000 нас.

Таблиця 7.17

Вхідні дані

Області	Показники					Кластер, до якого належить регіон
	x_1	x_2	x_3	x_4	x_5	
Вінницька	10,7	15,7	-5	-0,4	6,5	кластер 2
Волинська	14,1	13,1	1	0,3	6,5	кластер 2
Дніпропетровська	10,9	15,5	-4,6	-0,1	6,5	кластер 1
Донецька	9,4	15,9	-6,5	-0,7	6,8	кластер 1
Житомирська	11,9	16,5	-4,6	-0,4	6,4	кластер 2
Закарпатська	14,7	11,8	2,9	-1	6,5	кластер 2
Запорізька	10,2	14,9	-4,7	-0,6	6,3	кластер 1
Івано-Франківська	12,1	12,6	-0,5	0,7	6,5	кластер 2
Київська	11,9	15,8	-3,9	5,9	6,4	кластер 1
Кіровоградська	10,7	16,7	-6	-1,7	7,5	кластер 2
Луганська	9,1	15,9	-6,8	-0,8	6,1	кластер 1
Львівська	11,6	12,5	-0,9	0	6,3	кластер 1
Миколаївська	11,1	14,8	-3,7	-0,7	6,5	кластер 2
Одеська	12,1	14	-1,9	2,4	6,7	кластер 1
Полтавська	9,8	16,6	-6,8	0,3	6,7	кластер 1
Рівненська	15,1	12,6	2,5	-0,8	5,7	кластер 2
Сумська	9,1	16,9	-7,8	-1,3	6,4	кластер 2
Тернопільська	11,0	13,7	-2,7	-1	6,7	кластер 2
Харківська	9,7	14,4	-4,7	2,1	6,4	кластер 1
Херсонська	11,4	14,9	-3,5	-1,8	6,4	кластер 2
Хмельницька	11,1	15,7	-4,6	-0,7	6,6	кластер 2
Черкаська	9,6	16,2	-6,6	-0,5	6,2	кластер 2
Чернівецька	12,6	12,7	-0,1	1,6	7	кластер 2
Чернігівська	9,2	18,6	-9,4	-0,8	5,9	кластер 2

Завдання 5. На підставі даних підприємств (табл. 7.18), кожне з яких характеризується трьома ознаками: x_1 – інвестиції в основний капітал (млн грн), x_2 – чистий прибуток (млн грн), x_3 – коефіцієнт заборгованості (%), виділіть об'єкт-репрезентант і показник-репрезентант вибірки.

Таблиця 7.18

Вхідні дані

№ об'єкта	Ознаки		
	x_1	x_2	x_3
1	15,5	12,2	2,5
2	16,5	8,3	1,8
3	5,4	7,9	4,4
4	10,2	9,2	2,5
5	11,3	8,4	3,2
6	9,4	7,1	2,25
7	20,8	10,1	1,3
8	15,1	6,6	3,8
9	12,4	8,7	5,4
10	8,7	7,5	2,3

Наведіть економічну інтерпретацію отриманих результатів.

Розділ 8. Методи факторного аналізу

8.1. Сутність моделі факторного аналізу, його основні завдання.

8.2. Визначення структури та статистичне дослідження моделі факторного аналізу.

8.3. Метод головних факторів. Оцінювання факторів і задачі класифікації.

8.4. Метод головних компонент.

8.5. Приклад реалізації алгоритму методу головних компонент.

Ключові слова: методи факторного аналізу; метод головних компонент; методи вибору репрезентантів груп; метод "центра ваги"; процедури обертання; розкладення дисперсії; факторне навантаження; оцінка якості факторного відображення.

Література: [14; 15; 23; 34; 36].

8.1. Сутність моделі факторного аналізу, його основні завдання

Факторний аналіз – це багатовимірний метод, який застосовується для вивчення взаємозв'язків між значеннями змінних. Передбачається, що відомі змінні залежать від меншої кількості невідомих змінних і випадкової помилки.

Факторний аналіз – сукупність методів, які на основі реально існуючих зв'язків ознак (або об'єктів) дозволяють виявляти латентні (приховані) узагальнювальні характеристики організаційної структури та механізму розвитку досліджуваних явищ і процесів. Факторний аналіз дозволяє вирішити дві важливі проблеми дослідника: описати об'єкт вимірювання всебічно та водночас компактно. За допомогою факторного аналізу можливе виявлення прихованих змінних факторів, що відповідають за наявність лінійних статистичних зв'язків між спостережуваними змінними.

Дві основні мети факторного аналізу:

1) визначення взаємозв'язків між змінними, (класифікація змінних), тобто "об'єктивна R-класифікація";

2) скорочення числа змінних, необхідних для опису даних.

Завдання факторного аналізу наведено на рис. 8.1.

Рис. 8.1. **Завдання факторного аналізу**

У ході аналізу в один фактор об'єднуються змінні, які сильно корелюють між собою. Унаслідок відбувається перерозподіл дисперсії між компонентами та створюється максимально проста та наочна структура факторів. Після об'єднання корельованість компонент між собою усередині кожного фактора буде вищою, ніж їх корельованість з іншими факторами. Ця процедура також дозволяє виділити латентні змінні, що особливо важливе для аналізу економічних систем. Аналізуючи оцінки, отримані за декількома шкалами, подібними між собою, і тими, які мають високий коефіцієнт кореляції, можна припустити, що існує деяка латентна змінна, за допомогою якої можна пояснити схожість отриманих оцінок. Таку латентну змінну називають *фактором*. Такий фактор впливає на численні показники інших змінних, що приводить до можливості та необхідності виділити його як найбільш загальний, більш високого порядку.

Окреслимо історію розвитку факторного аналізу.

У 1901 р. виходить стаття англійського вченого К. Пірсона "На прямих і плоскостях, найбільш близьких до системи точок у просторі", у якій обговорювалась ідея головних осей.

У 1904 р. опубліковано фундаментальну статтю Ч. Спірмена "Загальний інтелект, об'єктивно визначений та виміряний";

40 – 50-ті роки – глибокі розробки американських статистиків і математиків: Л. Гуттмана, Г. Хотеллінга, Л. Терстоуна, К. Хользингера, С. Рао, англійських: С. Барта, Г. Томсона, Д. Лоулі, А. Максвелла та ін;

у 60-ті роки відбувається розвиток методів факторного аналізу рішеннями Г. Кайзера, Р. Йорескога та ін;

у 1960 р. виходить праця американського вченого Г. Хармана "Сучасний факторний аналіз".

Класифікація методів факторного аналізу подана на рис. 8.2.

Факторний аналіз розподіляють на такі види:

розвідувальний – він здійснюється під час дослідження прихованої структури фактора без припущення про кількість факторів і їх навантажень;

конфірматорний (підтверджувальний), призначений для перевірки гіпотез про кількість факторів і їх навантаженнях.

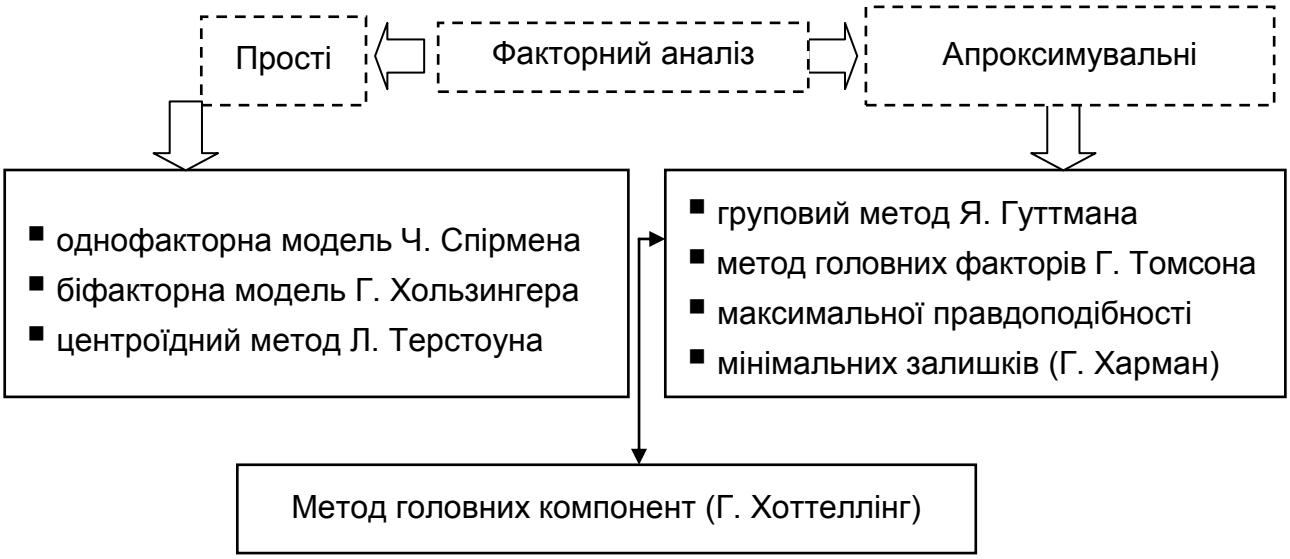


Рис. 8.2. **Класифікація методів факторного аналізу**

Практичне виконання факторного аналізу починається з перевірки його умов. **До обов'язкових умов факторного аналізу належать такі:** усі ознаки мають бути кількісними; кількість спостережень має бути не менше ніж удвічі більше числа змінних; вибірка має бути однорідна; вхідні змінні мають бути розподілені симетрично; факторний аналіз здійснюється за корельованими змінними.

8.2. Визначення структури та статистичне дослідження моделі факторного аналізу

Постановка завдання факторного аналізу. Нехай задана система змінних або ознак, значення яких відомі для кожного з N об'єктів. Уявімо вихідну інформацію у вигляді матриці розмірності $(n \times N)$. Передбачається, що кожен елемент матриці є результатом впливу деякого числа m гіпотетичних чинників і одного характерного. Для побудови моделі факторного аналізу слід провести стандартизацію вихідних даних:

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j},\ s_j = \sqrt{\frac{\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)^2}{n-1}}, \qquad (8.1)$$

де y_{ij} – стандартизоване значення j-ї змінної на і-му об'єкті (безрозмірна величина);

s_j – середньоквадратичне відхилення j-ї ознаки (виправлене).

Основна модель факторного аналізу:

$$y_{ij} = a_{j1}f_{1i} + a_{j2}f_{2i} + \cdots + a_{jm}f_{mi} + d_j v_{ji}, j = \overline{1,n}; r = \overline{1,m}; i = \overline{1,N}, \quad (8.2)$$

де y_{ij} – нормоване значення j-го показника (змінної) і-го об'єкта дослідження;

f_{mi} – значення m-го загального фактора на і-му об'єкті дослідження;

v_{ji} – значення j-го характерного фактора на і-му об'єкті дослідження;

a_{jm} – ваговий коефіцієнт j-й змінної на m-му загальному факторі або навантаження j-й змінної на r-му загальному факторі;

d_j – навантаження або ваговий коефіцієнт j-ї змінної на j-му характерному факторі.

Матрична форма моделі факторного аналізу має такий вигляд:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & \dots & 0 \end{pmatrix} +$$

$$+D = \begin{pmatrix} 0 & 0 & \dots & 0 & d_1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & d_n \end{pmatrix} = \quad (8.3)$$

$$= M = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & d_1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2m} & 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & \dots & d_n \end{pmatrix}$$

Після цих уточнень модель можна записати в матричній формі: $Y = AF$. Або $Y = MF^*$, з урахуванням повної факторної матриці М:

$$F = \begin{pmatrix} f_{11} & \dots & f_{1N} \\ \dots & \dots & \dots \\ f_{m1} & \dots & f_{mN} \end{pmatrix} + V = \begin{pmatrix} v_{11} & \dots & v_{1N} \\ \dots & \dots & \dots \\ v_{n1} & \dots & v_{nN} \end{pmatrix} = F^* = \begin{pmatrix} f_{11} & \dots & f_{1N} \\ \dots & \dots & \dots \\ f_{m1} & \dots & f_{mN} \\ v_{11} & \dots & v_{mN} \\ \dots & \dots & \dots \\ v_{m1} & \dots & v_{mN} \end{pmatrix} \quad (8.4)$$

Дисперсія ознаки y_j в факторному аналізі може бути подана як:

$$S_j^2 = h_j^2 + d_j^2 = 1,$$

де S_j^2 – дисперсія ознаки yj;

h_j^2 – внесок у дисперсію ознаки всіх m загальних факторів;

d_j^2 – внесок у дисперсію ознаки y_j характерного фактора v_j.

Дисперсія характерного фактора може бути подана як:

$$d_j^2 = b_j^2 + c_j^2 = 1,$$

де c_j^2 – компонента дисперсії, пов'язана зі специфікою параметра;

$d_j^2{}_j$ – компоненти, пов'язаної з помилками вимірювань.

Компоненти дисперсії в факторному аналізі розраховують за формулами, поданими в табл. 8.1.

Таблиця 8.1

Компоненти дисперсії

Частка дисперсії	Позначення	Формула
Повна дисперсія	S_j^2	$h_j^2 + b_j^2 + c_j^2 = h_j^2 + d_j^2 = 1$
Надійність	r_j^2	$h_j^2 + b_j^2 = 1 - c_j^2$
Спільність	h_j^2	$1 - d_j^2$
Характерність	d_j^2	$b_j^2 + c_j^2 = 1 - h_j^2$
Специфічність	b_j^2	$d_j^2 - c_j^2$
Дисперсія помилки	c_j^2	$1 - r_j^2$

Фундаментальна теорема факторного аналізу. У виразі $Y = MF^*$ невідомі дві матриці з трьох: M і F^*. Вихідні дані матриці Y дозволяють отримати матрицю R. Це матриця коефіцієнтів парної кореляції, або кореляційна матриця. Також для відтворення матриці парних кореляцій можна використовувати матрицю M:

$$R = \mathrm{MM}^{\mathrm{T}} = AA^T + DD^T;$$
$$R_h = AA^T, DD^T = D^2. \quad (8.5)$$

Таким чином, матриця коефіцієнтів парної кореляції, отримана з вихідних показників, може бути відтворена за допомогою матриці M:

$$R = R_h + D^2. \tag{8.6}$$

Подамо елемент матриці R у розгорнутому вигляді:

$$R = \frac{YY^T}{N}. \tag{8.7}$$

У матричній формі цей вираз можна записати так:

$$r_{jk} = 1/N(y_{j1}y_{k1} + y_{j2}y_{k2} + \cdots + y_{jN}y_{kN}). \tag{8.8}$$

Скориставшись формулою $Y = AF$, перетворимо R в редуковану матрицю:

$$R_h = \frac{1}{N}AF(AF)^T = \frac{1}{N}AFF^TA^T = A\frac{1}{N}FF^TA^T. \tag{8.9}$$

Позначимо $\frac{1}{N}FF^T = C$, відповідно: $R_h = ACA^T$. Цей вираз називають теоремою факторного аналізу. Якщо загальні фактори не корельовані між собою, то C буде одиничною матрицею, отже $R_h = AA^T$.

Методи обчислення спільності:

метод квадрата коефіцієнта множинної кореляції;

метод найбільшого коефіцієнта кореляції за рядком;

метод оцінки середнього коефіцієнта кореляції за рядком;

метод тріад для оцінки h_j.

$$h_j^2 = \frac{r_{jk}r_{jl}}{r_{kl}}, \tag{8.10}$$

де r_{jk} – коефіцієнт кореляції в рядку j, що має найбільше значення стохастичною зв'язку між ознакою Y_j і змінної Y_k;

r_{jl} – коефіцієнт кореляції, який має найбільше значення, що характеризує зв'язок ознаки Y_j с Y_l.

За методом першого центроїдного фактора обчислення h_j проводять таким чином:

$$\hat{h}_j^2 = \frac{\left(\sum_{k=1}^{n} r_{jk}\right)^2}{\sum_{k=1}^{n}\sum_{l=1}^{n} r_{kl}}. \quad (8.11)$$

Таким чином за обчисленими спільностями визначають головні фактори. Через значення повної дисперсії та спільності можна обчислити внесок остаточної частки дисперсії.

8.3. Метод головних факторів. Оцінювання факторів і задачі класифікації

Алгоритм методу головних факторів

1. Розрахунок матриці парних коефіцієнтів кореляції з одиницями на головній діагоналі.

2. Визначення спільності та знаходження матриці R_h, зі спільностями на головній діагоналі.

3. Визначення першого загального фактора за умови, щоб його внесок у дисперсію процесу v1 був максимальним:

$$v_1 = \bar{a}_1 \bar{a}_1 = \sum_{j=1}^{n} a_{j1}^2 = max. \quad (8.12)$$

Умови, за яких має бути забезпечено максимум, описують формулою:

$$r_{jk} = \sum_{r=1}^{m} a_{jr} a_{jk}, (j,k = \overline{1,n}).$$

$$\begin{cases} (h_1^2 - \lambda)a_{11} + r_{12}a_{21} + \cdots + r_{1n}a_{n1} = 0 \\ r_{21}a_{11} + (h_2^2 - \lambda)a_{21} + \cdots + r_{2n}a_{n1} = 0 \\ r_{n1}a_{11} + r_{n2}a_{21} + \cdots + (h_n^2 - \lambda)a_{n1} = 0 \end{cases} \quad \begin{vmatrix} (h_1^2 - \lambda) & \ldots & r_{1n} \\ \ldots & \ldots & \ldots \\ r_{n1} & \ldots & (h_n^2 - \lambda) \end{vmatrix} = 0.$$

Визначник матриці коефіцієнтів цієї системи рівнянь Q має дорівнювати нулю.

4. Знаходження найбільшого власного числа λ і його власного вектора визначають перший загальний фактор, що має максимальний внесок в дисперсію. Значення відповідних елементів матриці А розраховують за формулою:

$$a_{j1} = \frac{\alpha_{j1}\sqrt{\lambda_1}}{\sqrt{\alpha_{11} + \cdots + \alpha_{n1}}}, \lambda_1 = \sum_{j=1}^{n} a_{j1}^2, \quad (8.13)$$

де $\alpha_{11}, \ldots \alpha_{n1}$ – значення відповідного власного вектора.

Таким чином, отриманий перший загальний фактор, вагові коефіцієнти якого забезпечили йому максимальний внесок в сумарну спільність.

5. Знаходження другого загального фактора, який забезпечує максимальний внесок у сумарну дисперсію. Тут замість матриці R використовують матрицю залишків, яка дорівнює:

$$R_1 = R_h - R^1 = R_h - a_1 a_1^T. \quad (8.14)$$

6. Знаходження інших загальних факторів.

7. Числові значення факторів розраховуються за формулою:

$$F = \Lambda^{-1} A^T Y, \Lambda = A^T A. \quad (8.15)$$

Оцінка значущості моделі ФА визначається критерієм Бартлетта (перевіряється нульова гіпотеза, що виділеної кількості факторів достатньо для пояснення вибіркових коефіцієнтів кореляції). Для статистичної перевірки гіпотез знаходять розрахункове значення

$$\chi^2 = N ln \frac{|AA^T|}{|R|}, \quad (8.16)$$

де $|AA^T|$ – визначник відтвореної матриці кореляцій;

$|R|$ – визначник вихідної кореляційної матриці;

N – кількість об'єктів дослідження.

Якщо обчислене значення більше табличного з вибраним рівнем значущості ($\chi^2 > \chi_T^2$) і числі ступенів свободи, що дорівнює:

$$v = \frac{1}{2}[(n-m)^2 - n - m], \tag{8.17}$$

де n – кількість змінних;

m – кількість виділених загальних факторів;

то нульова гіпотеза відхиляється. Тоді слід додати ще один фактор і повторити процедуру перевірки.

Інтерпретація отриманих факторів. Після того як виділені головні фактори або головні компоненти, їх слід інтерпретувати. Принцип інтерпретації факторів: якщо під час аналізу матриці факторних навантажень установлено, що кожен їх головний фактор має помітно великі навантаження на своїй групі ознак, то інтерпретація факторів визначається виділеними таким чином групами ознак. У разі, коли не вдається пояснити сформовану систему факторів, вдаються до процедури обертання.

Проблема обертання. Подання всіх змінних у просторі загальних факторів або головних компонент у вигляді сукупності векторів називають конфігурацією (рис. 8.3).

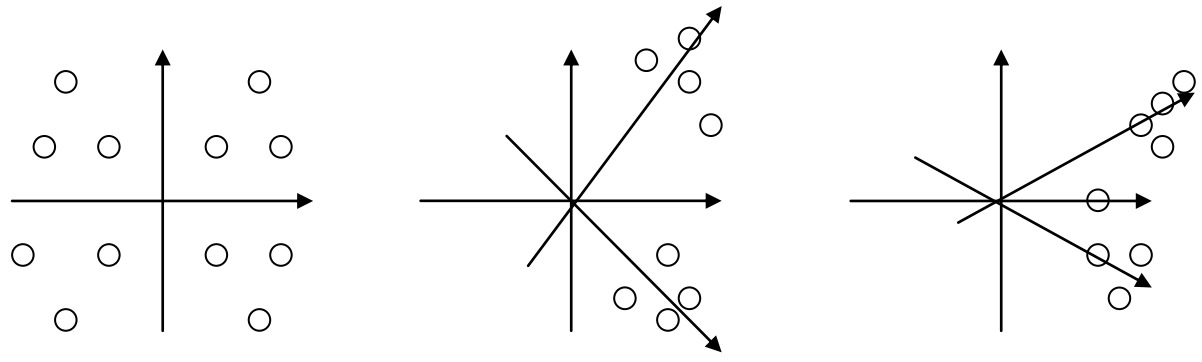


Рис. 8.3. **Конфігурація факторів**

Мета процедур обертання – отримання простої структури, за якої більшість спостережень знаходилась би поблизу координатних осей. Розглянемо приклад методу головних факторів. Задано матрицю стандартизованих вхідних даних (табл. 8.2).

Таблиця 8.2

Вхідні дані

№ п/п	y_1	y_2	y_3	y_4	y_5	y_6
1	0,30	0,04	0,69	0,20	0,23	0,56
2	0,76	0,10	0,57	0,62	0,42	0,52
3	1,36	0,62	1,56	1,27	0,23	2,02
4	0,22	1,43	0,97	0,30	1,92	0,16
5	1,15	0,63	0,57	1,11	0,03	0,38
6	1,04	1,47	0,69	1,28	1,01	0,40

Змінні: y_1 – безробіття; y_2 – імпорт; y_3 – експорт; y_4 – інфляція; y_5 – ставка депозитів; y_6 – податок на прибуток.

Розв'язання

1. Знаходження матриці парних кореляцій.

$$R = \begin{pmatrix} 1{,}00 & 0{,}03 & 0{,}32 & \mathbf{0,96} & -0{,}53 & 0{,}60 \\ 0{,}03 & 1{,}00 & 0{,}17 & 0{,}30 & \mathbf{0,77} & -0{,}24 \\ 0{,}32 & 0{,}17 & 1{,}00 & 0{,}28 & 0{,}07 & \mathbf{0,84} \\ \mathbf{0,96} & 0{,}30 & 0{,}28 & 1{,}00 & -0{,}32 & 0{,}46 \\ -0{,}53 & \mathbf{0,77} & 0{,}07 & -0{,}32 & 1{,}00 & -0{,}44 \\ 0{,}60 & -0{,}24 & \mathbf{0,84} & 0{,}46 & -0{,}44 & 1{,}00 \end{pmatrix}.$$

2. Визначення матриці R_h зі спільностями на головній діагоналі методом найбільшого елемента на рядку.

$$R_h = \begin{pmatrix} \mathbf{0,96} & 0{,}03 & 0{,}32 & 0{,}96 & -0{,}53 & 0{,}60 \\ 0{,}03 & \mathbf{0,77} & 0{,}17 & 0{,}30 & 0{,}77 & -0{,}24 \\ 0{,}32 & 0{,}17 & \mathbf{0,84} & 0{,}28 & 0{,}07 & 0{,}84 \\ 0{,}96 & 0{,}30 & 0{,}28 & \mathbf{0,96} & -0{,}32 & 0{,}46 \\ -0{,}53 & 0{,}77 & 0{,}07 & -0{,}32 & \mathbf{0,77} & -0{,}44 \\ 0{,}60 & -0{,}24 & 0{,}84 & 0{,}46 & -0{,}44 & \mathbf{0,84} \end{pmatrix}.$$

3. Визначення першого загального фактора за умови, що його внесок у сумарну дисперсію максимальний. Для цього знаходимо власні числа і власні вектори матриці Rh.

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
2,869	1,574	1,049	0,039	0,177	0,147

Графічне зображення власних чисел у порядку зменшення має назву "Графік кам'янистого осипу" (рис. 8.4).

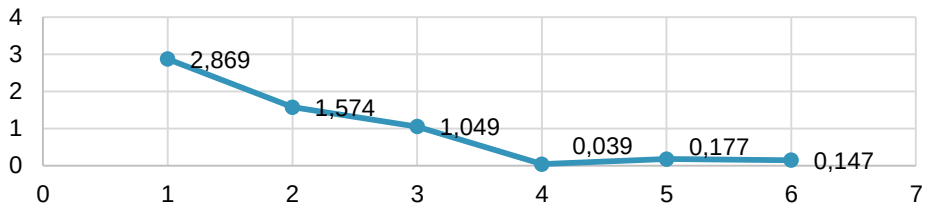


Рис. 8.4. **Графік кам'янистого осипу**

Як видно з графіка, першому загальному фактору відповідає перше власне значення матриці R_h, оскільки воно є максимальним.

4. Вибирається максимальне власне число та власний вектор і розраховуємо значення відповідного першому загальному фактору стовпця матриці вагових коефіцієнтів A, використовуючи формулу 8.13.

$a_1 =$

0,919545
-0,122197
0,577716
0,818082
-0,555222
0,835092

5. Знаходження другого загального фактора. Для цього замість матриці R_h використовується матриця залишків, яка дорівнює:

$$R_1 = R_h - R^1 = R_h - a_1 a_1^T$$

$$\begin{pmatrix} 0{,}11 & 0{,}15 & -0{,}21 & 0{,}20 & -0{,}02 & -0{,}17 \\ 0{,}15 & 0{,}75 & 0{,}24 & 0{,}40 & 0{,}70 & -0{,}14 \\ -0{,}21 & 0{,}24 & 0{,}51 & -0{,}20 & 0{,}39 & 0{,}36 \\ 0{,}20 & 0{,}40 & -0{,}20 & 0{,}29 & 0{,}13 & -0{,}22 \\ -0{,}02 & 0{,}70 & 0{,}39 & 0{,}13 & 0{,}46 & 0{,}03 \\ -0{,}17 & -0{,}14 & 0{,}36 & -0{,}22 & 0{,}03 & 0{,}15 \end{pmatrix}$$

6. Повторюючи процедури 3–5 знаходимо всі необхідні загальні фактори. Матриця вагових коефіцієнтів A:

	F_1	F_2	F_3
x_1	**0,9195**	0,0724	-0,3600
x_2	-0,1222	**0,8934**	-0,1820
x_3	0,5777	0,3817	**0,6494**
x_4	**0,8181**	0,3095	-0,4696
x_5	-0,5552	**0,7268**	0,1400
x_6	**0,8351**	-0,0331	0,4737

Шляхом перевірки за критерієм Бартлетта було встановлено, що виділення трьох загальних факторів достатньою. До того ж висновку можна було прийти, проаналізувавши графік "кам'янистої осипу". Звідки видно, що значення перших трьох власних чисел значно більше за інші. Інтерпретація факторів проводиться шляхом аналізу вагових коефіцієнтів (чим ближче до одиниці, тим більше вплив фактора на показник). Очевидно, що перший фактор має найбільший зв'язок з 1,4 і 6 показниками. Отже, їх доцільно розглядати як групу.

8.4. Метод головних компонент

З числа методів, що дозволяють узагальнювати значення елементарних ознак, метод головних компонент виділяється простою логічною конструкцією; на його прикладі стають зрозумілими загальна ідея і цільові настанови численних методів факторного аналізу. Для виявлення найбільш значущих факторів і їх структури, найбільш виправданим є метод головних компонент. Сутність методу полягає в заміні корельованих компонентів некорельованими факторами. Іншою важливою характеристикою методу є можливість виокремлення найбільш інформативних головних компонент і виключення інших з аналізу, що спрощує інтерпретацію результатів. Переваги методу також у тому, що він є математично обґрунтованим методом факторного аналізу. За твердженням ряду дослідників, метод головних компонент не є методом факторного аналізу, оскільки не розщеплює дисперсію індикаторів на загальну й унікальну.

Метод головних компонент дає можливість за m вхідними ознаками виділити m головних компонент, або узагальнених ознак. Простір головних компонентів є ортогональним. Математична модель головних компонент базується на логічному допущенні, що значення безлічі взаємозалежних ознак породжують деякий загальний результат.

Метод головних компонент використовується для вивчення взаємозв'язків між досліджуваними показниками. За його допомогою можна виявляти приховані показники (фактори), які відповідають за наявність лінійних статистичних зв'язків (кореляцій) між ними. Крім того, визначення більш впливових за умов проведення досліджень факторів серед первинно обраних показників, а також виявлення статистичного зв'язку визначають обґрунтованість висновків щодо ефективності тих чи інших впливів на досліджувану систему.

Алгоритм застосування та реалізації методу головних компонент стосовно дослідження груп кредитних ризиків відображено на рис. 8.5.

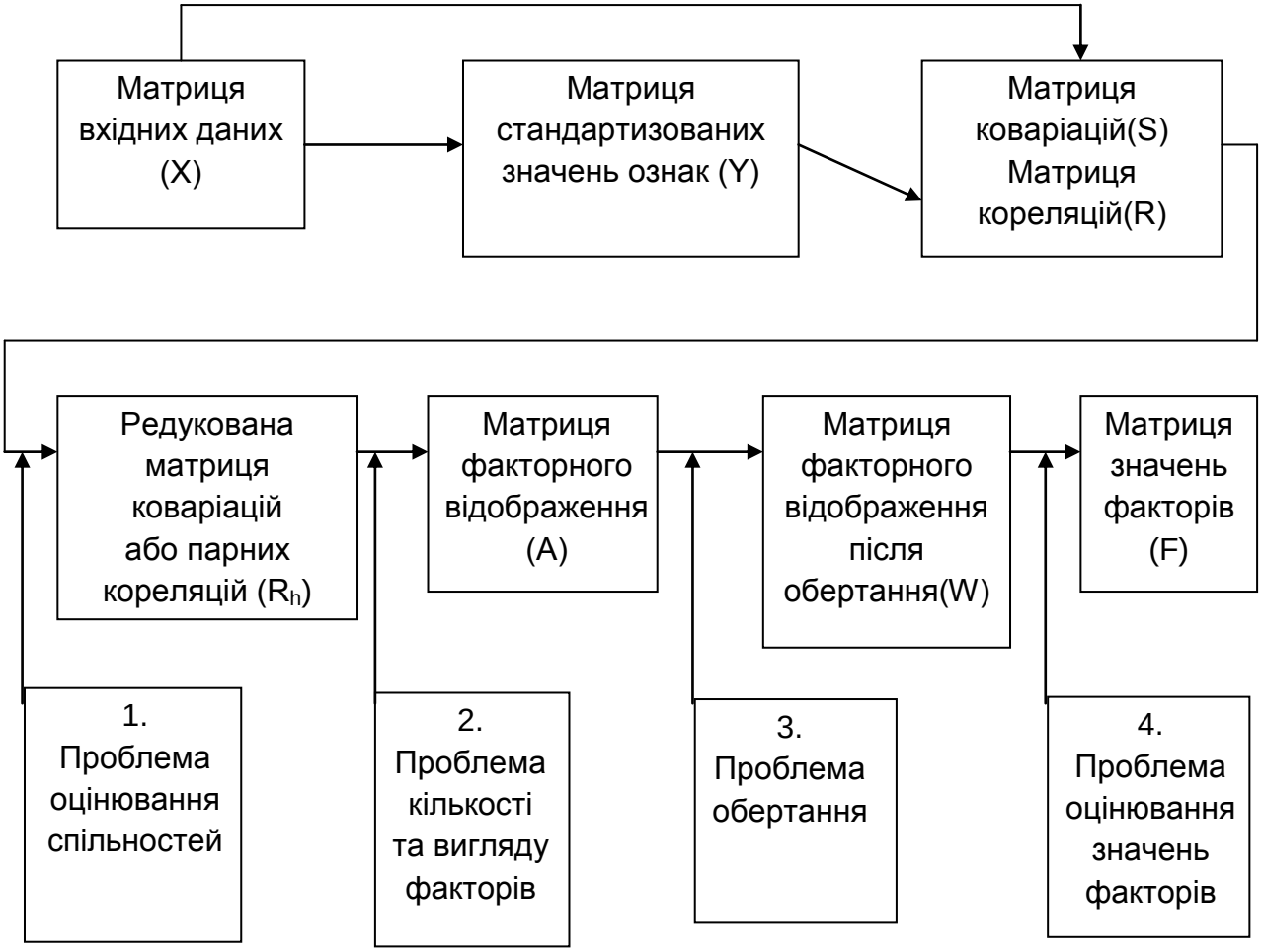


Рис. 8.5. **Загальна алгоритмічна схема реалізації методів факторного аналізу**

Оцінювання за цим методом починається з побудови матриці вихідних даних X і завершується отриманням матриць факторного відображення та значень факторів A і F. З урахуванням прийнятих позначень:

n – кількість спостережень, m – кількість аналітичних ознак X, r – кількість значущих узагальнених ознак (латентних факторів).

Розглянемо обчислювальні процедури методу головних компонент. Розв'язання задачі методом головних компонент зводиться до поетапного перетворення матриці вихідних даних X:

$$X \to Z \to R(S) \to \begin{Bmatrix} \Lambda \\ U \to V \end{Bmatrix} \to A \to F, \tag{8.18}$$

де X – матриця вихідних даних розмірністю $n \times m$; m – кількість елементарних ознак;

Z – матриця стандартизованих значень ознак, елементи матриці обчислюють за формулою: $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$;

R – матриця парних кореляцій $R = \frac{1}{n} \cdot Z'Z$; n – кількість об'єктів спостереження;

A – матриця факторного відображення, її елементи a_{rj} – вагові коефіцієнти.

Λ – діагональна матриця власних (характеристичних) чисел.

Якщо попередня стандартизація даних не проводилася, то на даному кроці отримують матрицю $S = \frac{1}{n} \cdot X'X$, елементи матриці X для розрахунку S будуть центрованими величинами: $x'_{ij} = x_{ij} - \bar{x}_j$;

Λ – діагональна матриця власних (характеристичних) чисел розраховують таким чином:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ 0 & 0 & 0 & \dots & \lambda_m \end{pmatrix}.$$

Значення λ_j знаходять рішенням характеристичного рівняння:

$$|R - \lambda E| = 0, \tag{8.19}$$

де λ_j – характеристики варіації; точніше, показники дисперсії кожної головної компоненти.

Сумарне значення $\sum \lambda_j$ дорівнюють сумі дисперсій елементарних ознак X_j. За умови стандартизації вихідних даних, коли $D(z_{ij}) = 1$, $\sum \lambda_j$ дорівнює кількості елементарних ознак m.

Розв'язання характеристичного рівняння відносно λ, коли число ознак m досить велике та матриця R великої розмірності, виникають труднощі щодо розрахунку визначника $|R|$. Найбільш ефективним є метод, що базується на рекурентних співвідношеннях Фадєєва. Якщо A – деяка симетрична матриця розмірністю $m \times m$, то її визначник знаходять за слідом матриць, похідних з A:

$$A_1 = A \qquad P_1 = t_r A_1 \qquad B_1 = A_1 - P_1 E$$
$$A_2 = AB_1 \qquad P_2 = \frac{1}{2} t_r A_2 \qquad B_2 = A_2 - P_2 E$$
$$\ldots \qquad \ldots \qquad \ldots$$
$$A_{m-1} = AB_{m-2} \qquad P_{m-1} = \frac{1}{m-1} t_r A_{m-1} \qquad B_{m-1} = A_{m-1} - P_{m-1} E$$
$$A_m = AB_{m-1} \qquad P_m = \frac{1}{m} t_r A_m \qquad B_m = A_m - P_m E$$

На заключному етапі розрахунків P_m і є визначник матриці $A(P_m = |A|)$. Для перевірки обчислень може використовуватися умова: $B_m = 0$.

Після обчислень рекурентних співвідношень записується характеристичний багаточлен:

$$P_m(\lambda) = \lambda^m - P_1 \lambda^{m-1} - P_2 \lambda^{m-2} - \cdots - P_m. \qquad (8.20)$$

Значення λ знаходять після того, як характеристичний багаточлен дорівнюють нулю, отримують характеристичне рівняння та розв'язують його щодо характеристичних коренів λ_j. Спочатку A має розмірність $m \times m$ – за кількістю елементарних ознак X_j; потім в аналізі залишається r найбільш значущих компонент, $r \leq m$. Обчислюють матрицю A за відомими даними матриці власних чисел Λ і нормованими власними векторами V за формулою $A = V\Lambda^{1/2}$.

F – матриця значень головних компонент розмірністю $r \times n$ розраховується таким чином:

$$F = A^{-1}Z' \text{ або } F = \Lambda^{-1}A'Z', \text{ або } F = \Lambda^{-1/2}V'Z', \qquad (8.21)$$

де V – матриця нормованих власних (характеристичних) векторів.

Матриця F у загальному виді записується так:

		Об'єкти			
		n_1	n_2	...	n_n
Головні компоненти	F_1	f_{11}	f_{12}	...	f_{1n}
	F_2	f_{21}	f_{22}	...	f_{2n}
	...	...	...	...	...
	F_r	f_{r1}	f_{r2}	...	f_{rn}

Кількість векторів V_j спочатку дорівнює m, тобто $j = \overline{1,m}$. Отримують V_j перетворенням ненормованих власних векторів U:

$$V_j = \frac{U_j}{|U_j|}, \tag{8.22}$$

де $|U_j|$ – норма вектора U , тобто $|U_j| = \left(u_{1j}^2 + u_{2j}^2 + \cdots + u_{mj}^2\right)^{1/2}$.

У свою чергу, власні вектори U_j знаходять з матричного рівняння: $(R - \lambda E)U = 0$. Реально це означає розв'язання m систем лінійних рівнянь для кожного λ_j за $j = \overline{1,m}$. У загальному випадку система рівнянь має вигляд:

$$\begin{gathered}
(1-\lambda_j)u_{1j} + r_{12}u_{2j} + r_{13}u_{3j} + \cdots + r_{1m}u_{mj} = 0; \\
r_{21}u_{1j} + (1-\lambda_j)u_{2j} + r_{23}u_{3j} + \cdots + r_{2m}u_{mj} = 0; \\
r_{31}u_{1j} + r_{32}u_{2j} + (1-\lambda_j)u_{3j} + \cdots + r_{3m}u_{mj} = 0. \\
\cdots \\
r_{m1}u_{1j} + r_{m2}a_{2j} + r_{m3}u_{3j} + \cdots + (1-\lambda_j)u_{mj} = 0.
\end{gathered}$$

Приведена система поєднує однорідні лінійні рівняння; оскільки кількість її рівнянь дорівнює кількості невідомих u_{mj} , вона має нескінченну множину рішень. Конкретні значення власних векторів можна знайти, задаючи принаймні величину компонента кожного вектора, і щоб не ускладнювати розрахунків; вона дорівнює одиниці.

8.5. Приклад реалізації алгоритму методу головних компонент

Розглянемо реалізацію алгоритму факторного аналізу на прикладах.

Приклад 1. Сукупність з чотирьох промислових підприємств оцінена за трьома характерними ознаками: X_1 – середньорічний виробіток, X_2 – рівень рентабельності; X_3 – рівень фондовіддачі. У результаті попередніх аналітичних розрахунків за вихідними даними отримана матриця парних кореляцій:

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Розв'язання

Використовуючи алгоритм методу головних компонент, знайдемо власні числа та власні вектори матриці R і побудуємо матриці з аналітичними результатами *(A* и *F):*

1. За рекурентним співвідношенням Фадєєва обчислимо визначник матриці парних кореляцій $|R|$.

Перший крок: $R = A$ і $A = A_1$, тоді $P_1 = t_r A_1 = 1 + 1 + 1 = 3$,

$$B = A_1 - P_1 E = \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix}.$$

Другий крок:

$$A_2 = AB_1 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \cdot \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix} =$$
$$= \begin{pmatrix} -1{,}638 & -0{,}513 & 0{,}101 \\ -0{,}513 & -1{,}469 & -0{,}350 \\ 0{,}101 & -0{,}350 & -1{,}783 \end{pmatrix}.$$

$$P_2 = \frac{1}{2} t_r A_2 = \frac{1}{2} \cdot (-1{,}638) + (-1{,}469) + (-1{,}783) = -2{,}445,$$

$$B_2 = A_2 - P_2 E = \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix}.$$

Третій крок:

$$A_3 = AB_2 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}524 & 0 & 0 \\ 0 & 0{,}524 & 0 \\ 0 & 0 & 0{,}524 \end{pmatrix}.$$

$$P_3 = \frac{1}{3}(3 \cdot 0{,}524) = 0{,}524, B_3 = A_3 - P_3 E = 0.$$

Отже: $|R| = 0{,}524$ і $B_3 = 0$.

2. Побудуємо характеристичне рівняння:

$$\lambda^3 - 3\lambda^2 + 2{,}445\lambda - 0{,}524 = 0,$$

$$\lambda_1 = 1{,}798;\ \lambda_2 = 0{,}875;\ \lambda_3 = 0{,}327.$$

Таким чином, наші вхідні елементарні ознаки X_1, X_2, X_3 можуть бути узагальнені значеннями трьох головних компонент, причому перша головна компонента F_1 пояснює приблизно 60 % усієї варіації X_j (1,798/3 = 0,599); друга головна компонента F_2 пояснює 29,2 % – загальної дисперсії (0,875/3 = 0,292); третя головна компонента F_3 охоплює залишкову, ще не пояснену варіацію вхідних ознак – 10,9 % (0,327/3 = 0,109). Усі головні компоненти пояснюють варіацію X_1, X_2, X_3 цілком на 100 % (59,9 + 29,2 + 10,9).

3. Власні вектори матриці парних кореляцій R знайдемо розв'язанням трьох систем лінійних рівнянь, відповідно для: $\lambda_1 = 1{,}798$; $\lambda_2 = 0{,}875$; $\lambda_3 = 0{,}327$.

Перша система рівнянь, для $\lambda_1 = 1{,}798$:

$$\left.\begin{matrix}(1-1{,}798)u_{11}+0{,}581u_{21}+0{,}154u_{31}=0\\ 0{,}581u_{11}+(1-1{,}798)u_{21}+0{,}439u_{31}=0\\ 0{,}154u_{11}+0{,}439u_{21}+(1-1{,}798)u_{31}=0\end{matrix}\right| \begin{matrix}u_{11}=1{,}262\\ u_{21}=1{,}469.\\ u_{31}=1{,}000\end{matrix}$$

Друга система рівнянь, для $\lambda_2 = 0{,}875$:

$$\left.\begin{matrix}(1-0{,}875)u_{12}+0{,}581u_{22}+0{,}154u_{32}=0\\ 0{,}581u_{12}+(1-1{,}875)u_{22}+0{,}439u_{32}=0\\ 0{,}154u_{12}+0{,}439u_{22}+(1-1{,}875)u_{32}=0\end{matrix}\right| \begin{matrix}u_{11}=-0{,}144\\ u_{21}=-0{,}234.\\ u_{31}=1{,}000\end{matrix}$$

Третя система рівнянь, для $\lambda_3 = 0{,}327$:

$$\left.\begin{matrix}(1-0{,}327)u_{13}+0{,}581u_{23}+0{,}154u_{33}=0\\ 0{,}581u_{13}+(1-0{,}327)u_{23}+0{,}439u_{33}=0\\ 0{,}154u_{13}+0{,}439u_{23}+(1-0{,}327)u_{33}=0\end{matrix}\right| \begin{matrix}u_{11}=1{,}307\\ u_{21}=-1{,}779.\\ u_{31}=1{,}000\end{matrix}$$

Матриця власних векторів приймає вид:

$$U=\begin{pmatrix}1{,}262 & -0{,}144 & 1{,}307\\ 1{,}469 & -0{,}234 & -1{,}779\\ 1{,}000 & 1{,}000 & 1{,}000\end{pmatrix}.$$

Пронормуємо вектори U_j, тобто знайдемо $V_j=\frac{U}{|U_j|}$, отримаємо матрицю нормованих значень власних векторів:

$$V=\begin{pmatrix}0{,}579 & -0{,}139 & 0{,}539\\ 0{,}674 & -0{,}225 & -0{,}734\\ 0{,}459 & 0{,}964 & 0{,}413\end{pmatrix}.$$

4. Матрицю факторного відображення *(A)* отримаємо з матричного рівняння $A=V\Lambda^{1/2}$.

$$A=\begin{pmatrix}0{,}579 & -0{,}139 & 0{,}539\\ 0{,}674 & -0{,}225 & -0{,}734\\ 0{,}459 & 0{,}964 & 0{,}413\end{pmatrix}\cdot\begin{pmatrix}\sqrt{1{,}798} & & \\ & \sqrt{0{,}875} & \\ & & \sqrt{0{,}327}\end{pmatrix}=$$

$$= \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{ccc} F_1 & F_2 & F_3 \\ 0{,}776 & -0{,}130 & 0{,}308 \\ 0{,}904 & -0{,}210 & -0{,}420 \\ 0{,}616 & 0{,}920 & 0{,}236 \end{array}.$$

Матриця *A* містить частинні коефіцієнти кореляції, що представляють зв'язки вихідних ознак X_j і головних компонент F_r. Відповідно, всі елементи a_{ij} можуть змінюватися у межах від -1 до +1.

З рівності $A'A = \Lambda$ випливає умова $\sum_j a_{jr}^2 = \lambda_r$.

Перевіримо, дану умову для матриці *A:*

$$\sum_j a_{j1}^2 = 0{,}776^2 + 0{,}904^2 + 0{,}616^2 = 1{,}798;$$
$$\sum_j a_{j2}^2 = 0{,}875;$$
$$\sum_j a_{j3}^2 = 0{,}327.$$

5. Тепер запишемо системи лінійних рівнянь залежності елементарних ознак Z_j і головних компонент, або узагальнених ознак F_r (максимально $r = j = 3$):

$$Z_1 = 0{,}776 \cdot F_1 - 0{,}130 \cdot F_2 + 0{,}308 \cdot F_3,$$
$$Z_2 = 0{,}904 \cdot F_1 - 0{,}210 \cdot F_2 - 0{,}420 \cdot F_3,$$
$$Z_3 = 0{,}616 \cdot F_1 + 0{,}902 \cdot F_2 + 0{,}236 \cdot F_3.$$

6. На завершальному кроці алгоритму обчислимо значення головних компонент для всіх об'єктів, що спостерігаються, і розрахуємо:

$$F = A^{-1}Z'.$$

Матриця Z відома з умови задачі:

$$F = \begin{pmatrix} 0{,}542 & 0{,}507 & 0{,}196 \\ -0{,}776 & -0{,}010 & 0{,}994 \\ 1{,}554 & -1{,}283 & -0{,}075 \end{pmatrix} \cdot \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix} =$$

$$= \begin{pmatrix} -0{,}233 & -1{,}533 & 1{,}143 & 0{,}623 \\ 0{,}823 & -0{,}372 & -1{,}687 & 1{,}236 \\ -2{,}218 & 0{,}892 & 1{,}205 & 0{,}121 \end{pmatrix}.$$

Більш звичною формою запису $n \times r$ значень головних компонент є транспонована матриця *F.*

Центр розподілу значень головних компонент *F* знаходиться в точці (0, 0, ..., 0). Звідси випливає правило рівності суми елементів кожного стовпця матриці $F^{/}$ нулю. У прикладі це правило витримується.

Результати розрахунків необхідні для прийняття рішення про кількість значущих ознак Z_j і головних компонент F_r і визначення назв головним компонентам.

Об'єкти	Головні компоненти F_1	F_2	F
n_1	$-0{,}233$	$0{,}823$	$-0{,}218$
n_2	$-1{,}533$	$-0{,}372$	$0{,}892$
n_3	$1{,}143$	$-1{,}687$	$1{,}205$
n_4	$0{,}623$	$1{,}236$	$1{,}205$
$\sum f_{ri}$	0	0	0

Розглянемо оцінку рівня інформативності та пошук назв для головних компонент.

Задачі розпізнавання головних компонент, визначення для них назв розв'язуються суб'єктивно на основі вагових коефіцієнтів a_{jr} з матриці відображення А.

Для кожної головної компоненти *F* значення a_{jr} умовно розбивається на чотири підмножини з нечіткими межами:

W_1 – підмножина незначущих вагових коефіцієнтів;

W_2 – підмножина значущих вагових коефіцієнтів;

W_3 – підмножина значущих вагових коефіцієнтів, які не формують назву головної компоненти;

$W_2 - W_3$ – підмножина значущих вагових коефіцієнтів, що формують назву головної компоненти.

Додаткове виділення підмножини W_3 пояснюється прагненням до більш простої структури головної компоненти, що завжди легше піддається інтерпретації.

Підтвердження значущості ознак (X_j або Z_j), які формують назви головних компонент, можна отримати розрахунковим шляхом під час визначенні коефіцієнта інформативності:

$$K_u = \frac{\sum_j a_{jr}^2\{W_2 - W_3\}}{\sum_{j=1}^{m} a_{jr}^2}. \qquad (8.23)$$

Набір пояснювальних ознак вважається задовільним, якщо K_u знаходиться у межах 0,75 – 0,95.

Приклад 2. На основі розглянутого алгоритму з семи показників діяльності підприємства розраховані факторні навантаження та визначено два головних фактора. Вихідні дані подані у табл. 8.3.

Таблиця 8.3

Вхідні дані

Коефіцієнти діяльності X_j	Головні компоненти	
	F_1	F_2
x_1 – рівень виробітку на одного робітника;	$a_{11} = 0,9$	$a_{12} = 0,1$
x_2 – рівень фондовіддачі;	$a_{21} = 0,8$	$a_{22} = 0,4$
x_3 – розмір оборотних виробничих засобів;	$a_{31} = 0,1$	$a_{32} = 0,8$
x_4 – розмір витрат на випуск одиниці товарної продукції;	$a_{41} = 0,8$	$a_{42} = 0,3$
x_5 – чисельність ПВП;	$a_{51} = 0,3$	$a_{52} = 0,7$
x_6 – рентабельність продукції;	$a_{61} = 0,7$	$a_{62} = 0,2$
x_7 – рівень енергоозброєності праці	$a_{71} = 0,2$	$a_{72} = 0,6$

Виділимо для першої головної компоненти F_1 підмножини вагових коефіцієнтів на основі простої візуальної оцінки аналітичних результатів:

$$W_1 = \begin{Bmatrix} a_{31} = 0,1 \\ a_{71} = 0,2 \end{Bmatrix};\ W_2 = \begin{Bmatrix} a_{11} = 0,9 & a_{51} = 0,3 \\ a_{21} = 0,8 & a_{61} = 0,7 \\ a_{41} = 0,8 & \end{Bmatrix};$$

$$W_3 = a_{51} = 0{,}3;\ W_2 - W_3 = \begin{Bmatrix} a_{11} = 0{,}9 & a_{41} = 0{,}8 \\ a_{21} = 0{,}8 & a_{61} = 0{,}7 \end{Bmatrix}.$$

Граничні значення для підмножини W_3 будуть: $a_{кр.1}$ = 0,3 і $a_{кр.2}$ = 0,7. У вирішальну підмножину $W_2 – W_3$ увійшли елементарні ознаки: X_1, X_2, X_4, X_6. Усі вони представляють характеристики ефективності виробничої діяльності. Назвемо F_1 ефективність виробництва. Значення коефіцієнта інформативності дає підстави стверджувати, що склад підмножини $W_2 – W_3$ для головної компоненти F_1 досить надійний:

$$K_{u2} = \frac{0{,}81 + 0{,}64 + 0{,}64 + 0{,}49}{0{,}81 + 0{,}64 + 0{,}01 + 0{,}64 + 0{,}09 + 0{,}49 + 0{,}04} = 0{,}948.$$

Тобто значеннями ознак X_1, X_2, X_4, X_6 склад головної компоненти F_1 визначається більше ніж на 94 %.

Для другої головної компоненти F_2:

$$W_1 = \begin{Bmatrix} a_{12} = 0{,}1 \\ a_{62} = 0{,}2 \end{Bmatrix};\ W_2 = \begin{Bmatrix} a_{22} = 0{,}4 & a_{42} = 0{,}3 \\ a_{32} = 0{,}8 & a_{52} = 0{,}7 \\ a_{72} = 0{,}6 & \end{Bmatrix};$$

$$W_3 = \begin{Bmatrix} a_{22} = 0{,}4 \\ a_{42} = 0{,}3 \end{Bmatrix};\ W_2 - W_3 = \begin{Bmatrix} a_{32} = 0{,}8 \\ a_{52} = 0{,}7 \\ a_{72} = 0{,}6 \end{Bmatrix}.$$

Назва головної компоненти F_2 визначається наявністю в її структурі значущих ознак X_3, X_5, X_7, тобто F_2 – це розмір виробничих ресурсів. Коефіцієнт інформативності підтверджує істотний склад цієї головної компоненти:

$$K_{u2} = \frac{0{,}64 + 0{,}49 + 0{,}36}{0{,}01 + 0{,}16 + 0{,}64 + 0{,}09 + 0{,}49 + 0{,}04 + 0{,}36} = 0{,}832.$$

Відбір значущих елементарних ознак для визначення назви головної компоненти визначається насамперед за абсолютною величиною вагового коефіцієнта a_{jr}. Знак коефіцієнта набуває значення за умови логічного пояснення складу та встановлення його несуперечності назві головної компоненти.

Завдання для самостійного опрацювання

Контрольні запитання для самодіагностики

1. У чому полягає сутність факторного аналізу?
2. Назвіть основні етапи факторного аналізу.
3. Які основні групи методів виділяють у факторному аналізі?
4. Які методи розрахунку спільностей ви знаєте?
5. Чим відрізняється редукована матриця від звичайної кореляційної матриці?
6. У чому відмінність методу головних компонент від методів факторного аналізу?
7. Як визначається необхідна та достатня кількість загальних факторів у методі головних факторів?
8. Які види обертання ви знаєте?
9. Яким чином здійснюється економічна інтерпретація отриманих чинників?
10. Який вигляд має лінійна модель методу головних компонент?
11. Назвіть основні поняття факторного аналізу.
12. Наведіть алгоритм методу головних факторів.
13. Як здійснюється розкладання дисперсії у факторному аналізі?
14. Назвіть критерії точності оцінки значень факторів.
15. Як здійснюється оцінювання рівня інформативності головних факторів?
16. Назвіть критерії вибору кількості факторів.

Тестові завдання

1. Обертання квартимакс передбачає:

а) максимальне спрощення в описі матриці факторного відображення;

б) обертання факторних осей, щоб величини факторних навантажень максимізували критерій

2. Для ортогонального перетворення факторного простору використовується:

а) квартимакс;

б) облімакс;

в) квартимін.

3. Значення власних чисел і власних векторів у методі головних компонент визначають для:

а) матриці вихідних даних;

б) нормованої матриці вихідних даних;

в) матриці парних кореляцій;

г) скороченої матриці кореляцій.

4. До простих методів факторного аналізу відносять методи:

а) головних факторів;

б) однофакторної моделі Ч. Спірмена;

в) метод максимальної правдоподібності.

5. Косокутність обертання простору факторів передбачає використання критерію:

а) облімакс;

б) варимакс;

в) квартимакс.

6. Надійність – це:

а) частка дисперсії характерного фактора без урахування помилки;

б) частка дисперсії, не пояснена загальними факторами;

в) частка загальної дисперсії, яка піддається поясненню через загальні фактори.

7. Спільність – це:

а) частка дисперсії, що не пояснене загальними факторами;

б) частка дисперсії, яка піддається поясненню через загальні фактори;

в) частка дисперсії характерного фактора без урахування помилки.

8. Перевірка гіпотези про достатність числа узагальнених ознак (факторів) здійснюється на основі критерію:

а) Уїлкса-χ^2;

б) Бартлетта;

в) Лоулі-χ^2.

9. Перевірка значущості матриці парних кореляції в методі головних чинників здійснюється на основі критерію:

а) Уілкса-χ^2;

б) Бартлетта;

в) Лоулі-χ^2.

10. Процедура обертання застосовується в методі головних компонент:

а) завжди;

б) тільки якщо спочатку виділені головні компоненти не вдається інтерпретувати;

в) якщо наявна незначна дисперсія фактора.

11. Специфічність – це:

а) частка дисперсії, обумовлена варіабельною специфікою ознаки x_j;

б) частка дисперсії, обумовлена недосконалістю вимірювань;

в) частка дисперсії характерного фактора без урахування помилки.

12. Факторні навантаження a_{ik}, отримані в методі головних компонент, є:

а) відстанню показника X_i і головної компоненти F_k;

б) частинними коефіцієнтами кореляції показника X_i і головної компоненти F_k;

в) коефіцієнтами детермінації показника X_i і головної компоненти F_k.

13. Характерність – це:

а) частка дисперсії, що не пояснена загальними факторами;

б) частка дисперсії, яка піддається поясненню через загальні фактори;

в) частка дисперсії характерного фактора без урахування помилки.

14. Коефіцієнт інформативності головних компонент оцінюється:

а) на основі вагових коефіцієнтів ознак;

б) на основі матриці кореляцій;

в) на основі власних чисел.

15. Методи факторного аналізу належить до групи методів:

а) класифікації;

б) зниження розмірності простору ознак;

в) розпізнавання.

16. Скорочена кореляційна матриця – це:

а) матриця зі спільностями на головній діагоналі;

б) матриця, скоригована на кількість факторів і спостережень;

в) відтворена матриця залишків.

17. До методів визначення спільностей відносять:

а) метод головних компонент;

б) метод тріад;

в) метод різниць.

18. Перевірка гіпотези про достатність числа головних компонент здійснюється на основі критерію:

а) χ^2–критерію Уїлкса;

б) χ^2–критерію Бартлетта;

в) критерію Фішера.

19. Матриця парних кореляцій в методі головних компонент розраховується за формулою:

а) $R = \frac{1}{N} \cdot Z'Z$;

б) $R = \frac{1}{n+1} \cdot Z'Z$;

в) $R = \frac{1}{n} \cdot Z$.

20. Матриця факторного відображення обчислюється за формулою:

а) $A = V\Lambda^{1/2}$;

б) $A = V\Lambda$;

в) $A = V\Lambda^2$.

Практичні завдання

Завдання 1. У табл. 8.4 наведено показники фінансово-економічної діяльності десяти промислових підприємств. Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Таблиця 8.4

Вхідні дані

№ підприємства	Питома вага покупних виробів	Коефіцієнт змінності устаткування	Фондоозброєність праці
1	0,40	1,37	6,4
2	0,26	1,49	7,8
3	0,40	1,44	9,76
4	0,50	1,42	7,9
5	0,40	1,35	5,35
6	0,19	1,39	9,9
7	0,25	1,16	4,5
8	0,44	1,27	4,88
9	0,17	1,16	3,46
10	0,39	1,25	3,6

Завдання 2. У табл. 8.5 наведено показники фінансово-економічної діяльності десяти промислових підприємств. Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Таблиця 8.5

Вхідні дані

№ підприємства	Питома вага покупних виробів	Коефіцієнт змінності устаткування	Премії і винагороди на одного працівника
1	0,40	1,37	1,23
2	0,26	1,49	1,04
3	0,40	1,44	1,8
4	0,50	1,42	0,43
5	0,40	1,35	0,88
6	0,19	1,39	0,57
7	0,25	1,16	1,75
8	0,44	1,27	1,70
9	0,17	1,16	0,84
10	0,39	1,25	0,60

Завдання 3. У табл. 8.6 наведено показники фінансово-економічної діяльності десяти промислових підприємств.

Таблиця 8.6

Вхідні дані

№ п/п	Продуктивність праці	Фондовіддача	Рентабельність
1	9,26	1,45	13,26
2	9,38	1,3	10,16
3	12,11	1,37	13,72
4	10,81	1,65	12,85
5	9,35	1,91	10,63
6	9,87	1,68	9,12
7	8,17	1,94	25,83
8	9,12	1,89	23,39
9	5,88	1,94	14,68
10	6,30	2,06	10,05

Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Завдання 4. У табл. 8.7 наведено показники фінансово-економічної діяльності десяти промислових підприємств. Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Таблиця 8.7

Вхідні дані

№ п/п	Премії і винагороди на одного працівника	Фондовіддача	Продуктивність праці
1	2	3	4
1	1,23	1,45	9,26
2	1,04	1,3	9,38
3	1,8	1,37	12,11

Закінчення табл. 8.7

1	2	3	4
4	0,43	1,65	10,81
5	0,88	1,91	9,35
6	0,57	1,68	9,87
7	1,75	1,94	8,17
8	1,70	1,89	9,12
9	0,84	1,94	5,88
10	0,60	2,06	6,30

Завдання 5. У табл. 8.8 наведено показники фінансово-економічної діяльності десяти промислових підприємств: фондовіддачу, середньорічну вартість ОВФ, продуктивність праці.

Таблиця 8.8

Вхідні дані

№ п/п	Фондовіддача	Середньорічна вартість ОВФ	Продуктивність праці
1	1,96	45,44	6,22
2	1,02	41,08	5,49
3	1,85	136,14	6,5
4	0,88	42,39	6,61
5	0,62	37,39	4,32
6	1,09	101,78	7,37
7	1,60	47,55	7,02
8	1,53	32,61	8,25
9	1,40	103,25	8,15
10	2,22	38,95	8,72

Необхідно за допомогою методу головних компонент розрахувати факторні навантаження, визначити значення факторів і виділити найбільш значущі показники.

Розділ 9. Лабораторний практикум

Лабораторна робота 1. Оцінка параметрів розподілу випадкових величин

Мета – закріплення теоретичного та практичного матеріалу з оцінювання параметрів розподілу випадкових величин, придбання навичок роботи в модулі *Basic Statistics / Tables.*

Завдання – необхідно провести аналіз варіаційного ряду для вибіркових даних у модулі *Basic Statistics* / *Tables* ППП *Statistica:*

1) розрахувати статистичні характеристики ряду (середнє, дисперсію, середнє квадратичне відхилення, моду, медіану, розмах варіації, коефіцієнти асиметрії та ексцесу);

2) побудувати гістограму та полігон розподілу випадкової величини, зробити висновки щодо характеру закону розподілу;

3) за допомогою критеріїв Пірсона та Колмогорова – Смірнова перевірити гіпотезу про нормальний закон розподілу;

4) зробити висновки щодо угрупування об'єктів за величиною відповідного показника.

Література: [5 – 9; 14; 41 – 44; 48; 49; 76].

Методичні рекомендації

Для розв'язання та аналізу задач розглядуваного типу в ППП *Statistica* передбачений модуль *Basic Statistics* / *Tables* (*Основні статистики й таблиці*). Розглянемо порядок роботи в даному модулі.

У меню програм слід вибрати програму *Statistica*, після її запуску виберіть у меню пункт *File* / *New для* підготовки власних даних. Перед вами з'явиться діалогове вікно, у якому необхідно вказати кількість змінних *(Number of variables)* і кількість випадків *(Number of Cases).* Після введення натисніть кнопку вікна *OK* (рис. 9.1).

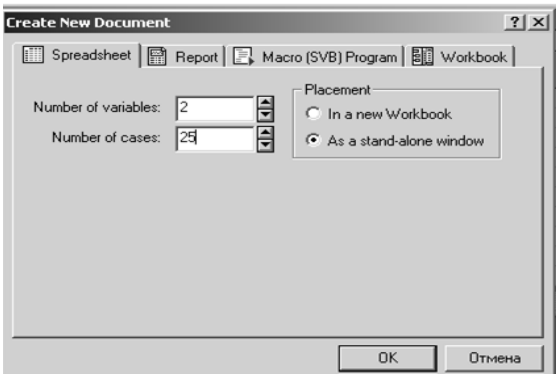


Рис. 9.1. **Визначення кількості змінних і спостережень**

Перед вами з'явиться порожнє поле, що містить таблицю розміром 25*2: 25 спостережень, 2 змінні (рис. 9.2). Кожен елемент даних, тобто значення показника, займає одну комірку поля даних.

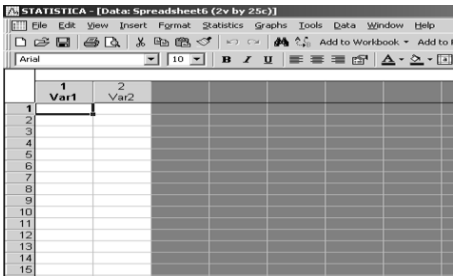


Рис. 9.2. **Поле даних**

Після заповнення всіх комірок поля даних ви отримаєте таблицю, наведену на рис. 9.3.

	1 Bank	2 % dohod
1	Приватбанк	1187477
2	Проминвест	793821
3	Аваль	876148
4	Ощадбанк	389719
5	Укрсоцбанк	459234
6	Укрсиббанк	451074
7	Укрэксим	328131
8	Райффайзенбанк	209010
9	Надра	273945
10	Брокбизнесбанк	167741
11	Укрпромбанк	232158
12	Финансы и кредит	175292
13	Первый укр. международный банк	111185
14	Хрещатик	70674
15	Форум	145468
16	Південний	132243
17	Правексбанк	120243
18	Кредитпромбанк	100261
19	Укргазбанк	104326
20	Кредитбанк	114054
21	Ситибанк	42602
22	Ингбанк Украина	35241
23	Вабанк	71296
24	КредитДнепро	91436
25	Донгорбанк	86384

Рис. 9.3. **Вихідні дані**

Розрахуємо основні статистичні характеристики ряду (середнє, дисперсію, середнє квадратичне відхилення, моду, медіану, розмах, коефіцієнти асиметрії й ексцесу).

Щоб почати обчислювальні процедури, необхідно ввійти в позицію меню *Statistics / Basic Statistics* / *Tables* (рис. 9.4).

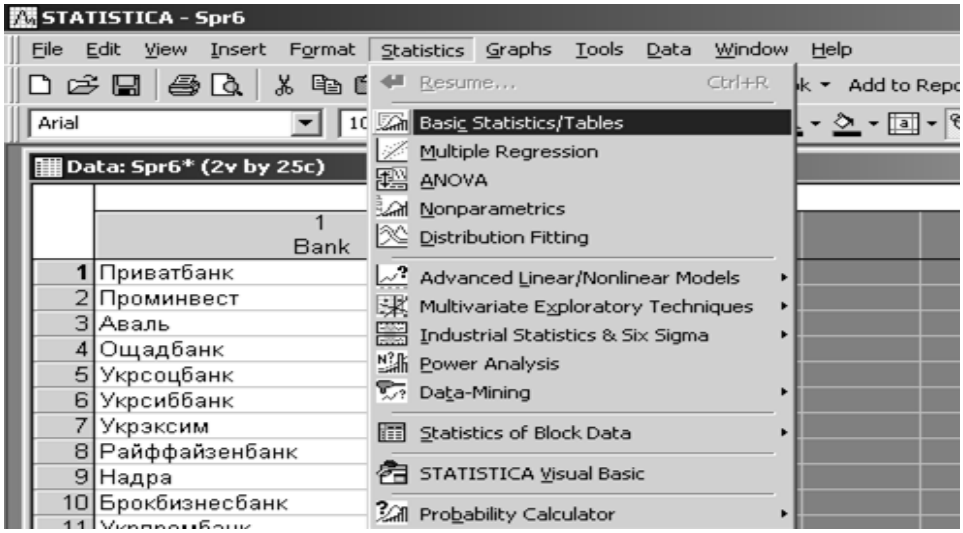


Рис. 9.4. **Вибір модуля**

Після підтвердження вибору модуля перед вами з'явиться діалогове вікно, що дозволяє задати напрям аналізу *Descriptive statistics (Описові статистики),* подане на рис. 9.5.

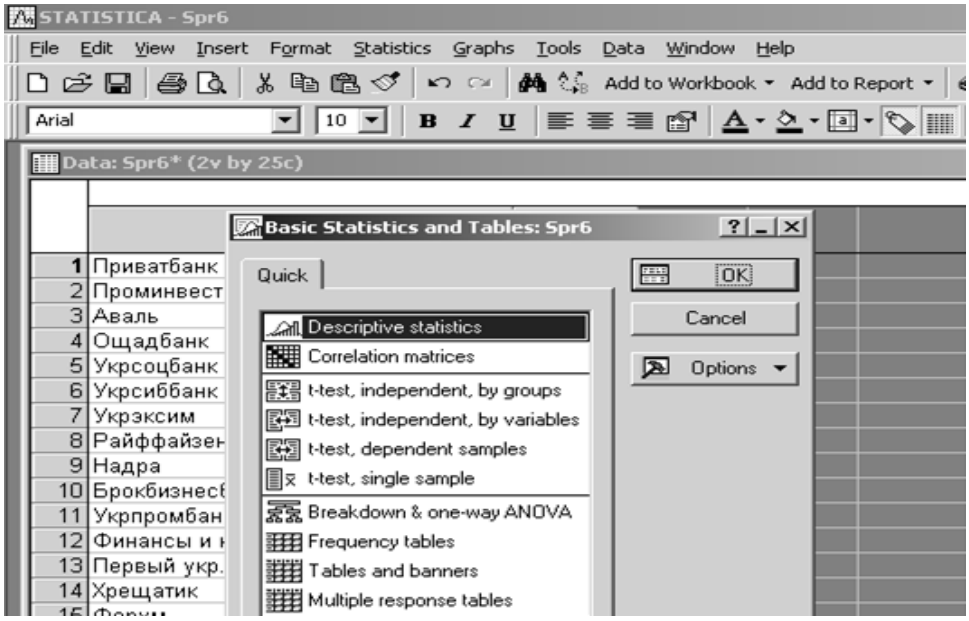


Рис. 9.5. **Вибір напряму аналізу**

Після вибору напряму аналізу з'явиться стартова панель модуля, де необхідно задати вихідні параметри: *Variable* (*Змінні*) та відповідний набір процедур для подальшого аналізу (рис. 9.6).

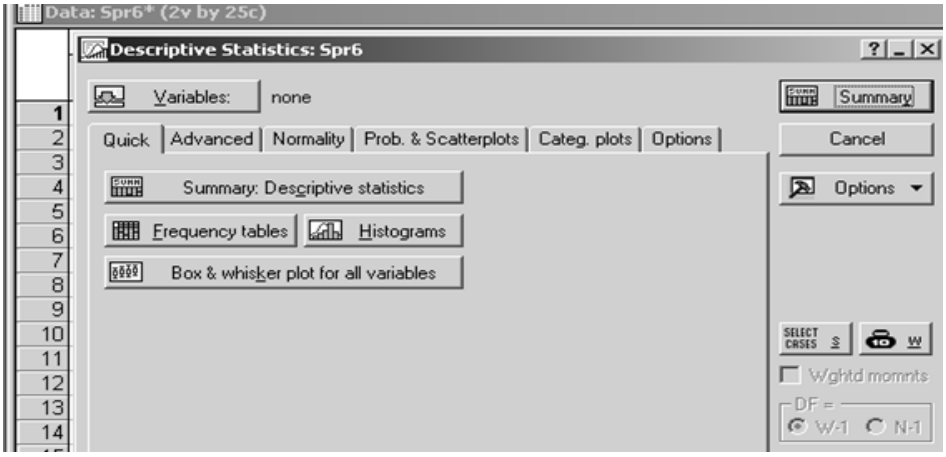


Рис. 9.6. **Стартова панель модуля**

Ініціюйте кнопку *Variable* (Змінні) й у вікні, що з'явилося, укажіть показники, за якими здійснюється аналіз. Після зазначення змінних підтвердьте свій вибір натисканням кнопки *OK.* Далі, ініціювавши вкладиш *Advanced*, необхідно виділити основні статистики для розрахунку (рис. 9.7).

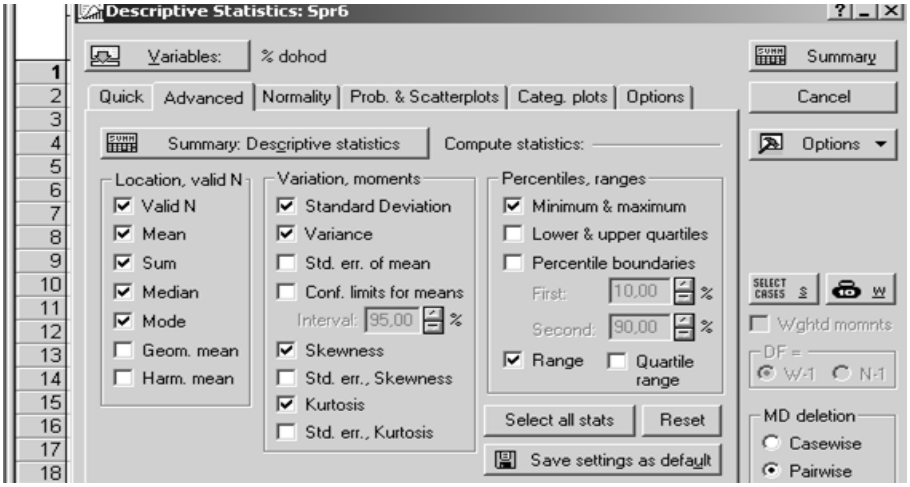


Рис. 9.7. **Вибір описових статистик**

Такими є: *Valid N* (кількість спостережень), *Mean* (середнє), *Sum* (сума значень), *Median* (медіана), *Mode* (мода), *Standard Deviation* (середнє квадратичне відхилення), *Variance* (дисперсія), *Skewness* (коефіцієнт

асиметрії), *Kurtosis* (коефіцієнт ексцесу), *Min & Max* (мінімум і максимум), *Range* (розмах вибірки). Результати розрахунку описових статистик для даної вибірки отримуємо натисканням клавіші *Summary* (рис. 9.8).

Descriptive Statistics (Spr6)

Variable	Valid N	Mean	Median	Sum	Minimum	Maximum	Range
% dohod	25	270766	145468	676916	35241,0	118747	115223

Variance	Std.Dev.	Skewness	Kurtosis
8,370020E+10	289309,9	1,999460	3,715184

Рис. 9.8. **Описові статистики**

Побудуємо гістограму та полігон розподілу випадкової величини; проведемо угрупування вибірки. Для наочності подання досліджуваної сукупності побудуємо полігон розподілу. Для цього необхідно зайти в меню *Graphs* / *2D Graphs* / *Scatterplots* (рис. 9.9), вибрати змінні та задати параметри графіка (рис. 9.10) і побудувати полігон розподілу випадкової величини (рис. 9.11).

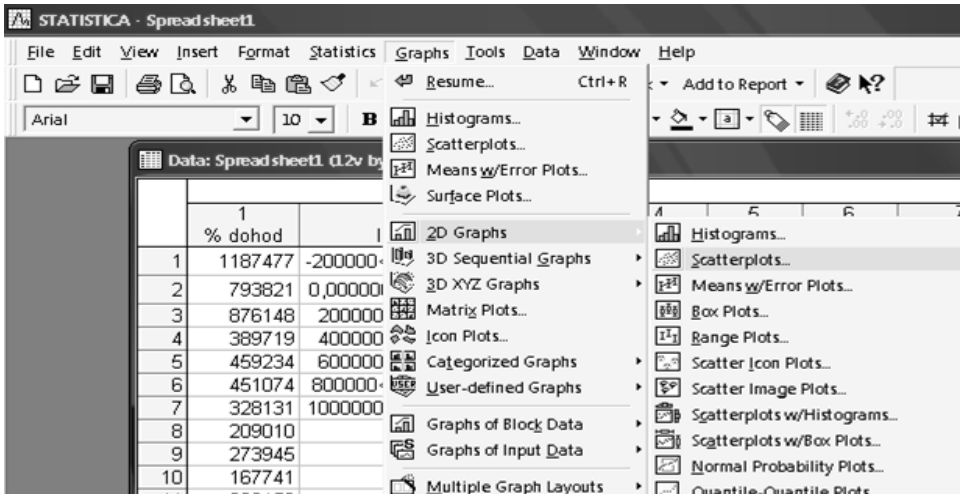


Рис. 9.9. **Вибір типу графіка**

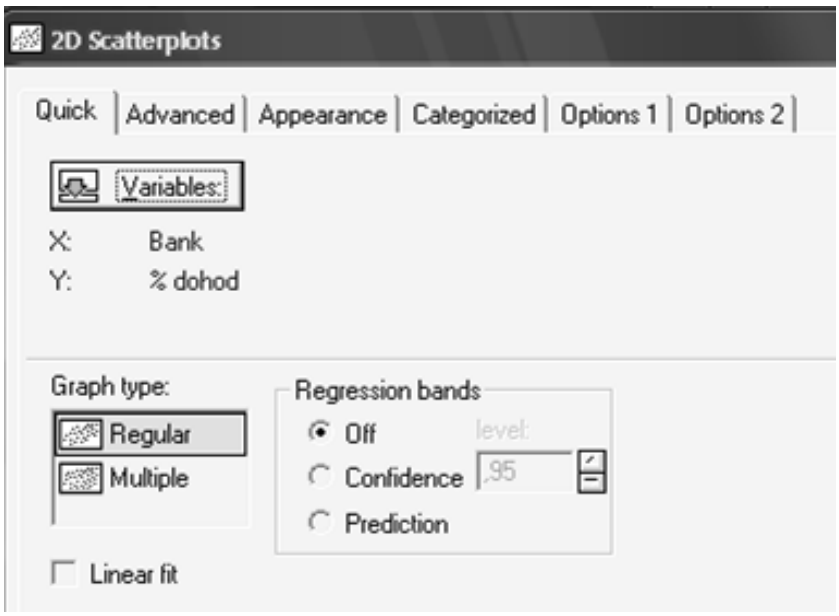


Рис. 9.10. **Вибір параметрів графіка**

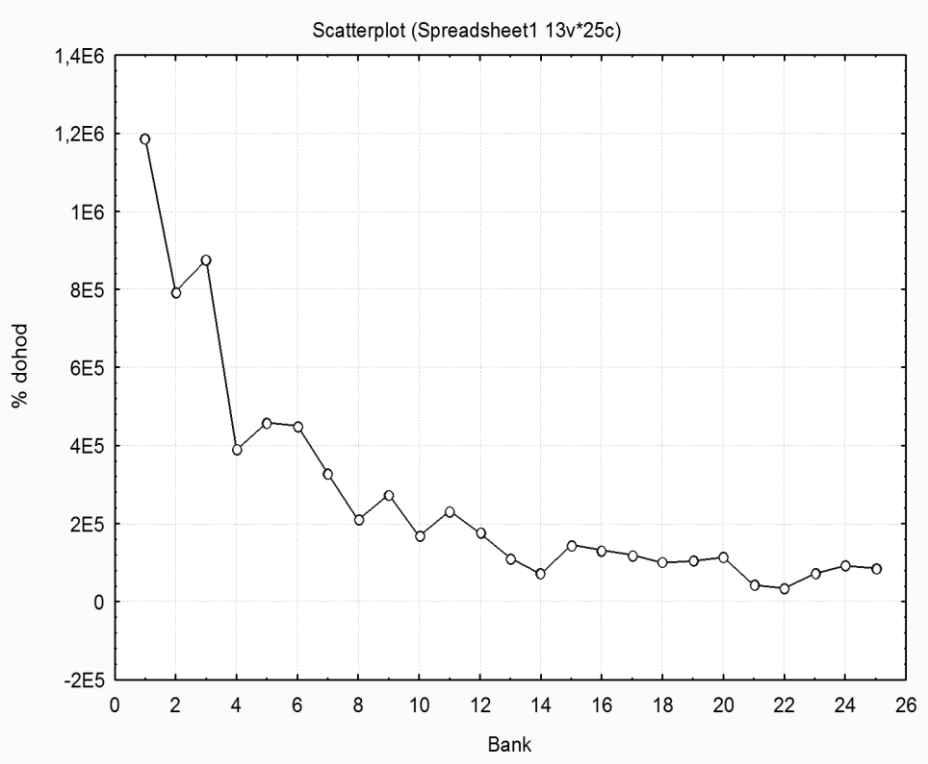


Рис. 9.11. **Полігон розподілу випадкової величини**

Подальший аналіз здійснюється в рамках перевірки вибірки на нормальний закон розподілу. Для проведення угрупування вибірки в стартовій панелі модуля вибираємо вкладиш *Normality*, де можна задавати бажану кількість інтервалів і критерій Колмогорова – Смірнова для тестування вибірки (рис. 9.12).

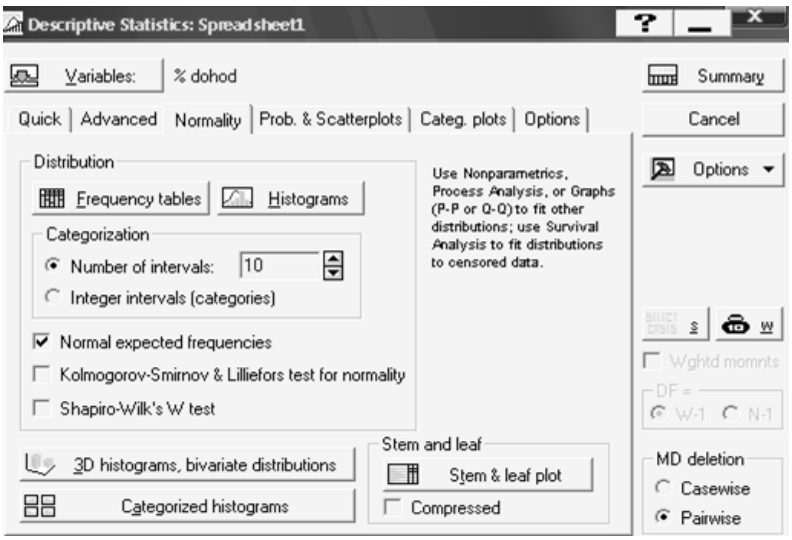


Рис. 9.12. **Вибір параметрів угрупування випадкової величини**

Ініціювавши клавішу *Frequency tables* (*Таблиці частот*), отримуємо наступну таблицю (рис. 9.13).

Frequency table: % dohod (Spr6)
K-S d=,23308, p<,15 ; Lilliefors p<,01

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All	Expected Count	Cumulative Expected	Percent Expected	Cumulative % Expected
-200000,<x<=0,000000	0	0	0,00000	0,0000	0,00000	0,0000	4,366527	4,36653	17,46611	17,46611
0,000000<x<=200000,0	15	15	60,00000	60,0000	60,00000	60,0000	5,718006	10,08453	22,87202	40,33813
200000,0<x<=400000,0	5	20	20,00000	80,0000	20,00000	80,0000	6,726783	16,81132	26,90713	67,24526
400000,0<x<=600000,0	2	22	8,00000	88,0000	8,00000	88,0000	4,999658	21,81097	19,99863	87,24389
600000,0<x<=800000,0	1	23	4,00000	92,0000	4,00000	92,0000	2,347086	24,15806	9,38835	96,63224
800000,0<x<=1000000,	1	24	4,00000	96,0000	4,00000	96,0000	0,695495	24,85355	2,78198	99,41422
1000000,<x<=1200000,	1	25	4,00000	100,0000	4,00000	100,0000	0,129962	24,98352	0,51985	99,93407
Missing	0	25	0,00000		0,00000	100,0000				

Рис. 9.13. **Результат угрупування вибірки**

Як видно, вихідна сукупність із двадцяти п'яти банків розбита на сім інтервалів. У кожному інтервалі розраховані такі характеристики: *Count*

(*частота*), *Cumulative Count* (*накопичена частота*), *Percent of Valid* (*% від загальної частоти*), *Cumul % of Valid* (*накопичений % від загальної частоти*), *% of all Cases* (*% від загального числа спостережень*), *Cumulative % of all Cases* (*накопичений % від загального числа спостережень*), *Expected Count* (*теоретична частота*), *Cumulative Expected* (*накопичена теоретична частота*), *% Expected* (*% від загальної теоретичної частоти*), *Cumulative % Expected* (*накопичений % від загальної теоретичної частоти*).

Ініціювавши клавішу *Histograms* (вкладиш *Normality*), створимо наступну гістограму розподілу з накладеною кривою нормального закону розподілу (рис. 9.14).

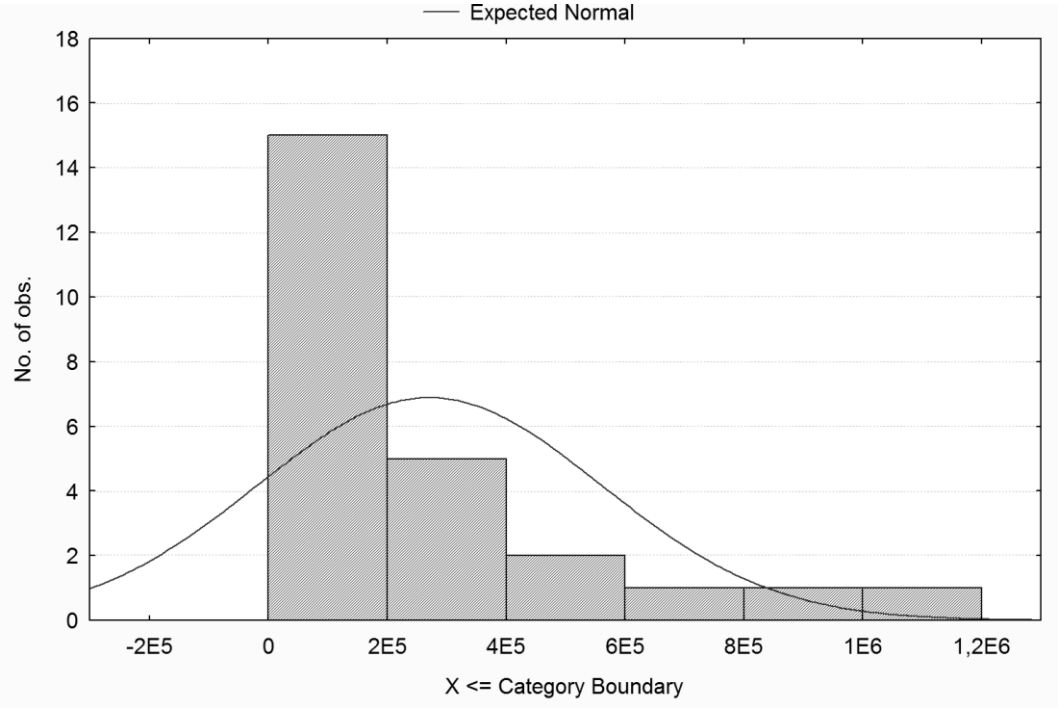


Рис. 9.14. **Гістограма розподілу**

Подальший аналіз вибірки передбачає розрахунок критерію Пірсона та Колмогорова – Смірнова для формування відповідних висновків про характер закону розподілу. Для визначення характеру закону розподілу та його відповідності нормальному закону дослідимо за допомогою графіків порівняння емпіричних і теоретичних частот і накопичених частот. Вихідні дані для побудови графіків та розраховані значення критерію Пірсона та Колмогорова – Смірнова наведені на рис. 9.15.

	1 % dohod	2 Interval	3 m	4 M	5 m'	6 M'	7 X2	8 M-M'
1	118747	-200000<x<=0,00000	0	0	4,36652	4,3665	4,3665	-4,3665
2	79382	0,000000<x<=20000	15	15	5,71800	10,0845	15,0673	4,9154
3	87614	200000<x<=400000	5	20	6,72678	16,8113	0,4432	3,1886
4	38971	400000<x<=600000	2	22	4,99965	21,8109	1,7997	0,1890
5	45923	600000<x<=800000	1	23	2,34708	24,1580	0,7731	-1,1580
6	45107	800000<x<=100000	1	24	0,69549	24,8535	0,1333	-0,8535
7	32813	1000000<x<=12000	1	25	0,12996	24,9835	5,8245	0,0164
8	20901					X^2	**28,407861**	
9	27394						**Kolmogorov**	**0,9830**

Рис. 9.15. **Аналіз закону розподілу випадкової величини**

Для побудови графіків інтервальних значень частоти розподілу досліджуваної сукупності необхідно зайти в меню *Graphs* / *2D Graphs* / *Scatterplots,* вибрати змінні та задати параметри графіка (рис. 9.16).

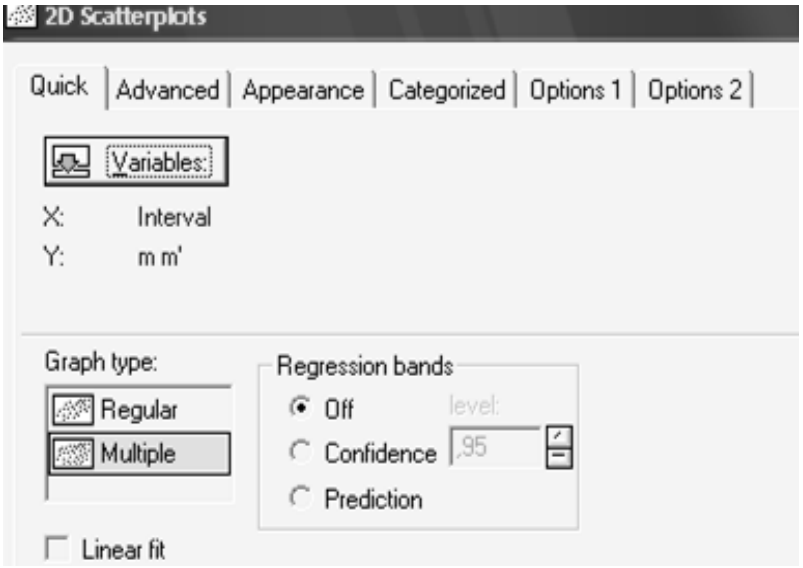


Рис. 9.16. **Вибір змінних для побудови графіків**

На рис. 9.17 і рис. 9.18 наведені графіки порівняння емпіричних і теоретичних частот і накопичених емпіричних і теоретичних частот, які дозволяють зробити висновки про відповідність нормальному закону розподілу та визначити розбіжність частот у кожному з досліджуваних інтервалів.

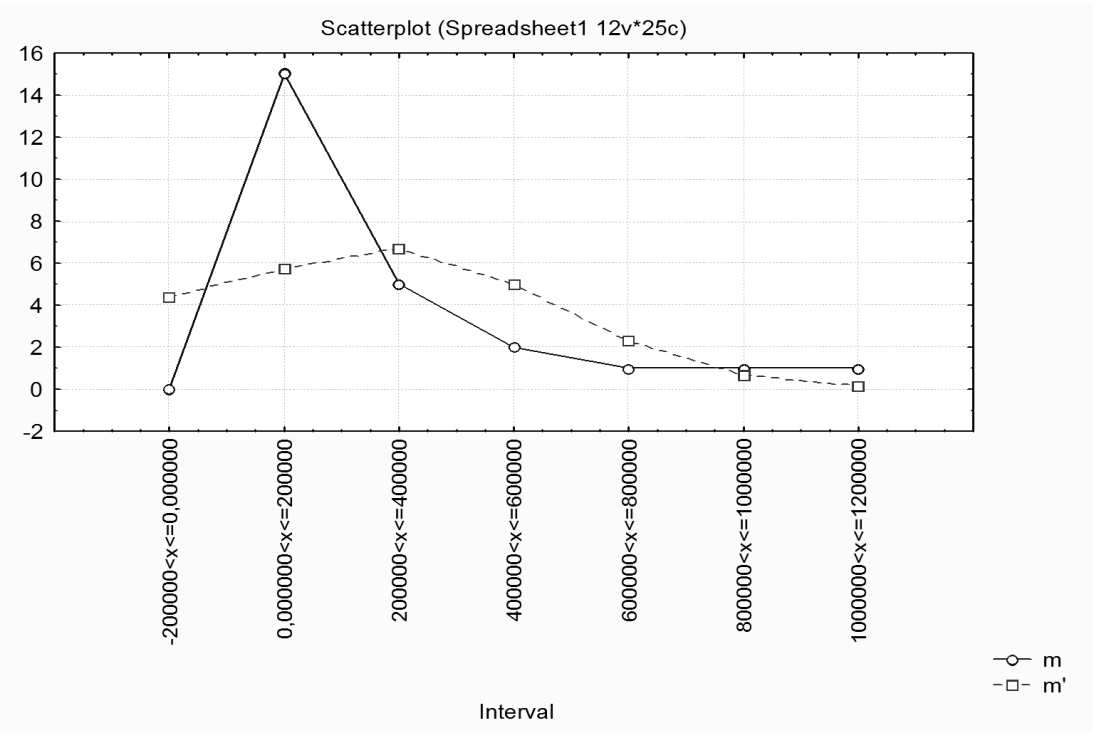


Рис. 9.17. **Графік порівняння емпіричних і теоретичних частот**

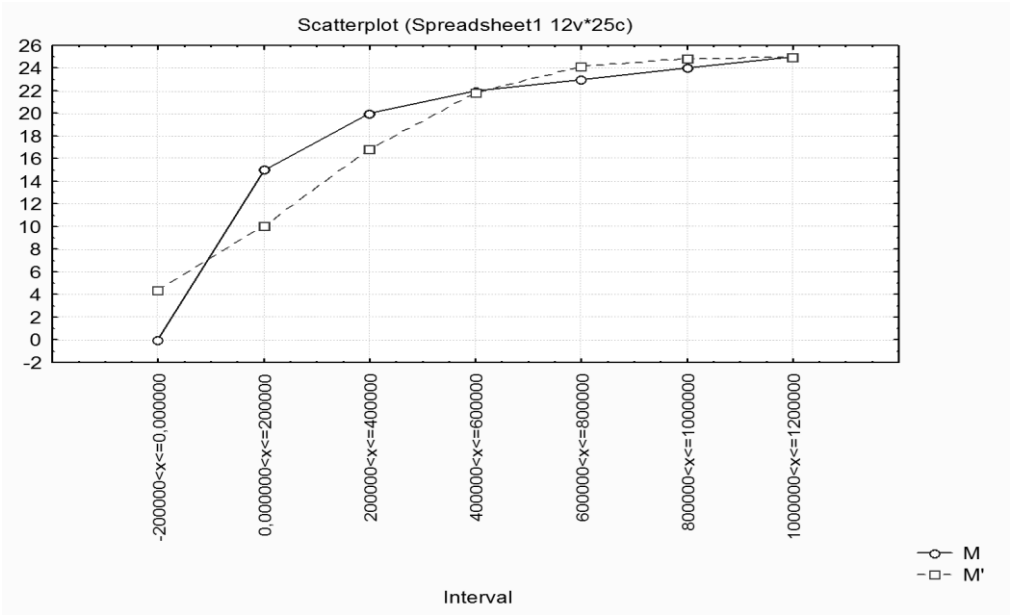


Рис. 9.18. **Графік порівняння накопичених емпіричних і теоретичних частот**

Далі робимо висновки про угрупування даних об'єктів за величиною показника % доходу. Порівнюємо отримані значення з табличними за відповідними критеріями та робимо висновки про характер закону розподілу.

Лабораторна робота 2. Методи та моделі кластерного аналізу. Класифікація без навчання

Мета – закріплення теоретичного та практичного матеріалу за темою "Методи кластерного аналізу. Класифікація без навчання"; набуття навичок роботи в модулі *Cluster Analysis.*

Завдання – необхідно побудувати моделі класифікації підприємств, використовуючи різні методи та стратегії класифікації для вибіркових даних в модулі *Cluster Analysis* ППП *Statistica:*

1) побудувати моделі кластерного аналізу, використовуючи ієрархічні (деревоподібні) методи кластерного аналізу;

2) порівняти результати дослідження, за різними правилами об'єднання та використовуючи різні метрики. Побудувати різні типи дендрограм класифікації. Зробити висновки;

3) провести класифікацію об'єктів за методом k-середніх, визначити характеристики моделі;

4) проаналізувати результати класифікації з різним значенням виділених кластерів, побудувати графіки, привести основні статистики та оцінювання змінних за отриманими моделями;

5) зробити висновки та подати економічну інтерпретацію отриманим результатам кластерних утворень.

Література: [5 – 9; 14; 41 – 44; 48; 49; 76].

Методичні рекомендації

Для розв'язання задач класифікації об'єктів у багатовимірному просторі та вивчення їх особливостей в ППП *Statistica* передбачений модуль *Cluster Analysis* (*Кластерний аналіз*). Розглянемо порядок роботи в даному модулі.

Таблиця вихідних даних для розв'язання задачі класифікації подана на рис. 9.19. Таким чином, задача дослідження полягає в отриманні класів однорідних об'єктів (підприємств) за такими показниками: x_1 – продуктивність праці; x_2 – коефіцієнт рентабельності капіталу; x_3 – коефіцієнт фондовіддачі. Щоб приступити до обчислювальних процедур, необхідно ввійти в позицію меню *Statistics* /*Multivariate Exploratory Techniques* / *Cluster Analysis* (рис. 9.20).

	1 X1	2 X2	3 X3
1	9,26	13,26	1,45
2	9,38	10,16	1,3
3	12,11	13,72	1,37
4	10,81	12,85	1,65
5	9,35	10,63	1,91
6	9,87	9,13	1,68
7	8,17	25,83	1,94
8	9,12	23,39	1,89
9	5,88	14,68	1,94
10	6,3	10,05	2,06
11	6,22	13,99	1,96
12	5,49	9,68	1,02
13	6,5	10,03	1,85
14	6,61	9,13	0,88
15	4,32	5,37	0,62
16	7,37	9,86	1,09
17	7,02	12,62	1,6
18	8,25	5,02	1,53
19	8,15	21,18	1,4
20	5,72	25,17	2,22
21	6,64	19,4	1,32
22	8,1	21	1,48
23	5,52	6,57	0,68
24	9,37	14,19	2,3
25	13,17	15,81	1,37

Рис. 9.19. **Вихідні дані**

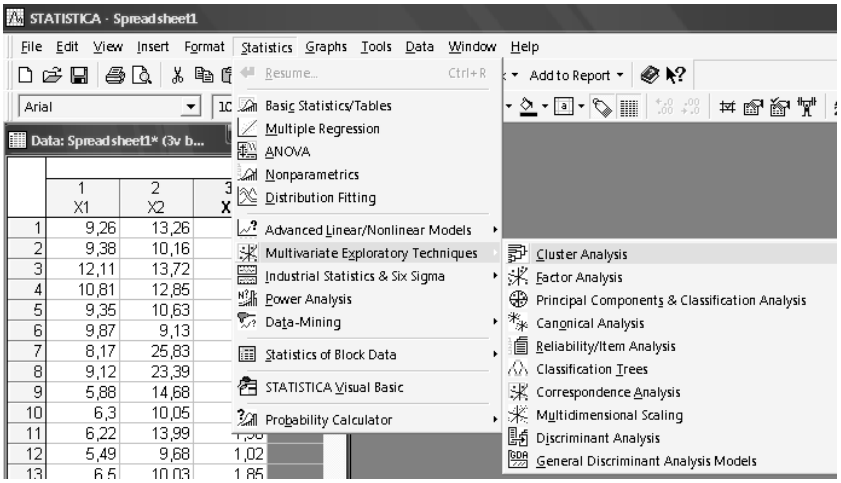


Рис. 9.20. **Вибір модуля**

Після підтвердження вибору модуля перед вами з'явиться стартова панель модуля (рис. 9.21), де необхідно вибрати напрям аналізу, тобто метод класифікації: *Joining tree clustering – (деревоподібна кластеризація); K-means clustering – (метод K-середніх); Two-way joining – (двовходова кластеризація).*

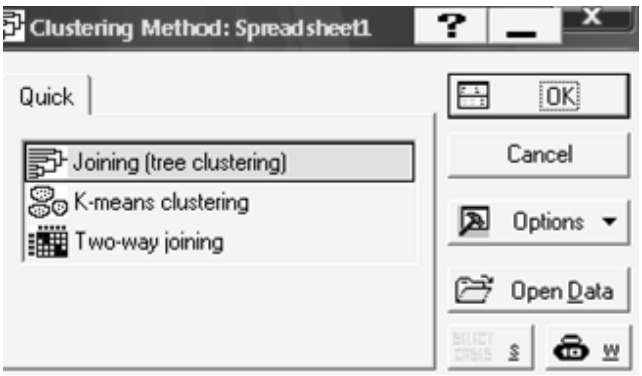


Рис. 9.21. **Стартова панель модуля**

Після підтвердження вибору методу (вибрано метод *Joining tree clustering*) необхідно задати вихідні параметри для проведення кластеризації (рис. 9.22): *Variable* (Змінні), *Cluster* (Об'єкти кластеризації), *Amalgamation rule* (Правила кластеризації), *Distance measure* (Міру подібності).

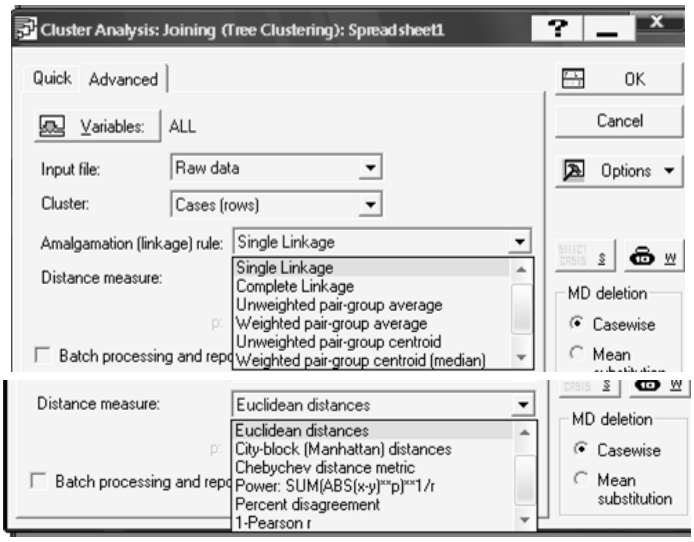


Рис. 9.22. **Визначення вихідних параметрів**

У розглядуваному модулі представлено такі правила ієрархічного об'єднання:

1) *Single linkage* – одиночного зв'язку;

2) *Complete linkage* – повних зв'язків;

3) *Unweighted pair-group average* – незваженого попарного середнього;

4) *Weighted pair-group average* – зваженого попарного середнього;

5) *Unweighted pair-group centroid* – незваженого центроїдного;

6) *Weighted pair-group centroid* – зваженого центроїдного;

7) Ward's method – метод Уорда.

В якості міри подібності використовують метрики:

1) *Euclidean distances* – евклідова метрика;

2) *Square Euclidean distances* – квадрат евклідової метрики;

3) *City-block (Manhattan) distances* – Мангеттенська відстань;

4) *Chebychev distances metric* – відстань Чебишева;

5) *Power metric* – степенева відстань Мінковського;

6) *Percent disagreement* – відсоток незгоди (для категоріальних даних);

7) *(1 – Personal r)* – (1 – коефіцієнт кореляції Пірсона).

Для дослідження виберемо процедуру *Single linkage (одиночного зв'язку)* та звичайну е*вклідову метрику* (*Euclidean distances).* Підтверджуючи вибір, ініціюйте клавішу (*OK*). З'явиться вікно результатів, де в верхній частині подана основна інформація та вибрані процедури дослідження. Опції в нижній частині вікна на вкладці *Advanced* призначені для аналізу результатів кластеризації (рис. 9.23): *Horizontal hierarchical tree plot* (горизонтальна деревоподібна діаграма); *Vertical icicle plot* (вертикальна деревоподібна діаграма – дендрограма); *Amalgamation schedule* (правило об'єднання в кластери); *Graf of amalgamation schedule* (графік порядку об'єднання); *Distance matrix* (матриця відстаней); *Descriptive statistics* (описові статистики).

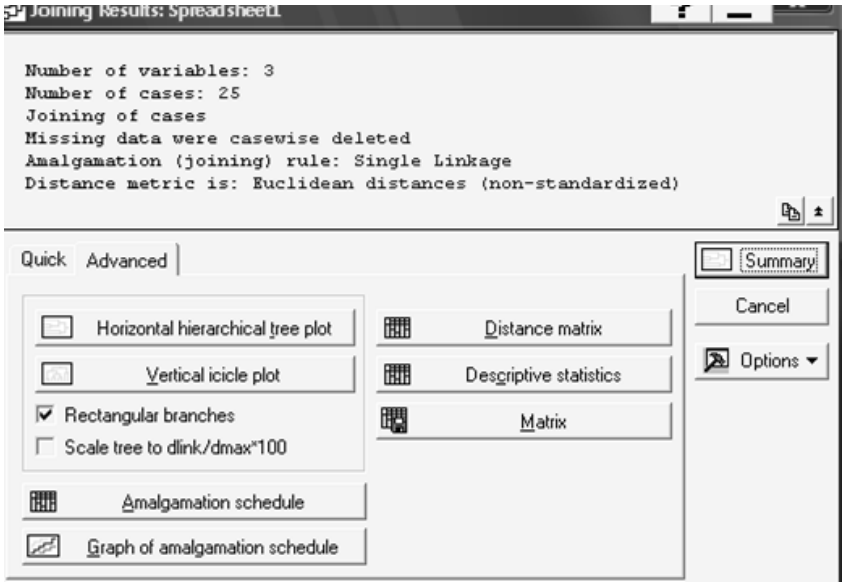


Рис. 9.23. **Опції аналізу результатів**

Ініціювавши клавішу *Vertical icicle plot,* отримаємо дендрограму класифікації (рис. 9.24), де на осі абсцис подані об'єкти дослідження, а на осі ординат – відстані між ними.

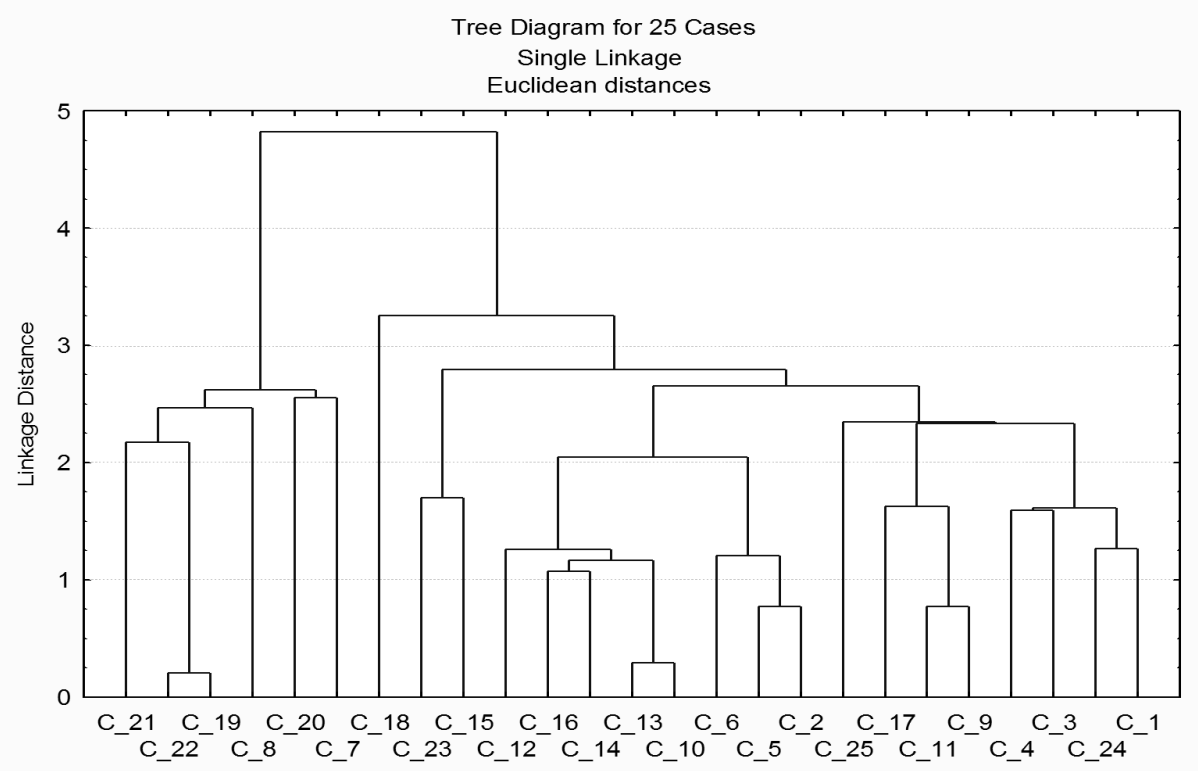


Рис. 9.24. **Дендрограма класифікації**

Ініціювавши клавішу *Distance matrix,* отримаємо матрицю відстаней, фрагмент якої наведено на рис. 9.25. На рис. 9.26 наведено фрагмент матриці об'єднання (*Amalgamation schedule)*.

	Euclidean distances (Spreadsheet1)												
Case No	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13
C_1	0,0	3,1	2,9	1,6	2,7	4,2	12,6	10,1	3,7	4,4	3,2	5,2	4,3
C_2	3,1	0,0	4,5	3,1	0,8	1,2	15,7	13,2	5,8	3,2	5,0	3,9	2,9
C_3	2,9	4,5	0,0	1,6	4,2	5,1	12,7	10,1	6,3	6,9	5,9	7,8	6,7
C_4	1,6	3,1	1,6	0,0	2,7	3,8	13,2	10,7	5,3	5,3	4,7	6,2	5,2
C_5	2,7	0,8	4,2	2,7	0,0	1,6	15,2	12,8	5,3	3,1	4,6	4,1	2,9
C_6	4,2	1,2	5,1	3,8	1,6	0,0	16,8	14,3	6,8	3,7	6,1	4,5	3,5
C_7	12,6	15,7	12,7	13,2	15,2	16,8	0,0	2,6	11,4	15,9	12,0	16,4	15,9
C_8	10,1	13,2	10,1	10,7	12,8	14,3	2,6	0,0	9,3	13,6	9,8	14,2	13,6
C_9	3,7	5,8	6,3	5,3	5,3	6,8	11,4	9,3	0,0	4,7	0,8	5,1	4,7
C_10	4,4	3,2	6,9	5,3	3,1	3,7	15,9	13,6	4,7	0,0	3,9	1,4	0,3
C_11	3,2	5,0	5,9	4,7	4,6	6,1	12,0	9,8	0,8	3,9	0,0	4,5	4,0

Рис. 9.25. **Матриця класифікації**

	Amalgamation Schedule (Spreadsheet1) Single Linkage Euclidean distances							
linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8
,203224(	C_19	C_22						
,290688	C_10	C_13						
,769480	C_9	C_11						
,770649	C_2	C_5						
1,074523	C_14	C_16						
1,167647	C_10	C_13	C_14	C_16				
1,202248	C_2	C_5	C_6					
1,255587	C_10	C_13	C_14	C_16	C_12			
1,264713	C_1	C_24						
1,589119	C_3	C_4						
1,615735	C_1	C_24	C_3	C_4				
1,626807	C_9	C_11	C_17					
1,698117	C_15	C_23						

Рис. 9.26. **Матриця об'єднання**

Дендрограма класифікації за методом Уорда наведена на рис. 9.27. Аналіз дендрограми дозволяє розпізнати три групи (кластери) однорідних станів в спостережуваній сукупності даних.

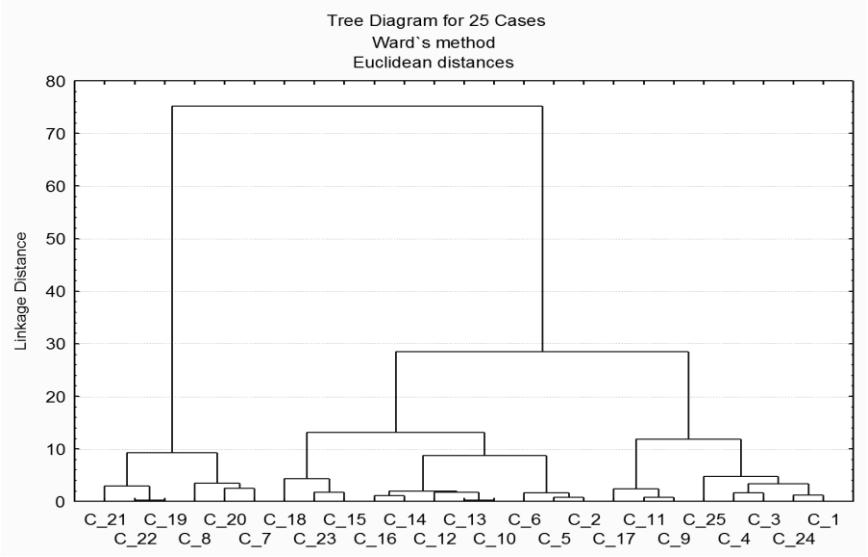


Рис. 9.27. **Дендрограма класифікації за методом Уорда**

Розглянемо реалізацію методу *К-середніх (K-means clustering)*. Вибір методу на стартовій панелі модуля (рис. 9.28). Ініціювавши дану опцію, отримаємо діалогове вікно реалізації методу, де на вкладці *Advanced* необхідно задати змінні для аналізу (*Variables)*, об'єкти кластеризації *(Cluster)*, число кластерів *(Number of clusters)*, число ітерацій *(Number of iterations)*, та початкові центри кластерів – опції (*Initial cluster centers*) (див. рис. 9.28).

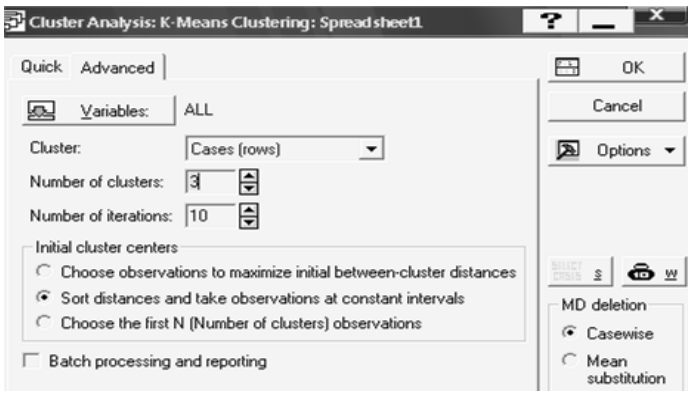


Рис. 9.28. **Діалогове вікно методу k-середніх**

Вікно аналізу результатів методу k-середніх наведено на рис. 9.29. Верхня частина є інформаційною, а нижня (вкладка *Advanced*) дозволяє отримати повну інформацію результатів аналізу. Розглянемо функціональне значення опцій даного вікна:

Summary: Cluster means & Euclidean distances (евклідові відстані та середні значення станів кластерів);

Analysis of variance (дисперсійний аналіз);

Graf of means (графік середніх значень);

Descriptive statistics for each cluster (описові статистики для кластерів);

Members of each cluster & distances (члени кластерів та їх відстані до центру кластера);

Save classifications and distances (збереження результатів кластеризації).

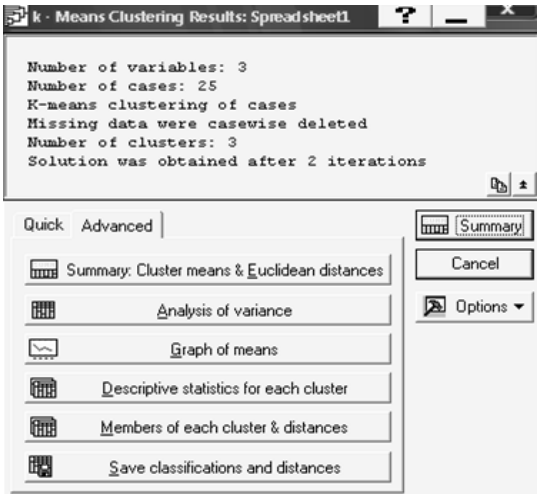


Рис. 9.29. **Опції аналізу результатів методу *k*-середніх**

Евклідові відстані між отриманими кластерами та середні значення для кожного досліджуваного показника подані на рис. 9.30, причому значення евклідових відстаней знаходиться під головною діагоналлю, а над – квадрат евклідових відстаней.

Cluster Number	Euclidean Distances between Distances below diagonal Squared distances above dia No. 1	No. 2	No. 3
No. 1	0,00000(	10,2320{	28,6556
No. 2	3,19876:	0,0000(	67,0729:
No. 3	5,35309'	8,1898:	0,0000(

Variable	Cluster Means (Spreadsheet1 Cluster No. 1	Cluster No. 2	Cluster No. 3
X1	9,2433:	6,96100(	7,6500(
X2	13,5277{	8,50000(	22,6616
X3	1,7277{	1,27100(	1,7083:

Рис. 9.30. **Евклідові відстані та середні значення станів кластерів**

Графік середніх значень для кластерів станів приведено на рис. 9.31. Як видно, найбільш кластери різняться за показником x_2, потім x_1 і дуже малі відмінності в середніх для показника x_3.

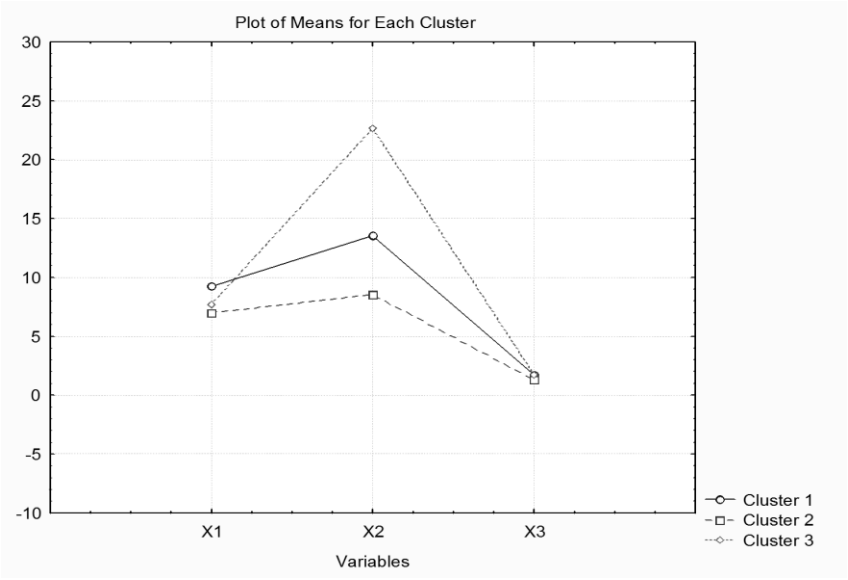


Рис. 9.31. **Графік середніх значень для кластерів станів**

Результати дисперсійного аналізу оцінки якості показників наведено на рис. 9.32. У таблиці наведені значення міжгрупових і внутрішньогрупових дисперсій ознак. Чим менше значення внутрішньогрупової дисперсії і більше значення міжгрупової, тим краще ознака характеризує приналежність об'єктів до кластера. Параметри *F* і *p* визначають внесок ознаки в класифікацію.

Variable	Analysis of Variance (Spreadsheet1)					
	Between SS	df	Within SS	df	F	signif. p
X1	25,375	2	87,2736	22	3,1983	0,06035
X2	752,533	2	86,6000	22	95,5873	0,00000
X3	1,211	2	3,6763	22	3,6233	0,04363

Рис. 9.32. **Таблиця дисперсійного аналізу**

На рис. 9.33. показані описові статистики для виділених кластерів, а саме: середнє, середньоквадратичне відхилення та дисперсія.

Variable	Descriptive Statistics for Cluste Cluster contains 9 cases		
	Mean	Standard Deviation	Variance
X1	9,2433	2,53757	6,43930
X2	13,5277	1,45677	2,12219
X3	1,7277	0,31920	0,10189

Variable	Descriptive Statistics for Cluste Cluster contains 6 cases		
	Mean	Standard Deviation	Variance
X1	7,6500	1,23521	1,52576
X2	22,6616	2,54776	6,49109
X3	1,7083	0,35957	0,12929

Variable	Descriptive Statistics for Cluste Cluster contains 10 cases		
	Mean	Standard Deviation	Variance
X1	6,96100	1,76793	3,12561
X2	8,50000	2,03215	4,12966
X3	1,27100	0,49606	0,24607

Рис. 9.33. **Описові статистики для кластерів**

Члени кластерів та їх відстані до центру відповідного кластеру наведено на рис. 9.34. Таким чином, дані таблиці дозволяють визначити склад кожного кластера.

	Members of Cluster Number 1 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 9 cases								
	Case No. C_1	Case No. C_3	Case No. C_4	Case No. C_5	Case No. C_9	Case No. C_11	Case No. C_17	Case No. C_24	Case No. C_25
Distance	0,22296	1,67159	0,98655	1,67746	2,05626	1,77088	1,38847	0,51056	2,63028

	Members of Cluster Number 2 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 10 cases									
	Case No. C_2	Case No. C_6	Case No. C_10	Case No. C_12	Case No. C_13	Case No. C_14	Case No. C_15	Case No. C_16	Case No. C_18	Case No. C_23
Distance	1,69391	1,73459	1,07423	1,09836	0,98126	0,47363	2,39413	0,82656	2,14779	1,43185

	Members of Cluster Number 3 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 6 cases					
	Case No. C_7	Case No. C_8	Case No. C_19	Case No. C_20	Case No. C_21	Case No. C_22
Distance	1,85853	0,95295	0,92021	1,85098	1,98405	1,00262

Рис. 9.34. **Члени кластерів та їх відстані до центру кластера**

Так, кластер № 3 (шість підприємств) має найвищій рівень рентабельності, а показники продуктивності праці та фондовіддача знаходяться

на середньому рівні. Кластеру № 1 (дев'ять підприємств) властиві найвищі показники продуктивності праці та фондовіддачі з середнім значенням рентабельності капіталу. Кластеру № 2 (десять підприємств) притаманні найнижчі значення за всіма досліджуваними показниками.

Лабораторна робота 3. Методи та моделі дискримінантного аналізу. Класифікація з навчанням

Мета – закріплення теоретичного та практичного матеріалу за темою "Методи та моделі дискримінантного аналізу"; набуття навичок роботи в модулі *Discriminant Analysis.*

Завдання – необхідно побудувати модель класифікації підприємств і провести розпізнавання для вибіркових даних у модулі *Discriminant Analysis* ППП *Statistica:*

1) побудувати модель дискримінантного аналізу на основі вибіркових даних;

2) дати оцінку якості моделей розпізнавання, значущості змінних та провести канонічний аналіз функцій;

3) побудувати моделі, використовуючи методи покрокового аналізу включення та виключення факторних змінних; дати оцінку якості побудованих моделей та дискримінації змінних;

4) проаналізувати результати розпізнавання (матриця класифікацій), подати теоретичну класифікацію за дискримінантною моделлю;

5) зробити висновки та прогнози (розпізнавання) за побудованою моделлю. Дати економічну інтерпретацію отриманим результатам.

Література: [5 – 9; 14; 41 – 44; 48; 49; 76].

Методичні рекомендації

Для розв'язання та аналізу задач класифікації багатомірних сукупностей за наявності навчальних вибірок (класифікація з навчанням) у ППП *Statistica* передбачений модуль *Discriminant Analysis* (*Дискримінантний аналіз*). Розглядуваний модуль має широкий набір засобів, які забезпечують проведення дискримінантного аналізу даних, візуалізації

та інтерпретації результатів. Розглянемо порядок роботи в даному модулі. Таблиця вихідних даних для розв'язання задачі розміщена на рис. 9.35.

	1 X1	2 X2	3 X3	4 X4	5 X5	6 Class
1	9,26	0,78	1,37	0,23	1,45	A
2	9,38	0,75	1,49	0,39	1,3	A
3	10,81	0,7	1,42	0,18	1,65	A
4	9,35	0,62	1,35	0,15	1,91	A
5	9,87	0,76	1,39	0,34	1,68	A
6	9,12	0,71	1,27	0,09	1,89	A
7	6,61	0,72	1,23	0,48	0,88	B
8	4,32	0,68	1,39	0,41	0,62	B
9	7,37	0,77	1,38	0,62	1,09	B
10	6,64	0,77	1,35	0,5	1,32	B
11	5,52	0,72	1,48	1,2	0,68	B
12	9,37	0,79	1,24	0,21	2,3	A
13	5,68	0,71	1,28	0,66	1,43	B
14	5,22	0,79	1,33	0,74	1,82	B
15	10,02	0,76	1,22	0,32	2,62	A
16	6,7	0,79	0,79	0,39	1,24	B
17	9,42	0,7	0,7	0,72	2,03	A

Рис. 9.35. **Вхідні дані**

В якості факторів впливу на рівень інвестиційної привабливості розглядаються такі коефіцієнти: продуктивність праці (x_1); питома вага робітників у складі промислово-виробничого персоналу (x_2); коефіцієнт змінності (x_3); коефіцієнт браку (x_4); коефіцієнт фондовіддачі ОВФ (x_5). Значення показників розподілене на навчальні вибірки (група А – інвестиційно привабливі підприємства, група В – підприємства-аутсайдери). Щоб приступити до обчислювальних процедур, необхідно ввійти в меню *Statistics* / *Multivariate Exploratory Techniques* / *Discriminant Analysis* (рис. 9.36).

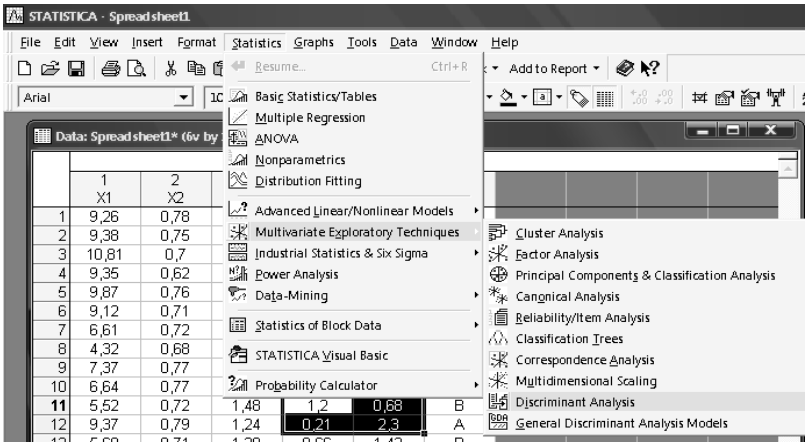


Рис. 9.36. **Вибір модуля**

Після підтвердження вибору модуля перед вами з'явиться стартова панель модуля, де необхідно задати вихідні параметри моделювання (рис. 9.37).

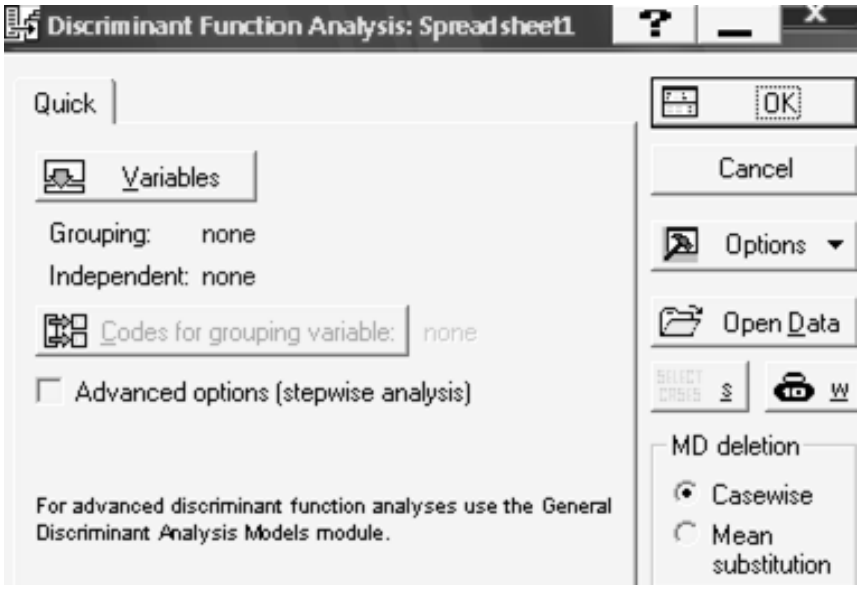


Рис. 9.37. **Стартова панель модуля**

Ініціюйте кнопку *Variables (Змінні);* у вікні, що з'явилося, вкажіть показники, за якими будете здійснювати аналіз (рис. 9.38). Після вибору залежних (групувальна – залежна змінна) та незалежних змінних підтвердьте свій вибір натисканням кнопки *OK*. Також у вікні можна задати коди для значень групувальної змінної.

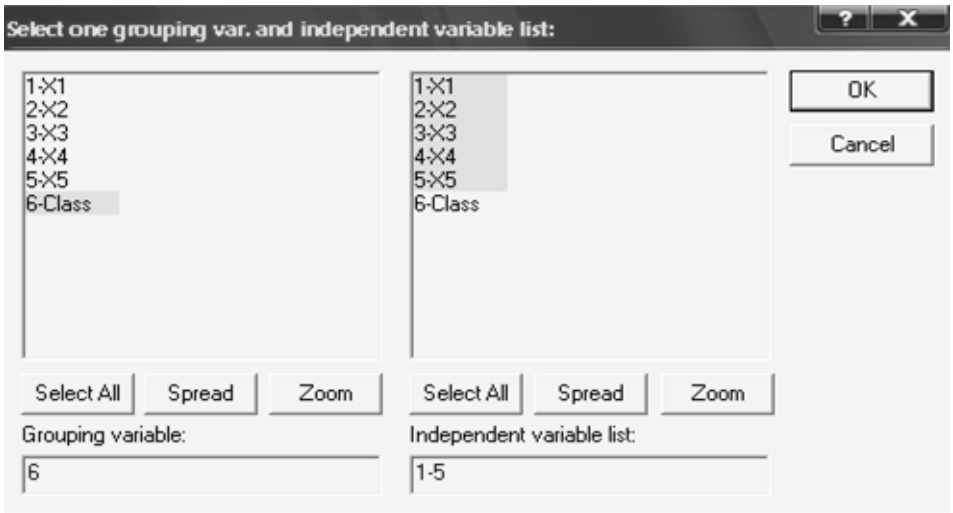


Рис. 9.38. **Вибір змінних для аналізу**

Стартова панель модуля *Discriminant Analysis* з опціями для аналізу наведена на рис. 9.39.

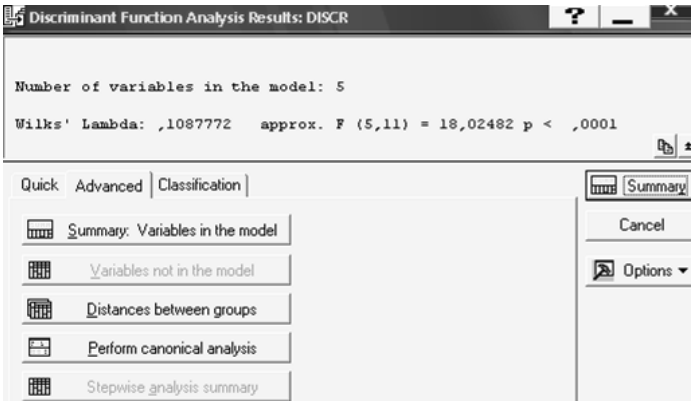


Рис. 9.39. **Результати дискримінантного аналізу**

У верхній інформаційній частині вікна показані значення лямбди Уїлкса (λ), яка характеризує якість дискримінації та змінюється в межах [0,1]. Значення, близькі до нуля, визначають гарну якість дискримінації. Значення критерію Фішера (*F*) – оцінки адекватності моделі наведено для порівняння з табличними значеннями.

У нижній частині вікна подано різноманітні опції для досконалого аналізу моделі та побудови графіків. Ініціювавши опцію *Summary: Variables in the model (Аналіз змінних у моделі)* на вкладці Advanced, отримаємо результати дискримінантного аналізу, наведені на рис. 9.40.

	Discriminant Function Analysis Summary (Spreadsheet1) No. of vars in model: 5; Grouping: Class (2 grps) Wilks' Lambda: ,10878 approx. F (5,11)=18,025 p< ,0001					
N=17	Wilks' Lambda	Partial Lambda	F-remove (1,11)	p-level	Toler.	1-Toler. (R-Sqr.)
X1	0,357719	0,304086	25,17404	0,000392	0,890477	0,109523
X2	0,125683	0,865487	1,70961	0,217705	0,847674	0,152326
X3	0,113186	0,961049	0,44583	0,518085	0,847547	0,152453
X4	0,109932	0,989494	0,11680	0,738972	0,968503	0,031497
X5	0,125269	0,868346	1,66776	0,223036	0,788661	0,211339

Рис. 9.40. **Результати дискримінантного аналізу**

Розглянемо стовпці таблиці:

Wilks' Lambda (лямбда Уїлкса) – лямбда Уїлкса, яка є результатом виключення відповідної змінної з моделі. Чим більше значення лямбди Уїлкса, тим значуща змінна в процедурі дискримінації;

Partial Lambda (частинна лямбда) – характеризує одиничний внесок відповідної змінної в дискримінацію моделі. Чим менше дана статистика, тим більший її вклад у загальну дискримінацію;

F-remove (*F*-виключення) – значення *F*-критерію для відповідної змінної;

p-level (рівень значущості) – значення *F*-критерію для відповідної змінної;

Toler (толерантність) – визначається як (1 - R^2), де R^2 – коефіцієнт множинної кореляції даної змінної зі всіма іншими змінними в моделі. Толерантність є мірою збитковості змінних у моделі.

Таким чином, можна дійти висновку, що найбільш значущою змінною для дискримінації є змінна x_1, x_2, x_5. Ініціювавши клавішу *Distances between groups (Відстані між групами)*, отримаємо таблицю відстаней (рис. 9.41), яка характеризує якість дискримінації спостережень і ступінь відмінностей (неоднорідність) груп.

	Squared Mahalano	
Class	A	B
A	0,0000(	32,8861
B	32,8861	0,0000(

Рис. 9.41. **Матриця відстаней між групами**

Ініціюванням опції *Perform canonical analysis (Канонічний аналіз)* відкриється меню напрямів канонічного аналізу (рис. 9.42).

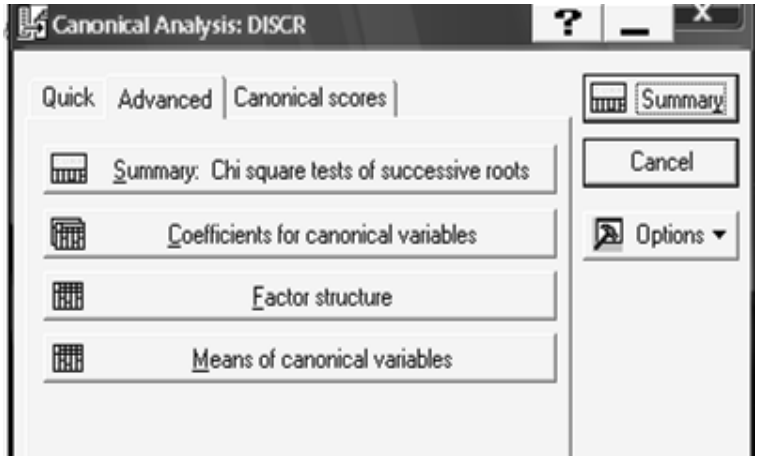


Рис. 9.42. **Опції канонічного аналізу функцій**

Summary: Chi square tests of successive roots (Підсумки: χ^2 -критерій послідовності коренів) (рис. 9.43).

Roots Removed	Chi-Square Tests with Successive Roots Removed (I					
	Eigen-value	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	8,19309	0,94404	0,10877	27,7306	5	0,00004

Рис. 9.43. **χ^2 -критерій канонічних коренів**

Отримані результати дозволяють провести оцінювання кількості значущих коренів для інтерпретації та статистичну значущість дискримінантних функцій.

Coefficients for canonical variables (Коефіцієнти канонічних змінних). Ініціювання даної опції дозволяє отримати таблиці нестандартизованих і стандартизованих коефіцієнтів дискримінантних функцій (рис. 9.44).

Variable	Raw Coef for Canor	Variable	Standardi for Canon
	Root 1		Root 1
X1	-1,1981	X1	-0,93642
X2	8,7874	X2	0,42196
X3	-1,0209	X3	-0,22708
X4	0,4891	X4	0,11032
X5	-1,0597	X5	-0,43279
Constant	5,7166		
Eigenval	8,1931	Eigenval	8,19309
Cum.Prop	1,0000	Cum.Prop	1,00000

Рис. 9.44. **Коефіцієнти канонічних змінних**

Ці результати використовують для визначення: значень канонічних змінних для кожного спостереження; ступеня та напряму впливу змінних у кожній дискримінантній функції. *Factor structure – факторна структура.* У таблиці на рис. 9.45 показані об'єднані внутрішньогрупові кореляції змінних з відповідними дискримінантними функціями, які використовуються для змістовної інтерпретації функцій. *Means of canonical variables*

(Середні значення канонічних змінних) (рис. 9.45) містить середні значення для дискримінантних функцій, які дозволяють визначити групи, які найкраще ідентифікуються конкретною дискримінантною функцією.

Variable	Factor Stru Correlatior (Pooled-w
	Root 1
X1	-0,85857
X2	0,05315
X3	0,00544
X4	0,27388
X5	-0,33411

Group	Means of
	Root 1
A	-2,5349
B	2,8518

Рис. 9.45. **Факторна структура та середні значення канонічних змінних**

Для побудови дискримінантних функцій необхідно ініціювати опцію *Classification functions (Функції класифікації)* на вкладці (*Classification)* (рис. 9.46).

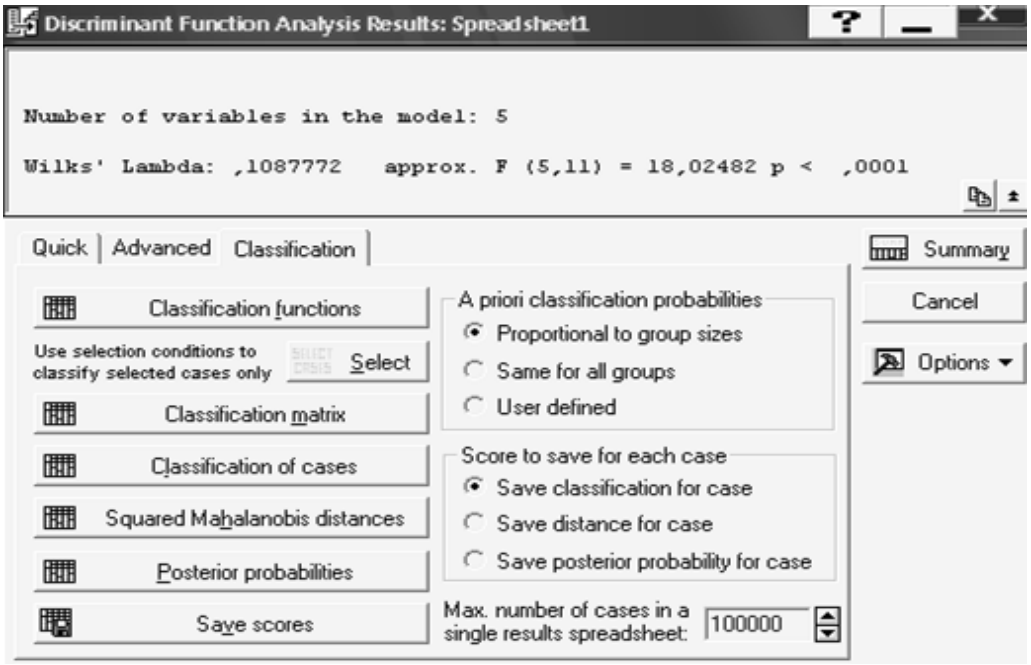


Рис. 9.46. **Опції результатів класифікації**

Дискримінантні функції для виділених класів станів підприємств (A, B) подані на рис. 9.47. Таким чином, лінійні дискримінантні функції мають такий вигляд:

$$y_1 = -192{,}44 + 13{,}05 \cdot x_1 + 252{,}88 \cdot x_2 + 42{,}08 \cdot x_3 + 15{,}18 \cdot x_4 + 8{,}3 \cdot x_5;$$
$$y_2 = -162{,}62 + 6{,}59 \cdot x_1 + 300{,}21 \cdot x_2 + 36{,}58 \cdot x_3 + 17{,}82 \cdot x_4 + 2{,}59 \cdot x_5.$$

	Classification Functi	
Variable	A p=,52941	B p=,47059
X1	13,045	6,591
X2	252,878	300,214
X3	42,081	36,582
X4	15,184	17,819
X5	8,301	2,592
Constant	-192,444	-162,621

Рис. 9.47. **Дискримінантні функції**

Класифікаційна матриця (*Classification matrix)* містить інформацію про кількість і відсоток коректно класифікованих спостережень у кожній з груп. Рядки матриці – вихідні класи, стовпці – розпізнані класи за моделлю (рис. 9.48).

	Classification Matrix (Spreads Rows: Observed classification Columns: Predicted classificat		
Group	Percent Correct	A p=,52941	B p=,4705
A	100,000	9	0
B	100,000	0	8
Total	100,000	9	8

Рис. 9.48. **Матриця класифікацій**

Для визначення приналежності об'єкта до одного з виділених класів на основі побудованої дискримінантної моделі необхідно скористатися такими опціями:

1) *Classification of cases (Класифікація спостережень)* – таблиця класифікації для кожного спостереження;

2) *Squared Mahalanobis distances (Квадрати відстаней Махаланобіса)* – таблиця квадратів відстаней Махаланобіса для кожного спостереження до центру відповідної групи;

3) *Posterior probabilities (Апостеріорні ймовірності)* – таблиця апостеріорних ймовірностей приналежності кожного спостереження до відповідної групи.

Реалізація даного напряму аналізу подана на рис. 9.49. Об'єкти, які неправильно класифіковані, будуть відмічені (*****).

| | Classification of Cases (Spread
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | 1 p=,52941 | 2 p=,47059 |
| 1 | A | A | B |
| 2 | A | A | B |
| 3 | A | A | B |
| 4 | A | A | B |
| 5 | A | A | B |
| 6 | A | A | B |
| 7 | B | B | A |
| 8 | B | B | A |
| 9 | B | B | A |
| 10 | B | B | A |
| 11 | B | B | A |
| 12 | A | A | B |
| 13 | B | B | A |
| 14 | B | B | A |
| 15 | A | A | B |
| 16 | B | B | A |
| 17 | A | A | B |

| | Squared Mahalanobis Distanc
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | A p=,52941 | B p=,47059 |
| 1 | A | 3,7782 | 19,9943 |
| 2 | A | 3,1758 | 22,5451 |
| 3 | A | 4,1865 | 51,0808 |
| 4 | A | 6,5368 | 44,5157 |
| 5 | A | 1,0908 | 29,3406 |
| 6 | A | 1,5051 | 27,2026 |
| 7 | B | 23,5515 | 1,8762 |
| 8 | B | 55,9241 | 7,6365 |
| 9 | B | 16,8742 | 3,5854 |
| 10 | B | 21,1234 | 1,3401 |
| 11 | B | 47,1895 | 8,1167 |
| 12 | A | 2,8158 | 27,8854 |
| 13 | B | 28,4881 | 1,6361 |
| 14 | B | 41,5080 | 5,7259 |
| 15 | A | 3,7939 | 42,9479 |
| 16 | B | 35,0980 | 7,7022 |
| 17 | A | 10,4973 | 33,0227 |

| | Posterior Probabilities (Spread
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | A p=,52941 | B p=,47059 |
| 1 | A | 0,99973 | 0,00026 |
| 2 | A | 0,99994 | 0,00005 |
| 3 | A | 1,00000 | 0,00000 |
| 4 | A | 1,00000 | 0,00000 |
| 5 | A | 0,99999 | 0,00000 |
| 6 | A | 0,99999 | 0,00000 |
| 7 | B | 0,00002 | 0,99997 |
| 8 | B | 0,00000 | 1,00000 |
| 9 | B | 0,00146 | 0,99853 |
| 10 | B | 0,00005 | 0,99994 |
| 11 | B | 0,00000 | 1,00000 |
| 12 | A | 0,99999 | 0,00000 |
| 13 | B | 0,00000 | 0,99999 |
| 14 | B | 0,00000 | 1,00000 |
| 15 | A | 1,00000 | 0,00000 |
| 16 | B | 0,00000 | 0,99999 |
| 17 | A | 0,99998 | 0,00001 |

Рис. 9.49. **Матриці розпізнавання стану підприємств**

Більш детальний аналіз змінних передбачає побудову моделей за методами покрокового включення та виключення незалежних змінних.

У даному модулі реалізовано метод покрокового включення змінних *(Forward stepwise)* і метод покрокового виключення *(Backward stepwise).* Вибір методів здійснюється на стартовій панелі модуля ініціюванням опції *Advanced options (stepwise analysis)* (рис. 9.50).

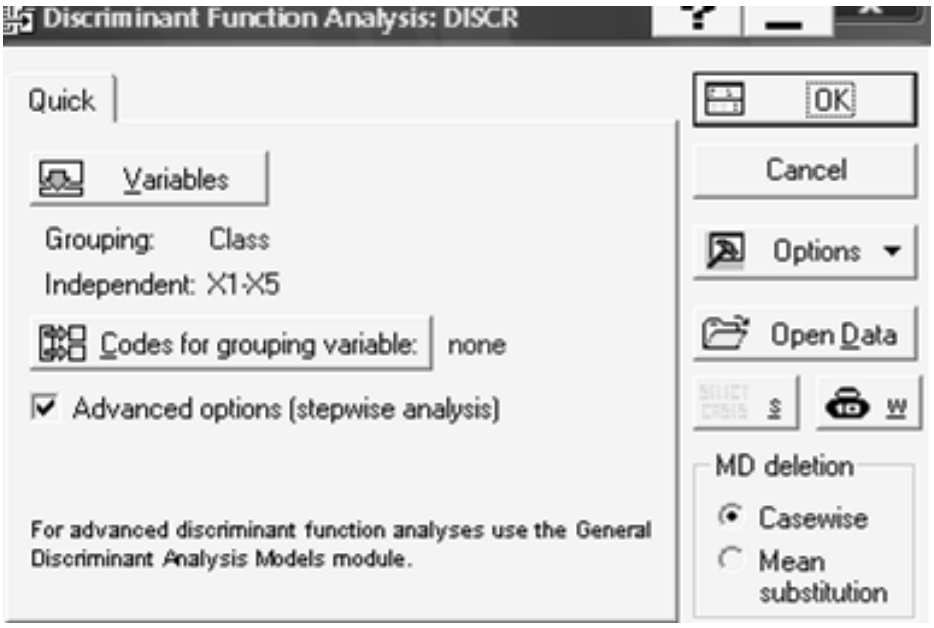


Рис. 9.50. **Вибір опцій покрокового аналізу**

Вибір методу оцінювання, порогові значення *F*-критерію включення або виключення, послідовність подання результатів вибирається у вкладці *Advanced* (рис. 9.51).

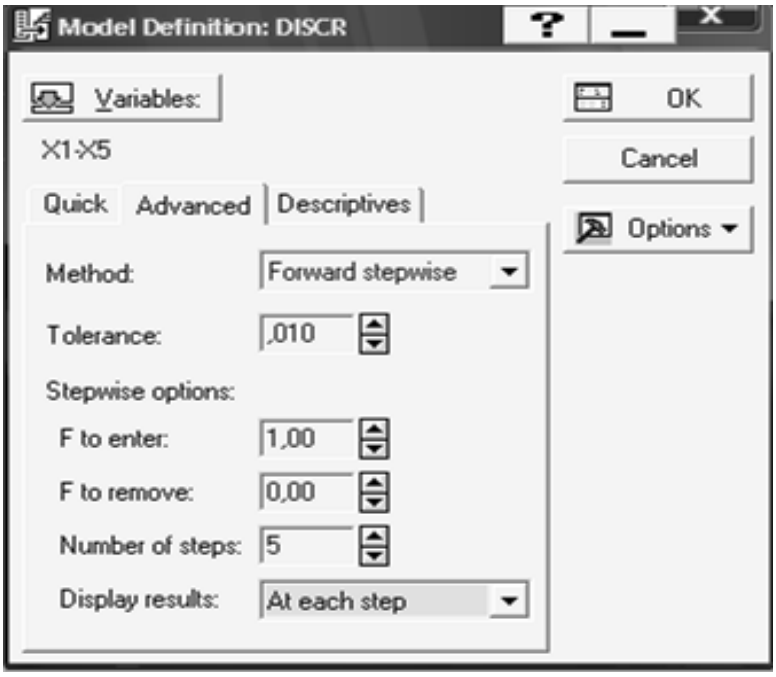


Рис. 9.51. **Вибір методу покрокового дискримінантного аналізу**

Послідовність етапів реалізації алгоритму покрокового включення *(Forward stepwise)* подано на рис. 9.52.

Stepwise Analysis - Step 0
Number of variables in the model: 0
Wilks' Lambda: 1,000000

Stepwise Analysis - Step 1
Number of variables in the model: 1
Last variable entered: X1 F (1,16) = 90,59234 p < ,0000
Wilks' Lambda: ,1420558 approx. F (1,15) = 90,59234 p < ,0000

Stepwise Analysis - Step 2
Number of variables in the model: 2
Last variable entered: X2 F (1,15) = 1,488011 p < ,2414
Wilks' Lambda: ,1284077 approx. F (2,14) = 47,51384 p < ,0000

Stepwise Analysis - Step 3
Number of variables in the model: 3
Last variable entered: X5 F (1,14) = 1,528165 p < ,2367
Wilks' Lambda: ,1149010 approx. F (3,13) = 33,38029 p < ,0000

Stepwise Analysis - Step 3 (Final Step)
Number of variables in the model: 3
Last variable entered: X5 F (1,13) = ,5423793 p < ,4745
Wilks' Lambda: ,1149010 approx. F (3,13) = 33,38029 p < ,0000

Рис. 9.52. **Реалізація моделі покрокового включення змінних**

Опції аналізу змінних, які ввійшли в модель *(Variables in the model)* і виключені з моделі *(Variables not in the model)*, показані на рис. 9.53. Результати аналізу змінних за методом покрокового включення змінних подані на рис. 9.54, 9.55.

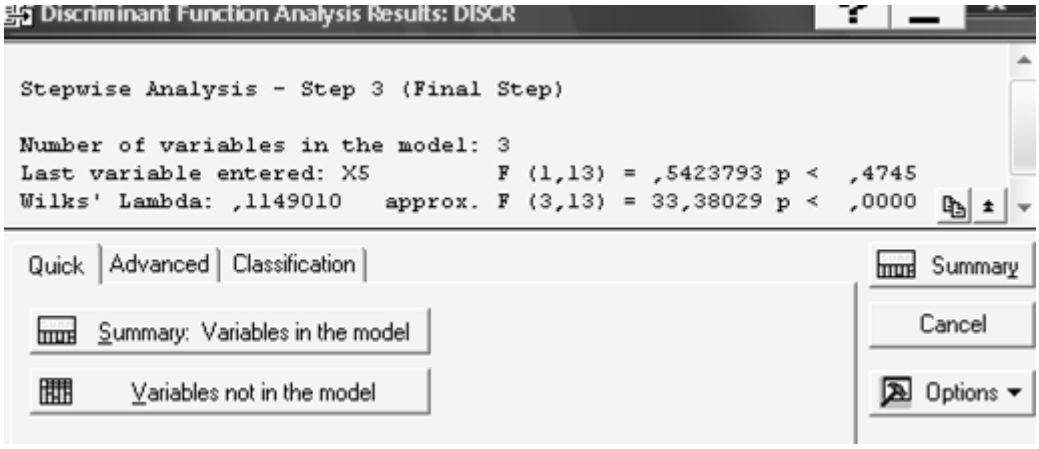


Рис. 9.53. **Опції аналізу змінних**

N=17	Discriminant Function Analysis Summary (Spreadsheet1) Step 3, N of vars in model: 3; Grouping: Class (2 grps) Wilks' Lambda: ,11490 approx. F (3,13)=33,380 p< ,0000					
	Wilks' Lambda	Partial Lambda	F-remove (1,13)	p-level	Toler.	1-Toler. (R-Sqr.)
X1	0,471135	0,243881	40,30468	0,000025	0,922465	0,077535
X2	0,133311	0,861902	2,08293	0,172616	0,852662	0,147338
X5	0,128408	0,894814	1,52816	0,238259	0,911070	0,088930

Рис. 9.54. **Аналіз змінних у моделі**

N=17	Variables currently not in the model (Spreadsheet1) Df for all F-tests: 1,12					
	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr.)
X3	0,10993	0,95675	0,54237	0,47560	0,85375	0,14624
X4	0,11318	0,98507	0,18182	0,67735	0,97559	0,02440

Рис. 9.55. **Аналіз змінних, виключених з моделі**

Аналіз змінних у моделі методом покрокового виключення подано на рис. 9.56.

N=17	Discriminant Function Analysis Summary (DISCR) Step 4, N of vars in model: 1; Grouping: Class (2 grps) Wilks' Lambda: ,14206 approx. F (1,15)=90,592 p< ,0000					
	Wilks' Lambda	Partial Lambda	F-remove (1,15)	p-level	Toler.	1-Toler. (R-Sqr.)
X1	1,00000	0,14205	90,5923	0,00000	1,00000	0,00

Рис. 9.56. **Аналіз змінних у моделі методом покрокового виключення**

Для розпізнавання нових об'єктів (підприємств) на основі відомих значень коефіцієнтів необхідно в початкову таблицю вхідних даних додати нові спостереження (рис. 9.57).

16	6,7	0,79	0,79	0,39	1,24	B
17	9,42	0,7	0,7	0,72	2,03	A
18	12,11	0,68	1,44	0,43	1,37	
19	5,49	0,74	1,1	0,05	1,02	

Рис. 9.57. **Вхідні дані для розпізнавання**

Для визначення приналежності нових об'єктів до виділених класів необхідно скористатися такими опціями: *Classification of cases (Класифікація спостережень)*; *Squared Mahalanobis distances (Квадрати відстаней Махаланобіса)*; *Posterior probabilities (Апостеріорні ймовірності)*. Результати розпізнавання наведено на рис. 9.58.

| | Classification of Cases (Spread
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | 1 p=,52941 | 2 p=,47059 |
| 1 | A | A | B |
| 2 | A | A | B |
| 3 | A | A | B |
| 4 | A | A | B |
| 5 | A | A | B |
| 6 | A | A | B |
| 7 | B | B | A |
| 8 | B | B | A |
| 9 | B | B | A |
| 10 | B | B | A |
| 11 | B | B | A |
| 12 | A | A | B |
| 13 | B | B | A |
| 14 | B | B | A |
| 15 | A | A | B |
| 16 | B | B | A |
| 17 | A | A | B |
| 18 | --- | A | B |
| 19 | --- | B | A |

| | Squared Mahalanobis Distanc
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | A p=,52941 | B p=,47059 |
| 1 | A | 3,7782 | 19,9943 |
| 2 | A | 3,1758 | 22,5451 |
| 3 | A | 4,1865 | 51,0808 |
| 4 | A | 6,5368 | 44,5157 |
| 5 | A | 1,0908 | 29,3406 |
| 6 | A | 1,5051 | 27,2026 |
| 7 | B | 23,5515 | 1,8762 |
| 8 | B | 55,9241 | 7,6365 |
| 9 | B | 16,8742 | 3,5854 |
| 10 | B | 21,1234 | 1,3401 |
| 11 | B | 47,1895 | 8,1167 |
| 12 | A | 2,8158 | 27,8854 |
| 13 | B | 28,4881 | 1,6361 |
| 14 | B | 41,5080 | 5,7259 |
| 15 | A | 3,7939 | 42,9479 |
| 16 | B | 35,0980 | 7,7022 |
| 17 | A | 10,4973 | 33,0227 |
| 18 | --- | 17,2723 | 78,5461 |
| 19 | --- | 44,9851 | 9,3938 |

| | Posterior Probabilities (Spreads
Incorrect classifications are ma | | |
|---|---|---|---|
| Case | Observed Classif. | A p=,52941 | B p=,47059 |
| 1 | A | 0,99973 | 0,00026 |
| 2 | A | 0,99994 | 0,00005 |
| 3 | A | 1,00000 | 0,00000 |
| 4 | A | 1,00000 | 0,00000 |
| 5 | A | 0,99999 | 0,00000 |
| 6 | A | 0,99999 | 0,00000 |
| 7 | B | 0,00002 | 0,99997 |
| 8 | B | 0,00000 | 1,00000 |
| 9 | B | 0,00146 | 0,99853 |
| 10 | B | 0,00005 | 0,99994 |
| 11 | B | 0,00000 | 1,00000 |
| 12 | A | 0,99999 | 0,00000 |
| 13 | B | 0,00000 | 0,99999 |
| 14 | B | 0,00000 | 1,00000 |
| 15 | A | 1,00000 | 0,00000 |
| 16 | B | 0,00000 | 0,99999 |
| 17 | A | 0,99998 | 0,00001 |
| 18 | --- | 1,00000 | 0,00000 |
| 19 | --- | 0,00000 | 1,00000 |

Рис. 9.58. **Результати розпізнавання**

Таким чином, можна зробити висновок, що підприємство № 18 з зазначеними характеристиками належить до класу А (інвестиційно привабливі підприємства), а підприємство № 19 – до класу В (підприємства-аутсайдери).

Діаграму розсіву об'єктів у просторі канонічних коренів можна побудувати ініціюванням клавіші *Scatterplot of canonical scores* в опції аналізу *Perform canonical analysis (Канонічний аналіз)* на вкладці *Canonical scores* (рис. 9.59). Діаграма розсіву подана на рис. 9.60.

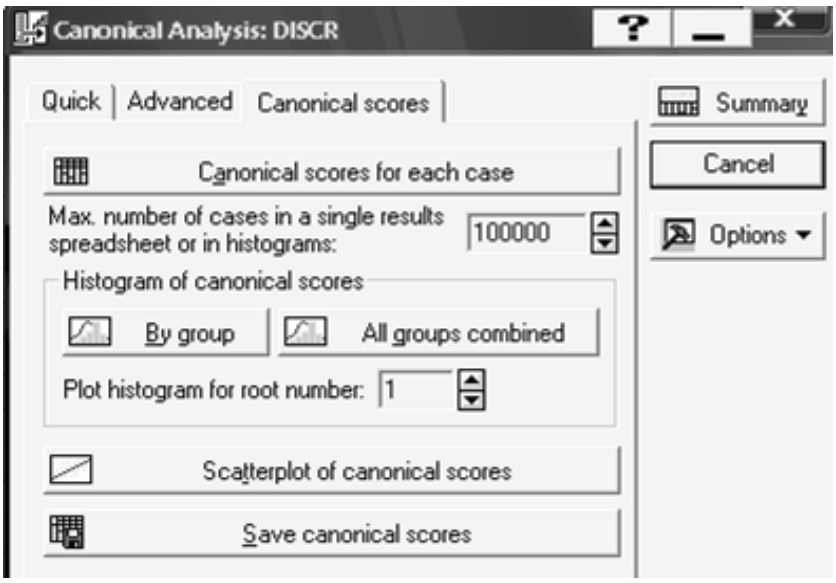


Рис. 9.59. **Опції побудови діаграми**

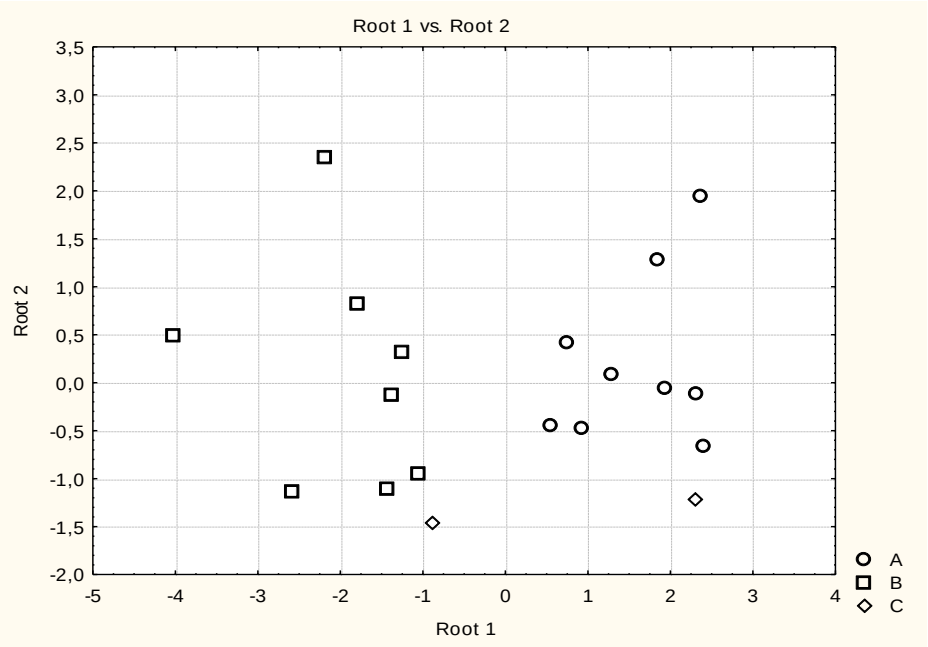


Рис. 9.60. **Діаграма розсіву об'єктів у просторі канонічних коренів**

Таким чином, результати діаграми підтверджують якість проведеної класифікації (виділені класи в просторі не перетинаються і мають значно відмінні значення канонічних коренів, які визначаються за дискримінантними функціями). Два нові підприємства, які було задано як клас С, за результатами діаграми належать, відповідно, до класу А та класу В.

Лабораторна робота 4. Методи редукції

Мета – закріплення теоретичного та практичного матеріалу за темою "Методи редукції"; набуття придбання навичок вибору показників-репрезентантів і побудови таксономічного показника рівня розвитку в середовищі *Microsoft Excel*.

Завдання 1. Необхідно вибрати найбільш значущі для оцінювання стабільності фінансового стану досліджуваних підприємств показники (показники-репрезентанти) за допомогою методу "центра ваги", дати економічну інтерпретацію отриманим результатам:

1) стандартизувати вихідні дані;

2) розрахувати матриці евклідових відстаней для кожної групи фінансових показників;

3) вибрати на основі суми відстаней показник-репрезентант у кожній групі фінансових показників, дати економічну інтерпретацію отриманим результатам.

Завдання 2. Необхідно здійснити впорядкування підприємств за стабільністю фінансового стану на основі методу рівня розвитку, дати економічну інтерпретацію отриманим результатам:

1) на основі стандартизованих вихідних даних визначити еталон розвитку для досліджуваних підприємств;

2) розрахувати евклідові відстані між окремими підприємствами та еталоном;

3) розрахувати таксономічний показник рівня розвитку на основі евклідових відстаней для кожного досліджуваного підприємства;

4) здійснити впорядкування підприємств за стабільністю фінансового стану на основі таксономічного показника рівня розвитку, дати економічну інтерпретацію отриманим результатам.

Література: [5 – 9; 14; 41 – 44; 48; 49; 76].

Методичні рекомендації

Завдання 1. У багатьох економічних дослідженнях виникає потреба в зменшенні числа ознак, що описують досліджувану область дійсності. Однак зі скороченням числа змінних повинні дотримуватися деякі вимо-

ги, для того щоб створений опис не спотворював дійсності. Існує досить великий спектр методів багатовимірного аналізу, що дозволяє розв'язувати завдання зі скорочення розмірності простору ознак. Такі методи підрозділяють на дві групи: методи побудови узагальнених показників, методи зменшення числа ознак.

Перша група методів спрямована на обчислення інтегральної оцінки об'єктів, що мають багатоознакову природу, у вигляді деякої функції $f(y_1, y_2, \dots y_q)$, що відбиває вплив усіх ознак. Такий спосіб дозволяє впорядкувати досліджувані об'єкти.

Сутність роботи другої групи методів полягає в заміні первісного набору q ознак $y = (y_1, y_2, \dots y_q)$ набором s діагностичних ознак $x = (x_1, x_2, \dots x_m)$ $(s < q)$, які мають такі властивості: ознаки не корельовані або слабо корельовані між собою; сильно корельовані з ознаками, що не входять у діагностичний набір. Таким чином, друга група методів дозволяє виключити з первісної системи ознак ті, які дублюють інформацію, а також забезпечує вибір ознак, що найбільш повно відображають стан досліджуваних процесів.

Одним з методів другої групи є метод "центра ваги". Розглянемо порядок розрахункових процедур для вибору показників-репрезентантів на основі алгоритму даного методу в середовищі *Microsoft Excel*.

Алгоритм методу "центру ваги" містить такі основні кроки.

Крок 1. На першому кроці алгоритму формуються матриці вхідних даних за кожною групою показників стану об'єкта дослідження $y_1, y_2, \dots y_q$, де q – кількість груп показників. Для k-ї групи показників структура цієї матриці може бути визначена таким чином: $y_k = (y_{ij}), i == [1; m], [1; n]$, де y_{ij} – значення і-го показника в j-му досліджуваному періоді; m – кількість показників, що входять у k-ту групу; n – кількість досліджуваних періодів.

Вхідні дані для вибору показників-репрезентантів наведені на рис. 9.61: x_1 – коефіцієнт поточної ліквідності, x_2 – коефіцієнт швидкої ліквідності, x_3 – коефіцієнт абсолютної ліквідності, x_4 – коефіцієнт забезпеченості власними оборотними коштами, x_5 – коефіцієнт маневреності власного капіталу, x_6 – коефіцієнт заборгованості, x_7 – коефіцієнт оборотності активів, x_8 – коефіцієнт оборотності оборотних коштів, x_9 – коефіцієнт оборотності запасів, x_{10} – коефіцієнт рентабельності активів, x_{11} – коефіцієнт рентабельності власного капіталу.

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
2	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
3	1	2,713	0,984	0,001	0,855	0,024	0,93	0,052	0,287	0,998	0,005	0,006
4	2	3,079	1,006	0,002	0,979	0,026	0,926	0,032	0,189	0,837	0,023	0,025
5	3	3,553	1,091	0,02	1,174	0,024	0,634	0,024	0,158	0,735	0,058	0,064
6	4	2,304	1,244	0,015	0,885	0,002	0,885	0,164	0,897	1,775	0,044	0,05
7	5	1,572	0,877	0,002	0,46	0,022	0,91	0,087	0,447	1,184	0,084	0,093
8	6	1,659	0,751	0,001	0,41	0,032	0,912	0,043	0,204	0,91	0,075	0,082

Рис. 9.61. **Вихідні дані**

Крок 2. Оскільки показники можна виразити в абсолютних і відносних величинах, а також мати різні одиниці вимірювання, то на другому кроці здійснюється процедура їхньої стандартизації за формулою:

$$z_{ij} = \frac{y_{ij} - \bar{y}_i}{s_i},$$

де z_{ij} – стандартизоване значення і-го показника в j-му досліджуваному періоді;

$\bar{y}_i$ – середнє арифметичне значення і-го показника;

s_i – стандартне відхилення і-го показника.

Для стандартизації вихідних даних спочатку необхідно розрахувати середнє арифметичне значення і-го показника та стандартне відхилення і-го показника в середовищі *Microsoft Excel* за допомогою функцій СРЗНАЧ і СТАНДОТКЛОН, як це показано на рис. 9.62.

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
2	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
3	1	2,713	0,984	0,001	0,855	0,024	0,93	0,052	0,287	0,998	0,005	0,006
4	2	3,079	1,006	0,002	0,979	0,026	0,926	0,032	0,189	0,837	0,023	0,025
5	3	3,553	1,091	0,02	1,174	0,024	0,634	0,024	0,158	0,735	0,058	0,064
6	4	2,304	1,244	0,015	0,885	0,002	0,885	0,164	0,897	1,775	0,044	0,05
7	5	1,572	0,877	0,002	0,46	0,022	0,91	0,087	0,447	1,184	0,084	0,093
8	6	1,659	0,751	0,001	0,41	0,032	0,912	0,043	0,204	0,91	0,075	0,082
9	середнє значення	=СРЗНАЧ(B3:B8)										
10	стандартне відхилення	=СТАНДОТКЛОН(B3:B8)										

Рис. 9.62. **Розрахунок середнього арифметичного та стандартного відхилення значення і-го показника**

Результати розрахунку показників наведені на рис. 9.63.

	A	B	C	D	E	F	G	H	I	J	K	L
1	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
2	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
3	1	2,713	0,984	0,001	0,855	0,024	0,93	0,052	0,287	0,998	0,005	0,006
4	2	3,079	1,006	0,002	0,979	0,026	0,926	0,032	0,189	0,837	0,023	0,025
5	3	3,553	1,091	0,02	1,174	0,024	0,634	0,024	0,158	0,735	0,058	0,064
6	4	2,304	1,244	0,015	0,885	0,002	0,885	0,164	0,897	1,775	0,044	0,05
7	5	1,572	0,877	0,002	0,46	0,022	0,91	0,087	0,447	1,184	0,084	0,093
8	6	1,659	0,751	0,001	0,41	0,032	0,912	0,043	0,204	0,91	0,075	0,082
9	**середнє значення**	**2,4800**	**0,9922**	**0,0068**	**0,7938**	**0,0217**	**0,8662**	**0,0670**	**0,3637**	**1,0732**	**0,0482**	**0,0533**
10	**стандартне відхилення**	**0,7866**	**0,1702**	**0,0084**	**0,2999**	**0,0102**	**0,1148**	**0,0523**	**0,2813**	**0,3761**	**0,0304**	**0,0334**

Рис. 9.63. **Результати розрахунку середнього арифметичного та стандартного відхилення значення і-го показника**

Для стандартизації вихідних даних необхідно створити таблицю такої ж розмірності, як таблиця вхідних даних. Стандартизацію вихідних даних в середовищі *Microsoft Excel* можна здійснити за допомогою функції НОРМАЛИЗАЦИЯ, аргументами якої є вихідні дані, середнє значення та стандартне відхилення показників (рис. 9.64).

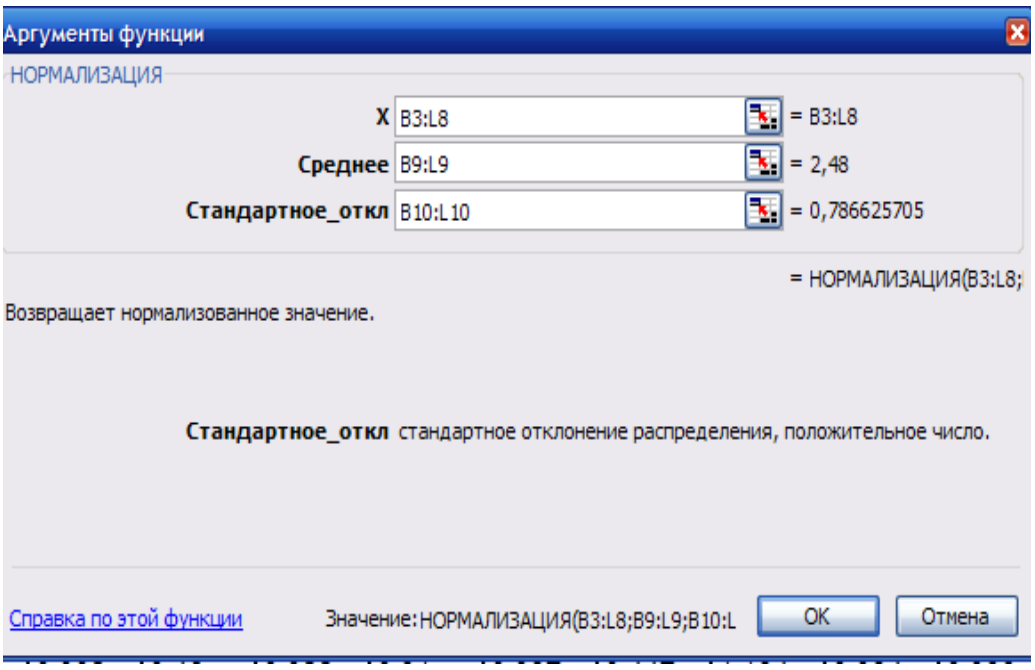


Рис. 9.64. **Аргументи функції НОРМАЛИЗАЦИЯ**

Для стандартизації вихідних даних необхідно в першій комірці створеної таблиці задати аргументи функції НОРМАЛИЗАЦИЯ; потім, починаючи з першої комірки, виділити діапазон, в якому будуть розраховані стандартизовані дані; натиснути клавішу F2, а потім одночасно натиснути клавіші Ctrl + Shift + Enter (рис. 9.65).

12	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
13	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
14	1	=НОРМАЛИЗАЦИЯ(B3:L8;B9:L9;B10:L10)										
15	2											
16	3											
17	4											
18	5											
19	6											
20												
21												

Рис. 9.65. **Розрахунок стандартизованих значень показників**

Результати стандартизації наведені на рис. 9.66.

12	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
13	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
14	1	0,2962	-0,0480	-0,6925	0,2040	0,2281	0,5559	-0,2867	-0,2726	-0,1999	-1,4209	-1,4190
15	2	0,7615	0,0813	-0,5737	0,6175	0,4236	0,5210	-0,6690	-0,6210	-0,6279	-0,8284	-0,8494
16	3	1,3641	0,5806	1,5630	1,2677	0,2281	-2,0218	-0,8220	-0,7312	-0,8991	0,3237	0,3198
17	4	-0,2237	1,4795	0,9694	0,3040	-1,9223	0,1640	1,8542	1,8962	1,8661	-0,1371	-0,0999
18	5	-1,1543	-0,6766	-0,5737	-1,1132	0,0326	0,3817	0,3823	0,2963	0,2947	1,1795	1,1892
19	6	-1,0437	-1,4168	-0,6925	-1,2800	1,0100	0,3991	-0,4588	-0,5677	-0,4338	0,8832	0,8594
20												

Рис. 9.66. **Таблиця стандартизованих значень показників**

Крок 3. Описані обчислювальні процедури є основою для розрахунку матриць відстаней $p_1, p_2, \ldots p_q$, елементи яких відбивають ступінь близькості показників усередині кожної групи. Як міра відстані використовується евклідова відстань, що визначається за формулою:

$$\rho_E(z_i, z_j) = \sqrt{\sum_{l=1}^{n} (z_{il} - z_{jl})^2},$$

де $\rho_E(z_i, z_j)$ – відстань між і-м і j-м показником групи;

z_{il}, z_{jl} – стандартизовані значення і-го та j-го показників групи в періоді l.

Матриці відстаней розраховуються для груп з числом показників більше двох. Це групи показників ліквідності, фінансової стабільності та ділової активності. Розрахунок відстаней для показників ліквідності наведено на рис. 9.67.

Матриця відстаней показників є симетричною з нульовими діагональними елементами (оскільки відстань від показника до самого себе дорівнює нулю).

21	№ підприємства/	Показники ліквідності			Розрахунок відстаней		
22	група показників	X1	X2	X3	(x1-x2)^2	(x1-x3)^2	(x2-x3)^2
23	1	0,2962	-0,0480	-0,6925	0,11846	0,97744	0,415347
24	2	0,7615	0,0813	-0,5737	0,46269	1,78283	0,429044
25	3	1,3641	0,5806	1,5630	0,61376	0,03956	0,964982
26	4	-0,2237	1,4795	0,9694	2,90094	1,42366	0,260144
27	5	-1,1543	-0,6766	-0,5737	0,22821	0,33704	0,010576
28	6	-1,0437	-1,4168	-0,6925	0,13921	0,12337	0,524696
29				**Сума**	**4,46327**	**4,68391**	**2,604787**
30				**Відстань**	**2,11265**	**2,16423**	=КОРЕНЬ(G29)
31							КОРЕНЬ(число)

Рис. 9.67. **Розрахунок відстаней для показників ліквідності**

Матриця відстаней для групи показників ліквідності наведена на рис. 9.68.

32		Показники ліквідності		
33		X1	X2	X3
34	X1	0	2,1126	2,1642
35	X2	2,1126	0	1,6139
36	X3	2,1642	1,6139	0

Рис. 9.68. **Матриця відстаней для групи показників ліквідності**

Аналогічно розраховуються матриці відстаней для груп показників фінансової стабільності та ділової активності (рис. 9.69).

	Показники фінансової стабільності		
	X4	X5	X6
X4	0	3,5542	4,0035
X5	3,5542	0	3,1665
X6	4,0035	3,1665	0

	Показники ділової активності		
	X7	X8	X9
X7	0	0,1782	0,1537
X8	0,178229	0	0,2288
X9	0,153727	0,2288	0

Рис. 9.69. **Матриці відстаней для груп показників фінансової стабільності та ділової активності**

Крок 4. На четвертому кроці здійснюється вибір так званих показників-репрезентантів груп, які містять найбільш значущу інформацію, властиву групі за правилами:

у групах з одного елемента показники, що їх утворюють, мають властивості, які сильно відрізняються від показників інших груп; тому вони належать до числа показників-еталонів (репрезентантів);

у групах, де число показників більше двох, розраховується сума відстаней кожного показника до інших показників групи:

$$\rho_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} \rho(z_i, z_j),$$

де m – кількість показників групи.

До складу показників-репрезентантів належить показник з найменшою сумою відстаней: $\rho_s = \min_i \rho_i$.

У групах, які складаються з двох показників, спочатку обчислюють суму відстаней показників, що належать до групи, від показників-репрезентантів, вибраних за описаними правилами: $\rho_i = \sum_{\substack{j=1 \\ j \neq i}}^{k} \rho(z_i, z_j)$, де k – кількість груп показників. Репрезентантом є той показник, у якого сума відстаней до виокремлених репрезентантів інших груп елементів є максимальною: $\rho_s = max_i \rho_i$.

Таким чином, результатом четвертого кроку є набір показників-репрезентантів $x = (x_1, x_2, \dots x_k)$, що описують найбільш важливі аспекти стану об'єкта дослідження.

У таблиці вхідних даних відсутні групи з одним елементом, отже пропустимо цей крок і розглянемо групи з числом показників більше двох. Розрахуємо суми відстаней та виберемо показники-репрезентанти у цих групах на основі найменшої суми відстаней (рис. 9.70).

Отже, як показник-репрезентант у групі показників ліквідності був виділений показник швидкої ліквідності (x_2); у групі показників фінансової стабільності – коефіцієнт маневреності власного капіталу (x_5); у групі показників ділової активності – коефіцієнт оборотності активів (x_7).

	Показники ліквідності			Сума
	X1	X2	X3	
X1	0	2,1126	2,1642	4,27688
X2	2,1126	0	1,6139	3,72658
X3	2,1642	1,6139	0	3,77817

	Показники фінансової стабільності			Сума
	X4	X5	X6	
X4	0	3,5542	4,0035	7,55771
X5	3,5542	0	3,1665	6,72073
X6	4,0035	3,1665	0	7,16997

	Показники ділової активності			Сума
	X7	X8	X9	
X7	0	0,1782	0,1537	0,33196
X8	0,1782	0	0,2288	0,40704
X9	0,1537	0,2288	0	0,38254

Рис. 9.70. **Вибір показників-репрезентантів груп з кількістю показників більше двох**

Оскільки група показників рентабельності включає тільки два показники, то для вибору показника-репрезентанта цієї групи знайдемо відстані кожного показника до раніше виділених показників-репрезентантів (рис. 9.71). Як показник-репрезентант вибирається той, у якого сума відстаней від показників-репрезентантів, виділених із груп елементів із числом більше двох, максимальна.

	Показники рентабельності			Сума
	X2	X5	X7	
X10	3,7586	2,9688	3,0043	9,73175
X11	3,7377	3,0039	2,9707	9,71233

Рис. 9.71. **Вибір показника-репрезентанта групи, де кількість показників дорівнює двом**

Максимальною є сума відстаней від коефіцієнта рентабельності активів (x_{10}) до вибраних раніше репрезентантів. Отже, цей показник і буде репрезентантом групи показників рентабельності.

Таким чином, до показників-репрезентантів відносять: показник швидкої ліквідності (x_2), коефіцієнт маневреності власного капіталу (x_5), коефіцієнт оборотності активів (x_7), коефіцієнт рентабельності активів (x_{10}).

Завдання 2. Для зіставлення об'єктів, які характеризуються більшою кількістю ознак, найчастіше застосовують таксономічні процедури. Одним з методів дослідження багатовимірних об'єктів є таксономічний показник рівня розвитку, запропонований З. Хельвігом. Це синтетична величина, "рівнодіюча" всіх ознак, що характеризують об'єкти, яка дозволяє лінійно впорядкувати елементи досліджуваної сукупності.

Першим кроком процесу побудови таксономічного показника рівня розвитку є визначення елементів матриці спостережень:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{\omega 1} & x_{\omega 1} & \dots & x_{\omega j} & \dots & x_{\omega m} \end{pmatrix},$$

де ω – кількість досліджуваних об'єктів,

m – кількість ознак,

x_{ij} – значення j-*ї* ознаки для i-*го* об'єкта.

Оскільки ознаки, що включають в матрицю спостережень неоднорідні, то проводять стандартизацію їх значень. Матриця стандартизованих вихідних даних наведена на рис. 9.66.

Наступний крок у розглянутій процедурі полягає в диференціації ознак матриці спостережень. Усі змінні розподіляють на стимулятори та дестимулятори. Підставою розподілу ознак на дві групи є характер впливу кожного з них на рівень розвитку досліджуваних об'єктів. Ознаки, що роблять позитивний, стимуляційний вплив на рівень розвитку об'єктів, називають стимуляторами, на відміну від ознак-дестимуляторов (табл. 9.1).

Таблиця 9.1

Класифікація ознак на стимулятори та дестимулятори

Умовне позначення показника	Показники	Характер впливу на рівень стабільності фінансового стану	Група
x_1	Коефіцієнт поточної ліквідності	Позитивний	Стимулятор
x_2	Коефіцієнт швидкої ліквідності	Позитивний	Стимулятор
x_3	Коефіцієнт абсолютної ліквідності	Позитивний	Стимулятор
x_4	Коефіцієнт забезпеченості власними оборотними коштами	Позитивний	Стимулятор
x_5	Коефіцієнт маневреності власного капіталу	Позитивний	Стимулятор
x_6	Коефіцієнт заборгованості	Негативний	Дестимулятор
x_7	Коефіцієнт оборотності активів	Позитивний	Стимулятор
x_8	Коефіцієнт оборотності оборотних коштів	Позитивний	Стимулятор
x_9	Коефіцієнт оборотності запасів	Позитивний	Стимулятор
x_{10}	Коефіцієнт рентабельності активів	Позитивний	Стимулятор
x_{11}	Коефіцієнт рентабельності власного капіталу	Позитивний	Стимулятор

Розподіл ознак на стимулятори та дестимулятори є основою для побудови так званого еталона розвитку, що є точкою з координатами:

$$P_0(z_{01}z_{02}, \dots, z_{0s}, \dots, z_{0m}),$$

де $z_{0s} = \max_r z_{rs}$, якщо $s \in I$;

$z_{0s} = \min_r z_{rs}$, якщо $s \notin I$, $(s = 1, \dots, m)$, де I – множина стимуляторів;

z_{rs} – стандартизоване значення ознаки s для об'єкта r.

Визначення еталону розвитку в середовищі *Microsoft Excel* здійснюється за допомогою функцій МАКС і МИН (рис. 9.72).

12	№ підприємства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності	
13	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
14	1	0,2962	-0,0480	-0,6925	0,2040	0,2281	0,5559	-0,2867	-0,2726	-0,1999	-1,4209	-1,4190
15	2	0,7615	0,0813	-0,5737	0,6175	0,4236	0,5210	-0,6690	-0,6210	-0,6279	-0,8284	-0,8494
16	3	1,3641	0,5806	1,5630	1,2677	0,2281	-2,0218	-0,8220	-0,7312	-0,8991	0,3237	0,3198
17	4	-0,2237	1,4795	0,9694	0,3040	-1,9223	0,1640	1,8542	1,8962	1,8661	-0,1371	-0,0999
18	5	-1,1543	-0,6766	-0,5737	-1,1132	0,0326	0,3817	0,3823	0,2963	0,2947	1,1795	1,1892
19	6	-1,0437	-1,4168	-0,6925	-1,2800	1,0100	0,3991	-0,4588	-0,5677	-0,4338	0,8832	0,8594
20	Об'єкт-еталон (P0)	1,3641	1,4795	1,5630	1,2677	1,0100	-2,0218	1,8542	1,8962	1,8661	1,1795	1,1892
21		max	max	max	max	max	min	max	max	max	max	max

Рис. 9.72. **Визначення еталону розвитку**

Відстань між окремими точками-одиницями та точкою P_0, що є еталоном розвитку, позначається c_{i0} і розраховується на основі евклідової відстані за формулою:

$$\rho_{i0} = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{0j}\right)^2}.$$

Розрахунок відстаней між окремими підприємствами й об'єктом-еталоном наведено на рис. 9.73.

12	№ підприєм ства/	Показники ліквідності			Показники фінансової стабільності			Показники ділової активності			Показники рентабельності			
13	група показників	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11		
14	Z1	0,2962	-0,0480	-0,6925	0,2040	0,2281	0,5559	-0,2867	-0,2726	-0,1999	-1,4209	-1,4190		
15	Z2	0,7615	0,0813	-0,5737	0,6175	0,4236	0,5210	-0,6690	-0,6210	-0,6279	-0,8284	-0,8494		
16	Z3	1,3641	0,5806	1,5630	1,2677	0,2281	-2,0218	-0,8220	-0,7312	-0,8991	0,3237	0,3198		
17	Z4	-0,2237	1,4795	0,9694	0,3040	-1,9223	0,1640	1,8542	1,8962	1,8661	-0,1371	-0,0999		
18	Z5	-1,1543	-0,6766	-0,5737	-1,1132	0,0326	0,3817	0,3823	0,2963	0,2947	1,1795	1,1892		
19	Z6	-1,0437	-1,4168	-0,6925	-1,2800	1,0100	0,3991	-0,4588	-0,5677	-0,4338	0,8832	0,8594		
20	Об'єкт-еталон (Z0)	1,3641	1,4795	1,5630	1,2677	1,0100	-2,0218	1,8542	1,8962	1,8661	1,1795	1,1892		
21													Сума	Відстань
22	(Z1-Z0)^2	1,1403	2,3331	5,0869	1,1316	0,6115	6,6443	4,5835	4,7034	4,2681	6,7619	6,8026	**44,0671**	**6,6383**
23	(Z2-Z0)^2	0,3631	1,9550	4,5655	0,4228	0,3439	6,4660	6,3666	6,3360	6,2201	4,0316	4,1558	**41,2264**	**6,4208**
24	(Z3-Z0)^2	0,0000	0,8079	0,0000	0,0000	0,6115	0,0000	7,1616	6,9030	7,6464	0,7324	0,7558	**24,6187**	**4,9617**
25	(Z4-Z0)^2	2,5211	0,0000	0,3523	0,9288	8,5987	4,7777	0,0000	0,0000	0,0000	1,7335	1,6618	**20,5738**	**4,5358**
26	(Z5-Z0)^2	6,3421	4,6486	4,5655	5,6689	0,9554	5,7768	2,1664	2,5596	2,4693	0,0000	0,0000	**35,1526**	**5,9290**
27	(Z6-Z0)^2	5,7973	8,3885	5,0869	6,4907	0,0000	5,8608	5,3497	6,0704	5,2896	0,0878	0,1087	**48,5304**	=КОРЕНЬ(
28														M27)
29														

Рис. 9.73. **Розрахунок відстаней між окремими підприємствами й об'єктом-еталоном**

Результати розрахунку евклідових відстаней від кожного підприємства до еталону наведені на рис. 9.74.

	№ підприємства	Евклідова відстань до еталону (ci0)
31		
32	1	6,6383
33	2	6,4208
34	3	4,9617
35	4	4,5358
36	5	5,9290
37	6	6,9664

Рис. 9.74. **Евклідові відстані від кожного підприємства до еталону**

Отримані відстані слугують вихідними величинами, використовуваними для розрахунку показника рівня розвитку:

$$d_i^* = 1 - \frac{c_{i0}}{c_0},$$

$$c_0 = \bar{c}_0 + 2s_0;\ \bar{c}_0 = \frac{1}{w}\sum_{i=1}^{w} c_{i0};\ s_0 = \sqrt{\frac{1}{w}\sum_{i=1}^{w}(c_{i0} - \bar{c}_0)^2}.$$

Розрахунок показника рівня розвитку наведено на рис. 9.75.

	№ підприємства	Евклідова відстань до еталону (ci0)	$(c_{io} - \overline{c_o})^2$	d_i^*
30				
31				
32	1	6,6383	=(B32-$B$38)^2	=1-B32/$B$40
33	2	6,4208	=(B33-$B$38)^2	=1-B33/$B$40
34	3	4,9617	=(B34-$B$38)^2	=1-B34/$B$40
35	4	4,5358	=(B35-$B$38)^2	=1-B35/$B$40
36	5	5,929	=(B36-$B$38)^2	=1-B36/$B$40
37	6	6,9664	=(B37-$B$38)^2	=1-B37/$B$40
38	$\overline{c_0}$	=СРЗНАЧ(B32:B37)	=СРЗНАЧ(C32:C37)	
39	S0	=КОРЕНЬ(C38)		
40	C0	=B38+2*B39		
41				

Рис. 9.75. **Розрахунок таксономічного показника рівня розвитку**

Отримані результати розрахунку наведені на рис. 9.76.

	№ підприємства	Евклідова відстань до еталону (ci0)	$(c_{io}-\overline{c_o})^2$	d_i^*
31				
32	1	6,6383	0,53238	**0,1354**
33	2	6,4208	0,26226	**0,1637**
34	3	4,9617	0,89669	**0,3538**
35	4	4,5358	1,88465	**0,4092**
36	5	5,9290	0,00041	**0,2278**
37	6	6,9664	1,11875	**0,0927**
38	$\overline{c_0}$	5,9087	0,7825	
39	S0	0,8846		
40	C0	7,6779		

Рис. 9.76. **Значення таксономічного показника рівня розвитку підприємств**

Інтерпретація показника рівня розвитку є такою: чим ближче значення показника рівня розвитку до одиниці, тим на більш високому рівні розвитку перебуває об'єкт.

Упорядкуємо підприємства за рівнем розвитку (рис. 9.77).

	d_i^*	№ підприємства
43		
44	0,4092	4
45	0,3538	3
46	0,2278	5
47	0,1637	2
48	0,1354	1
49	0,0927	6
50		

Рис. 9.77. **Упорядковані підприємства за рівнем розвитку**

Таким чином, найбільш стійким є четверте підприємство, а найгірше фінансове становище характерне для шостого.

Лабораторна робота 5. Методи та моделі багатовимірного шкалювання

Мета – закріплення теоретичного та практичного матеріалу за темою "Моделі багатовимірного шкалювання"; набуття навичок роботи в модулі *Multidimensional Scaling.*

Завдання – необхідно провести шкалювання простору ознак для вибіркових даних у модулі *Multidimensional Scaling* ППП *Statistica*:

1) побудувати матричний файл зі структурою відповідно до постановки задачі;

2) оцінити параметри якості моделей для різних типів шкального простору;

3) побудувати шкальний простір ознак, вибрати остаточний варіант розбивки на шкали, зробити висновки щодо характеру шкалювання;

4) представити результати графічного аналізу дослідження якості моделі (графік простору, діаграма Шепарда, графіки монотонних перетворень);

5) зробити висновки щодо угрупування об'єктів за шкальним простором, дати інтерпретацію отриманих результатів і компонентного складу латентних шкал.

Література: [5 – 9; 14; 41 – 44; 48; 49; 76].

Методичні рекомендації

Для розв'язання задач дослідження просторової структури даних об'єктів у ППП *Statistica* існує особливий набір підпрограм, об'єднаних у групу методів *Multidimensional Scaling (Багатомірне шкалювання)*. У модулі можна отримати розрахунки за метричними та неметричними методами багатомірного шкалювання, заснованими на інформації про подібності або відмінності досліджуваних об'єктів. Розглянемо порядок роботи в даному модулі.

Щоб запустити модуль *Multidimensional Scaling*, необхідно в даному середовищі створити матричний файл, що має таку структуру:

кількість спостережень повинна дорівнювати числу змінних плюс чотири допоміжні рядки;

матриця повинна бути квадратною, імена спостережень мають збігатися з іменами змінних;

останні чотири спостереження повинні містити таку інформацію:

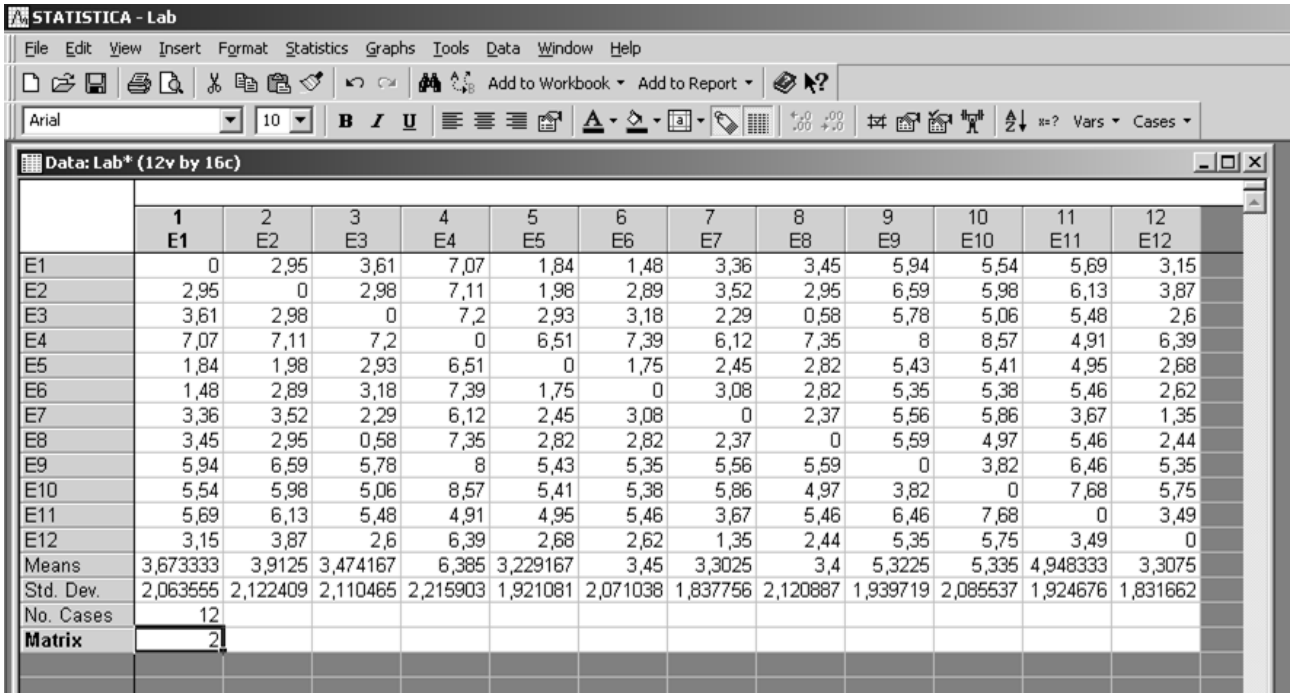
Means – для кожної змінної обчислюється середнє значення. Дана операція здійснюється за допомогою команди *Statistics of Block Data* / *Block columns* / *Means;*

Std. Dev. – для кожної змінної обчислюється середнє квадратичне відхилення за допомогою команди *Statistics of Block Data* / *Block columns* / *St's;*

No. Cases – кількість показників, на основі яких була розрахована відповідна матриця;

Matrix – даний рядок може приймати чотири значення: 1 – матриця кореляцій; 2 – матриця відстаней; 3 – матриця відмінностей; 4 – матриця коваріацій.

Отриманий матричний файл перетворюється у вигляд крупноформатної таблиці за допомогою команди *Format* / *Spreadsheets* / *System Default,* доступної шляхом запуску стартової панелі програми. Приклад перетворених вихідних даних для роботи в модулі поданий на рис. 9.78.



	1 E1	2 E2	3 E3	4 E4	5 E5	6 E6	7 E7	8 E8	9 E9	10 E10	11 E11	12 E12
E1	0	2,95	3,61	7,07	1,84	1,48	3,36	3,45	5,94	5,54	5,69	3,15
E2	2,95	0	2,98	7,11	1,98	2,89	3,52	2,95	6,59	5,98	6,13	3,87
E3	3,61	2,98	0	7,2	2,93	3,18	2,29	0,58	5,78	5,06	5,48	2,6
E4	7,07	7,11	7,2	0	6,51	7,39	6,12	7,35	8	8,57	4,91	6,39
E5	1,84	1,98	2,93	6,51	0	1,75	2,45	2,82	5,43	5,41	4,95	2,68
E6	1,48	2,89	3,18	7,39	1,75	0	3,08	2,82	5,35	5,38	5,46	2,62
E7	3,36	3,52	2,29	6,12	2,45	3,08	0	2,37	5,56	5,86	3,67	1,35
E8	3,45	2,95	0,58	7,35	2,82	2,82	2,37	0	5,59	4,97	5,46	2,44
E9	5,94	6,59	5,78	8	5,43	5,35	5,56	5,59	0	3,82	6,46	5,35
E10	5,54	5,98	5,06	8,57	5,41	5,38	5,86	4,97	3,82	0	7,68	5,75
E11	5,69	6,13	5,48	4,91	4,95	5,46	3,67	5,46	6,46	7,68	0	3,49
E12	3,15	3,87	2,6	6,39	2,68	2,62	1,35	2,44	5,35	5,75	3,49	0
Means	3,673333	3,9125	3,474167	6,385	3,229167	3,45	3,3025	3,4	5,3225	5,335	4,948333	3,3075
Std. Dev.	2,063555	2,122409	2,110465	2,215903	1,921081	2,071038	1,837756	2,120887	1,939719	2,085537	1,924676	1,831662
No. Cases	12											
Matrix	2											

Рис. 9.78. **Вихідні дані**

Щоб приступити до обчислювальних процедур, необхідно ввійти в позицію меню *Statistics* / *Multivariate Exploratory Techniques* / *Multidimensional Scaling* (рис. 9.79).

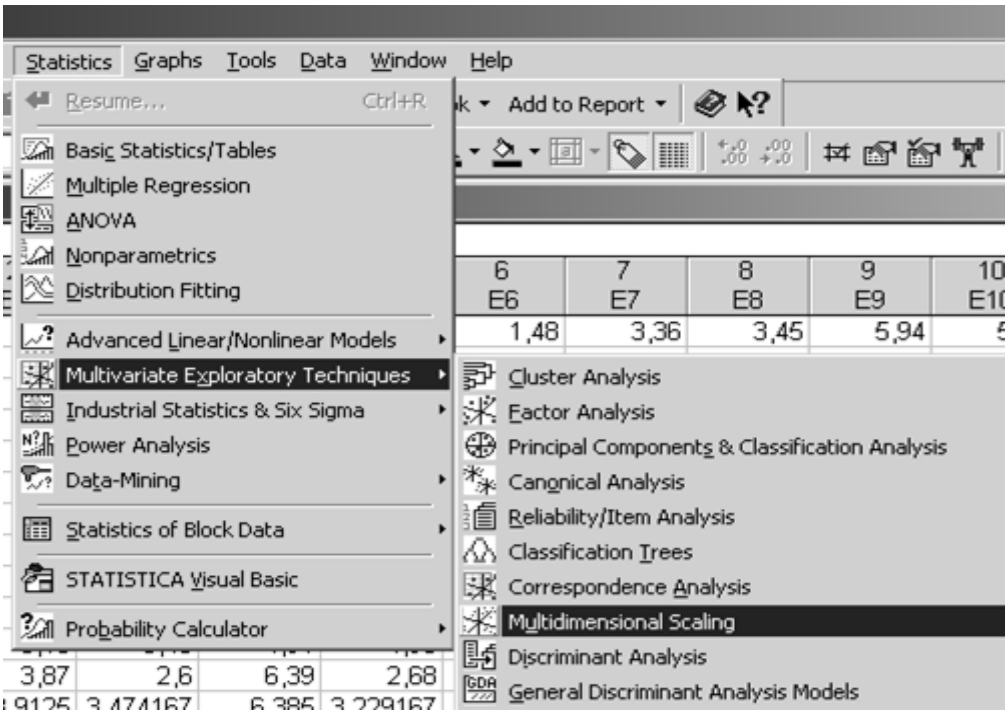


Рис. 9.79. **Вибір модуля**

Після запуску модуля з'явиться його стартова панель, де необхідно задати змінні для аналізу *(Variables)* і вибрати початковий розмір шкального простору *(Number of dimensions)* (рис. 9.80).

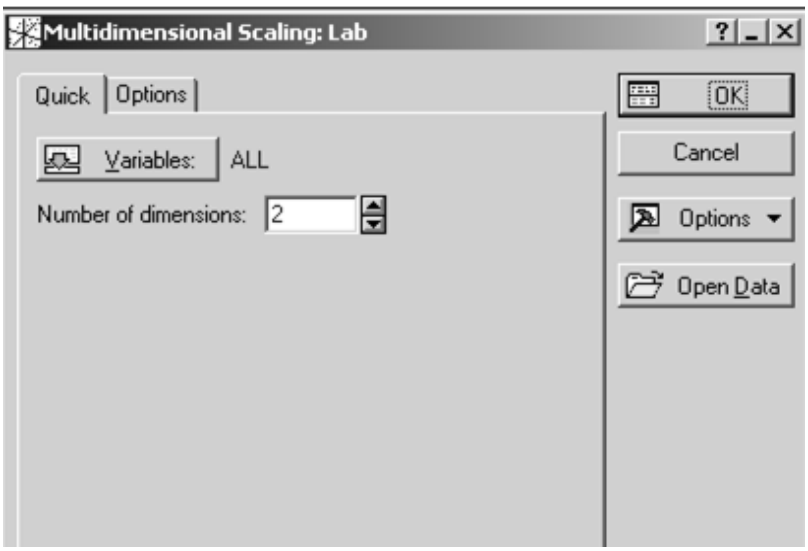


Рис. 9.80. **Стартова панель модуля**

У закладці (*Options)* задається стартова конфігурація для розрахунків. У розглянутому прикладі задана стандартна конфігурація моделі за Гутманом (рис. 9.81).

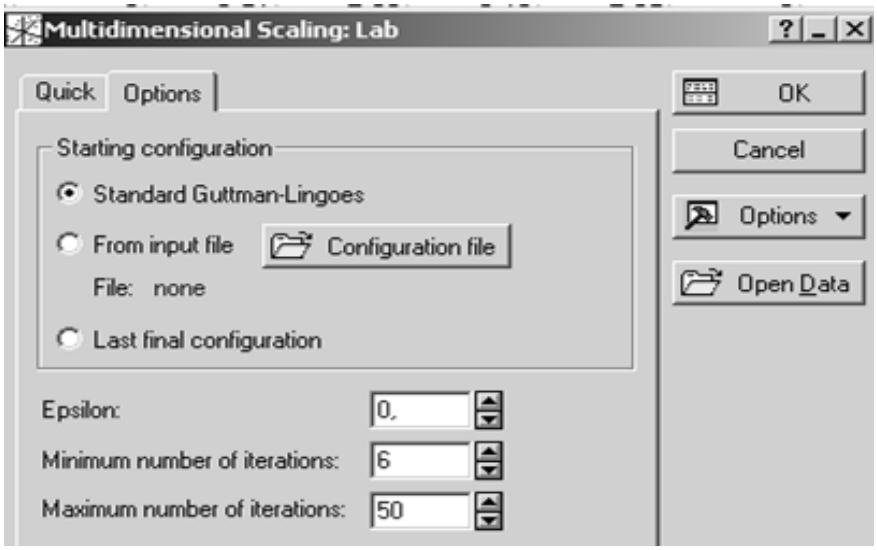


Рис. 9.81. **Стартова конфігурація**

Ініціюйте кнопку ОК, і перед вами з'явиться вікно, що містить параметри оцінювання (рис. 9.82). Оскільки визначення параметрів – ітеративна процедура, то у вікні будуть виділені найкращі значення параметрів конфігурації заданого розміру, а також подана інформація про збіжність процедури оцінювання *(Estimation procedure converged)*.

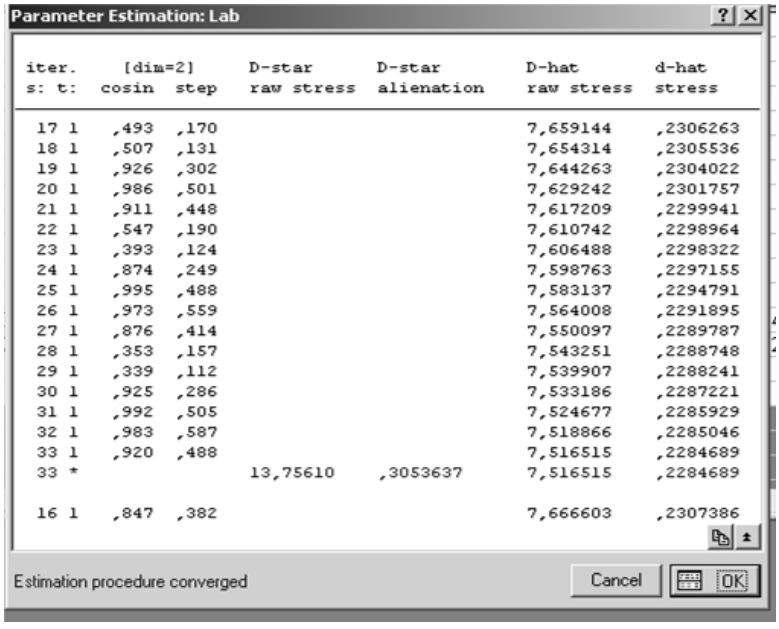


Рис. 9.82. **Вікно параметрів оцінювання**

Натисніть кнопку ОК для перегляду результатів аналізу. У верхній частині отриманого вікна міститься основна інформація про якість моделі: *Vars from file* (кількість змінних), *Number of dimensions* (кількість шкал), *Start config.* (стартова конфігурація моделі), *Last iteration* (остання ітерація), *Best iteration* (найкраща ітерація). Далі подані всі коефіцієнти для найкращої ітерації *D-star: Raw stress* (значення стресу за Гутманом), *D-hat: Raw stress* (значення стресу за Краскалом), Alienation (коефіцієнт відчуження), *Stress* (значення стресу). У нижній частині вікна знаходиться ряд опцій, кожна з яких пов'язана з певним напрямом аналізу побудованої моделі (рис. 9.83).

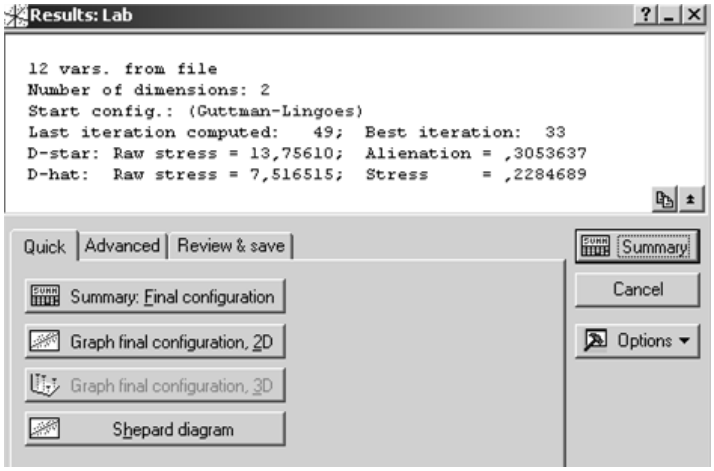


Рис. 9.83. **Вікно оцінки параметрів для двомірного простору**

Подальше дослідження моделі спрямоване на виявлення оптимального числа латентних ознак (шкал), на які необхідно розбити вихідну сукупність. Кількість шкал визначається за мінімальним значенням стресу для відповідного шкального простору, однак чим більше шкал, тим менше значення стресу. Значення показників стресу для досліджуваної моделі розміщені на рис. 9.84.

	1 D-star	2 D-hat	3 Stress
Dim. 1	48,16164	27,81277	0,439482
Dim.2	13,7561	7,516515	0,228469
Dim.3	5,876501	3,360456	0,152763
Dim.4	3,304883	1,464296	0,10084
Dim.5	1,41196	0,647873	0,067076
Dim.6	0,482888	0,167861	0,034142

Рис. 9.84. **Залежність значень стресу від розміру простору**

Для визначення оптимальної розмірності використовують графік залежності стресу від різної кількості шкал (графік "кам'янистого осипу") (рис. 9.85).

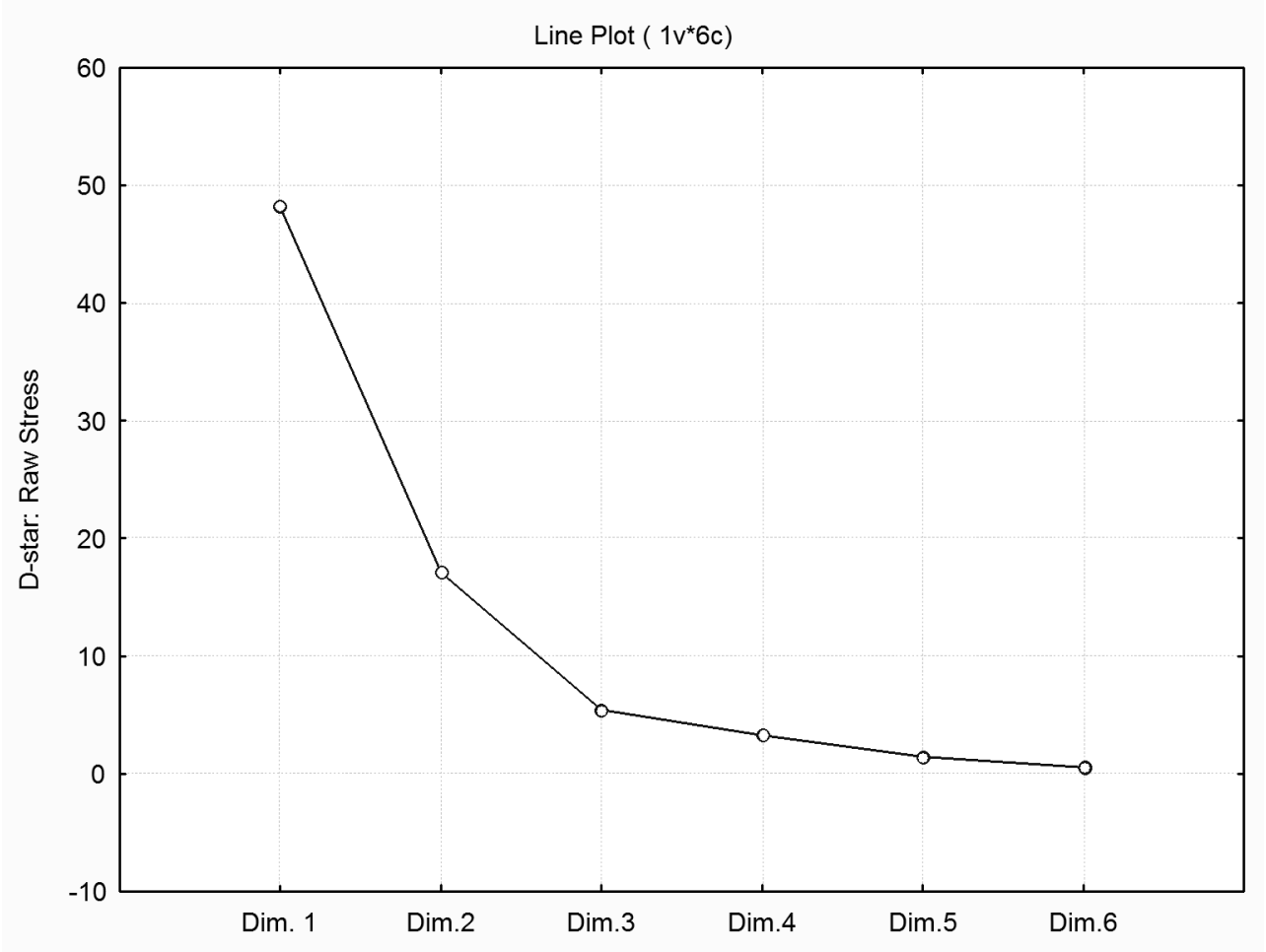


Рис. 9.85. **Графік "кам'янистого осипу"**

Висновок про розмірність можна зробити, знайшовши таку абсцису на графіку, в якій графік стресу починає візуально згладжуватися в напрямку правої, пологої його частини; таким чином, зменшення стресу максимально сповільнюється. Відповідно до даного критерію варто вибрати для відтворення моделі тривимірний простір.

Проведемо аналіз моделі тривимірного простору. Для цього у вікні початкових умов виберемо трирівневу розмірність. Після цього буде створена початкова конфігурація і відкрите вікно оцінки параметрів. Ініціювавши кнопку *Summary: Final configuration*, отримуємо матрицю координат стимулів (рис. 9.86).

	Final Configuration (Lab) D-star: Raw stress = 5,876501; Al D-hat: Raw stress = 3,360456; St		
	DIM. 1	DIM. 2	DIM. 3
E1	-0,62720	0,06683	-0,89668
E2	0,17350	-0,90185	-0,47084
E3	0,68971	-0,75511	0,60040
E4	-0,13550	-0,06111	0,16306
E5	-0,54589	0,31159	1,11518
E6	1,05004	0,42147	0,43634
E7	0,80320	0,21580	-0,90042
E8	-0,57972	1,11512	-0,09575
E9	-0,15858	-0,22622	0,36150
E10	0,04690	0,20457	-0,09131
E11	0,23466	0,41670	-0,17158
E12	-0,95114	-0,80780	-0,04989

Рис. 9.86. **Матриця координат стимулів**

Саме ці координати відображують просторовий розподіл підприємств у тривимірному просторі, графік якого можна створити, ініціювавши клавішу *Graf Final configuration, 3D* (рис. 9.87).

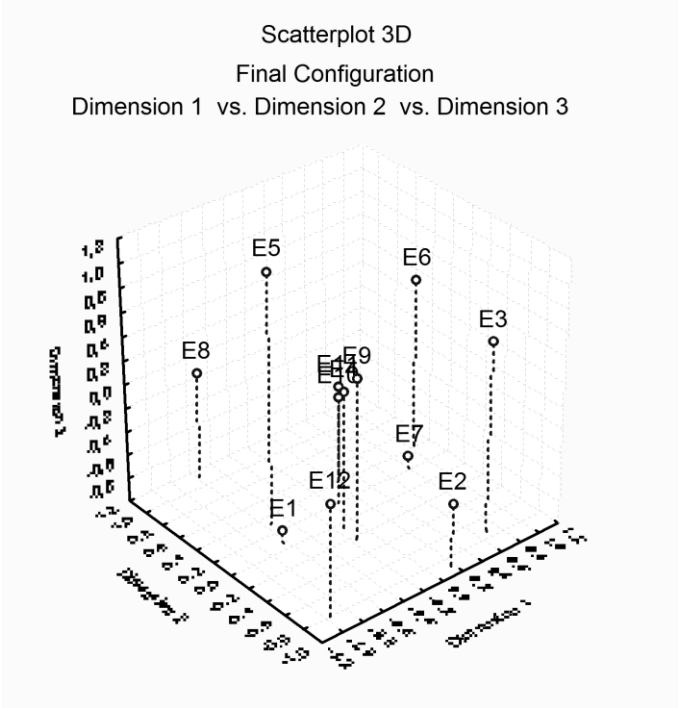


Рис. 9.87. **Просторовий розподіл підприємств**

Для оцінювання залежності відтворених відстаней від вихідних використовують діаграму Шепарда (рис. 9.88). На осі ординат діаграми позначені відтворені відстані, а на осі абсцис відкладаються істинні подібності між об'єктами. На цьому графіку також відображується графік степеневої функції, що містить величини *D-hats*, тобто результат монотонного перетворення вихідних даних за Краскалом.

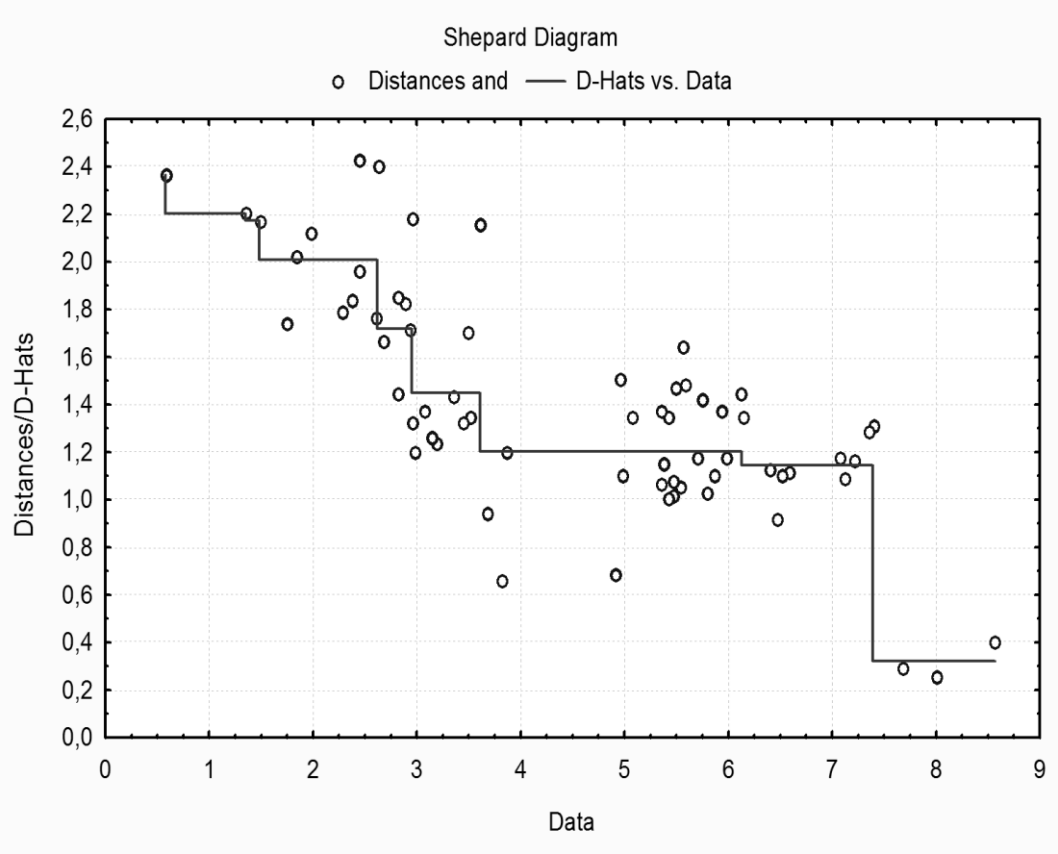


Рис. 9.88. **Діаграма Шепарда**

Якби всі відтворені результівні відстані лягли на степеневу лінію, то ранги спостережуваних відстаней були б у точності відтворені отриманим рішенням (просторовою моделлю). Відхилення від цієї лінії показують на погіршення якості згоди (тобто якості підгонки моделі). Щоб оцінити якість тривимірного рішення, необхідно проаналізувати таблицю фактичних і оцінюваних відстаней. Для цього, ініціювавши кнопку *Summary statistics,* створимо таблицю (рис. 9.89), що містить чотири стовпці. У першому стовпці зазначено, між якими об'єктами обчислена дана відстань. У другому стовпці відображені відтворені для даної конфігурації відстані.

Стовпці *D-hat* і *D-star* – монотонні перетворення вихідних даних: *D-star* відображає перетворення за Гутманом, а *D-hat* – регресію, оцінювану за Краскалом.

	Final Configuration (Lab) D-star: Raw stress = 5,876501 D-hat: Raw stress = 3,360456		
	Distance	D-star	D-hat
D(10, 4)	0,41057	0,25918	0,32140
D(9, 4)	0,25918	0,29444	0,32140
D(11,10)	0,29444	0,41057	0,32140
D(6, 4)	1,30885	0,65792	1,14467
D(8, 4)	1,28369	0,69088	1,14467
D(4, 3)	1,16357	0,92314	1,14467
D(4, 2)	1,09735	0,94594	1,14467
D(4, 1)	1,17524	1,00364	1,14467
D(9, 2)	1,12230	1,01707	1,14467
D(5, 4)	1,10175	1,02781	1,14467
D(11, 9)	0,92314	1,05924	1,14467
D(12, 4)	1,12613	1,06566	1,14467
D(11, 2)	1,35348	1,07553	1,20169
D(7, 4)	1,44528	1,09735	1,20169
D(10, 2)	1,17654	1,10175	1,20169

Рис. 9.89. **Таблиця відтворених і перетворених відстаней**

Рядки таблиці містять відстані, визначені в матриці відмінностей, відсортовані за величиною *D-star* або *D-hat*. Якщо якість даної моделі (кількість шкал) досить добра, то порядок відтворених відстаней приблизно дорівнює перетвореним вихідним даним (тобто або величині *D-star*, або *D-hat*). Наявність дуже відмінних елементів з упорядкування свідчить про недостатню точність моделі. Аналіз наведеної таблиці свідчить про задовільну якість моделі, тому що немає значних відмінностей значень відтворених і перетворених даних.

Додатково якість підгонки моделі можна проілюструвати графіками перетворення даних *D-hats* і *D-stars*, (графіки створюємо, ініціювавши кнопки *D-hat values* і *D-star values*) (рис. 9.90, 9.91).

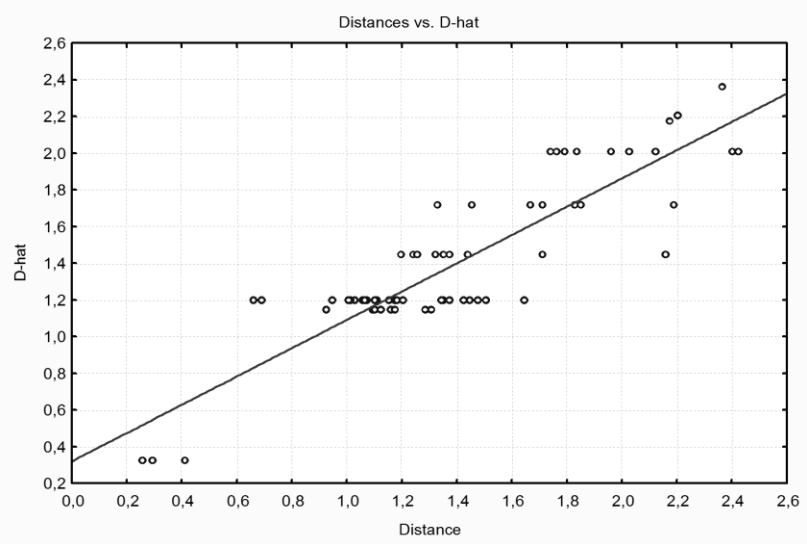


Рис. 9.90. **Графік монотонного перетворення за Краскалом**

Із графіків видно, що майже всі точки згруповані навколо ліній монотонних перетворень. Це може свідчити про те, що тривимірна конфігурація досить адекватно описує подібності між об'єктами.

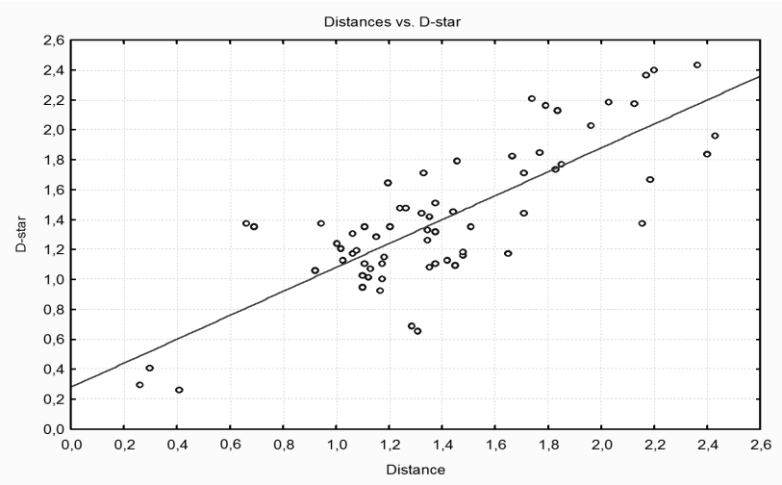


Рис. 9.91. **Графік монотонного перетворення за Гутманом**

Проведений аналіз тривимірної моделі шкального простору дає адекватну оцінку якості моделі, тобто розміщення підприємств із даними характеристиками в тривимірному просторі найбільш обґрунтоване та має логічну економічну інтерпретацію. Для визначення компонентного складу векторів необхідно застосувати факторний аналіз і визначити, які вихідні показники сформували латентні фактори, або шкали.

Глосарій

Абсолютні вимірювання – вимірювання, які засновані на прямих вимірюваннях однієї або декількох основних величин або на використанні значень фізичних констант.

Багатовимірне шкалювання – метод багатовимірного статистичного аналізу, що дозволяє подати дані в теоретичному просторі, описати процеси і явища, які через свою складність або нестабільність не піддаються моделюванню традиційними методами.

Багатовимірний дисперсійний аналіз – метод багатовимірного статистичного аналізу, що дозволяє оцінювати та досліджувати дисперсії комплексів ознак.

Багатовимірний коваріаційний аналіз – метод багатовимірного статистичного аналізу, що дозволяє оцінювати залежність варіації результативної ознаки від факторної.

Багатовимірний простір ознак – простір, в якому об'єкти подані значеннями двох і більше ознак.

Багатовимірний статистичний аналіз – сукупність формалізованих статистичних методів, які базуються на поданні вихідної інформації в багатовимірному геометричному просторі та дозволяють визначати неявні (латентні), але об'єктивно існуючі закономірності в організаційній структурі та тенденціях розвитку соціально-економічних явищ і процесів.

Багаторазове вимірювання – вимірювання фізичної величини одного розміру, результат якого отримано з кількох послідовних вимірювань, тобто воно складається з ряду однократних вимірювань.

Багатошарові персептрони – нейронні мережі прямого поширення. Вхідний сигнал у таких мережах поширюється в прямому напрямку, від шару до шару. Багатошаровий персептрон складається з таких елементів: множини вхідних вузлів, які утворюють вхідний шар; одного або декількох прихованих шарів обчислювальних нейронів; одного вихідного шару нейронів.

Вибірка – своєрідна мікромодель (проекція) всієї генеральної сукупності, яка за всіма основними досліджуваними якісними характеристиками та контрольними ознаками повинна своєю структурою максимально повторювати структуру генеральної сукупності.

Вибіркова сукупність (вибірка) – відібране за строго заданим правилом певне число елементів генеральної сукупності.

Вибіркове обстеження – дослідження об'єкта вибірки.

Вимірювання – сукупність дій, виконуваних за допомогою засобів вимірювань з метою знаходження числового значення вимірюваної величини в прийнятих одиницях вимірювання; кодування та співвідношення ступеня вираженості ознак емпіричних об'єктів або подій за допомогою чисел відповідно до певних правил.

Випадкові (статистичні) вимірювання – вимірювання, в яких вимірювана величина змінюється випадковим чином (випадковий процес).

Віддільність – ступінь перекриття кластерів і віддаленості кластерів один від одного в просторі.

Відносні вимірювання – вимірювання відношення величини до однойменної величини, що відіграє роль одиниці, або вимірювання величини за відношенням до однойменної величини, прийнятої за вихідну.

Внутрішньогрупова дисперсія – характеризує варіацію, обумовлену впливом неврахованих факторів у ході групування в кожній групі.

Впорядкованість шкали – коли одна позиція шкали, визначається числом і відповідна вираженості вимірюваної властивості, більша, менша або дорівнює іншій позиції.

Генеральна сукупність – сукупність усіх елементів, що володіють рядом загальних характеристик; охоплює повну множину елементів з точки зору вирішення певної проблеми.

Грубі помилки – дані, що різко виділяються на тлі досліджуваної сукупності спостережень.

Дендрит – ламана лінія, яка може розгалужуватися, з'єднуючи кожні дві точки досліджуваної сукупності, однак не може містити замкнутих ламаних ліній (контурів).

Дендрограма – дерево об'єднання кластерів з порядковими номерами об'єктів на горизонтальній осі та зі шкалою відстаней на вертикальній осі.

Дестимулятор – ознака, якій надає негативний вплив на досліджуваний об'єкт.

Детерміновані (регулярні) вимірювання – вимірювання, в яких вимірювана величина змінюється за певним законом (відомим).

Динамічні вимірювання – вид вимірювання за характером залежності вимірюваної величини від часу, за якого вимірювана величина змінюється і є непостійною в часі.

Дискримінантна валідність – оцінюється за здатністю кожного пункту вимірювального інструмента відображати мінливість вимірюваних характеристик сукупності.

Дискримінантна функція – лінійна комбінація змінних, отримана в результаті дискримінантного аналізу; дозволяє об'єднувати людей або об'єкти в одну або більше груп.

Дискримінантний аналіз – інструмент багатовимірного статистичного аналізу, який використовується для прийняття рішення про те, які змінні поділяють (тобто "дискримінують") набори даних, що виникають.

Дискримінантні змінні – ознаки, що дозволяють відрізняти один клас (підмножину) від іншого.

Дисперсійний аналіз – статистичний метод, що дозволяє аналізувати вплив різних факторів на досліджувану змінну.

Дисперсія кластера – ступінь розсіювання точок кластера в просторі щодо центру кластера.

Довірчий інтервал – величина, що характеризує точність оцінки вимірюваної величини.

Достовірність – характеризує довіру до результатів вимірювань і розподіляє їх на дві категорії: достовірні та недостовірні залежно від того, відомі або невідомі ймовірнісні характеристики їх відхилень від дійсних значень відповідних величин.

Експерт – суб'єкт, який з високим професіоналізмом дає судження про деякий об'єкт, явище, процес; простий спостерігач, який реєструє поведінку об'єкта.

Емпірична валідність – розглядає якість інструменту з позиції відповідності отриманих результатів якимсь стандартам.

Еталон – об'єкт, що має найбільш якісний набір ознак вихідної вибірки.

Загальна дисперсія – співвідношення між внутрішньогруповою і груповою дисперсією.

Засоби вимірювань – використовувані технічні засоби, що мають нормовані метрологічні властивості.

Змістовна валідність – показує, наскільки пункти для проведення вимірювання відповідають сутності вимірюваних характеристик.

Зовнішня валідність – характеризує сприйняття об'єктами тієї генеральної сукупності, яку передбачається дослідити.

Імовірнісні (випадкові) вибірки – вибірки, що формуються за допомогою таких підходів, які передбачають ретельне проходження алгоритму, не залишаючи місце безсистемності та "випадковості".

Імовірнісні нейронні мережі – вид нейронних мереж для задач класифікації, де щільність ймовірності приналежності до класів оцінюється за допомогою ядерної апроксимації.

Інтервальність шкали – означає, що інтервали між позиціями шкали рівні між собою.

Ітерація (лат. Iteratio – я повторюю) – в широкому сенсі: термін, що позначає повторення будь-якої дії, явища або процесу. У вузькому сенсі: термін, найчастіше застосовуваний для описання поетапного процесу, в якому результати виконання групи операцій в рамках кожного етапу використовуються наступним етапом (крім останнього, тому що він надає кінцевий результат).

Канонічна вага – вага стандартизованої вихідної змінної у відповідній канонічній змінній.

Канонічне крос-навантаження – кореляція вихідної змінної з протилежною канонічною змінною.

Канонічне навантаження – кореляція вихідної змінної з відповідною канонічною змінною.

Канонічний корінь – власні числа розв'язуваної системи рівнянь, дорівнює квадрату коефіцієнта канонічних кореляцій (один для одного рівняння канонічної кореляції).

Канонічні змінні – лінійні комбінації вихідних наборів змінних.

Класифікаційна матриця – матриця, в якій подані результати класифікації об'єктів, отримані на основі дискримінантних функцій, що дозволяє визначити частку правильно розпізнаних об'єктів і оцінити точність класифікації.

Класифікація об'єктів за допомогою функції відстані – з використанням функції відстані об'єкт належить до тієї групи, для якої відстань Махалонобіса найменша.

Кластер – клас, таксон, згущення, група, пучок – група елементів, які характеризуються будь-якими загальними властивостями.

Кластерний аналіз – таксономія, автоматична класифікація, стратифікація, класифікація без вчителя, розпізнавання з самонавчанням – сукупність обчислювальних процедур, що використовуються для класифікації (методи знаходження кластерів).

Кластерний метод – багатовимірна статистична процедура, яка виконує збирання даних, що містять інформацію про вибір об'єктів, і потім упорядковує об'єкти в порівняно однорідні групи.

Коваріаційна матриця – матриця, складена з попарних коваріацій елементів (добутків відхилень від середніх значень у групі та вибірці) двох векторів (значень об'єктів кластерів).

Коефіцієнт канонічної кореляції – міра зв'язку між двома множинами змінних. Максимальна величина цього коефіцієнта дорівнює 1. Чим більше величина коефіцієнта канонічної кореляції, тим краще розділова здатність дискримінантної функції.

Константа дискримінантної функції – межа, що розділяє дві сукупності.

Контрольно-перевірочні вимірювання – вид вимірювань, похибка яких з певною ймовірністю не повинна перевищувати деякого заданого значення.

Концептна валідність – розглядає відповідність вимірювального інструменту тому концепту (властивості), для вимірювання або оцінювання якого він був виконаний.

Критерій "лямбда Уїлкса" – критерій, що використовується для перевірки умови максимально чіткого розподілу груп.

Критерій F-включення, F-виключення – значення F-статистики для змінної вказує на її статистичну значущість при дискримінації між сумами; тобто вона є мірою вкладу змінної в передбачення приналежності об'єкта до сукупності.

Критерій відсіювання Р. Кеттела – критерій, застосовуваний для відбору числа факторів. Відповідно до даного критерію здійснюється побудова графіка власних значень. Кількість факторів визначається за точкою перегину на графіку власних значень.

Критерій згоди – варіант застосування критерію хі-квадрат, призначений для перевірки відмінностей емпіричного та теоретичного розподілів досліджуваної властивості за однією вибіркою.

Критерій Кайзера – критерій, застосовуваний для відбору числа факторів. Відповідно до даного критерію число факторів дорівнює числу компонент, власні значення яких більше одиниці.

Критерій Коаскела – Уоллі-са – непараметричний тест, який застосовується для трьох і більше незалежних вибірок з даними в порядковій або інтервальній шкалі, якщо не задовольняються припущення про використання дисперсійного аналізу.

Критерій Манна – Уїтні – непараметричний критерій, який використовується для порівняння двох незалежних вибірок.

Критерій незалежності – застосування критерію хі-квадрат для порівняння розподілів декількох незалежних вибірок.

Критерій Уїлкоксона – непараметричний критерій, заснований на обчисленні різниці вимірів у кожній і-парі елементів пов'язаних вибірок.

Критерій частки відтворюваної дисперсії – критерій, застосовуваний для відбору числа факторів. Фактори ранжуються за часткою детермінованої дисперсії; коли відсоток дисперсії виявляється несуттєвим, виділення слід зупинити. Бажано, щоб виділені фактори пояснювали більше 80 % розсіяння.

Критерії, вільні від розподілу – це критерії, форма розподілу для яких несуттєва.

Латентна змінна – змінна, значення якої в ході спостереження не доступні для безпосереднього вимірювання, а лише можуть бути оцінені відповідно до висунутої гіпотези за значеннями явних змінних.

Матриця вихідних даних – матриця X розмірністю n × m – прямокутна таблиця, кожен рядок якої є результатом вимірювання m- розглянутих ознак в одному з об'єктів.

Матриця відображення (матриця вагових коефіцієнтів) – матриця A, елементи якої є ваговими коефіцієнтами i-го головного фактора на j-ту змінну.

Матриця відстаней – матриця, елементи якої відображають ступінь близькості об'єктів, показників, ознак тощо в досліджуваній сукупності.

Мережі прямого поширення – всі зв'язки спрямовані строго від вхідних нейронів до вихідних. Прикладами таких мереж є перцептрон Розенблатта, багатошаровий перцептрон, мережі Ворда.

Метод багатовимірного групування (кластерний аналіз) – метод багатовимірного статистичного аналізу, що дозволяє провести групування багатовимірних об'єктів.

Метод вимірювань – сукупність прийомів використання принципів і засобів вимірювань.

Метод головних компонент і факторний аналіз – методи багатовимірного статистичного аналізу, що дозволяють звести множину елементарних ознак до невеликого числа значущих "узагальнених ознак" і виявити латентні фактори щодо як ознак, так і об'єктів.

Метод "далекого сусіда" – ступінь подібності оцінюється за ступенем подібності між найбільш віддаленими (несхожими) об'єктами кластерів.

Метод канонічних кореляцій – метод багатовимірного статистичного аналізу, що дозволяє стиснути дані та моделювати зв'язки узагальнених ознак.

Метод "найближчого сусіда" – ступінь подібності оцінюється за ступенем схожості між найбільш схожими (найближчими) об'єктами цих кластерів.

Метод середнього зв'язку – ступінь подібності оцінюється як середня величина ступенів подібності між об'єктами кластерів.

Методи обертання – метою є отримання зрозумілої (інтерпретованої) матриці навантажень, тобто факторів, які ясно позначені високими навантаженнями для деяких змінних і низькими – для інших. До методів обертання належать варимакс, квартимакс, еквімакс.

Метрична система заходів – загальна назва міжнародної десяткової системи одиниць, заснованої на використанні метра та кілограма.

Метричне багатовимірне шкалювання – вид багатовимірного шкалювання, що вимагає вимірювання близькості за шкалою рівня не нижче інтервального.

Міжгрупова дисперсія – характеризує варіацію, обумовлену впливом фактора, покладеного в основу угруповання. Міра варіації окремих середніх за групами щодо загальної середньої.

Міра відстані – формат відстаней, розрахованих для матриці відстаней; може бути, наприклад, евклідовою, мангетенською та ін.

Множинний кореляційно-регресійний аналіз – метод багатовимірного статистичного аналізу, що дозволяє провести вимірювання та моделювання зв'язків досліджуваних ознак або об'єктів.

Модель з випадковими факторами (Модель II) – модель, в якій кожному випадковому, обраному значенню одного фактора відповідає своя підмножина значень другого фактора.

Модель з фіксованими рівнями чинників (Модель I) – модель, в якій кожен рівень одного фактора поєднується при плануванні експерименту з кожної градацією іншого фактора.

Надійність – характеристика, що відображає стійкість і узгодженість отриманих результатів вимірювання.

Надійність-еквівалентність – характеризує ідентичність результатів, отриманих декількома аналогічними інструментами.

Надійність-стійкість – характеризує стабільність результатів у часі.

Надійність-узгодженість – характеризує узгодженість пунктів інструменту.

Неметричне багатовимірне шкалювання – вид багатовимірного шкалювання, який використовує шкали нижчого рівня (наприклад, рангова шкала).

Необхідний обсяг вибірки – функція варіації змінних параметрів генеральної сукупності та точності оцінки цих параметрів, необхідної для дослідника.

Непараметричні тести – тести, що ґрунтуються на більш слабких допущених щодо аналізованих даних. Головним мотивом застосування є небажання робити припущення, необхідне для використання параметричних процедур.

Неповна редукція – полягає в побудові деяких синтетичних величин, що мають багатоознакову природу, у вигляді деякої функції f (y1, y2, ..., yn), яка відображає вплив усіх ознак і, таким чином, дозволяє впорядкувати досліджувані об'єкти.

Непрямі вимірювання – вимірювання, за яких шукана величина безпосередньо не вимірюється; її значення знаходять на основі відомої залежності між цією величиною і величинами, отриманими в результаті прямих вимірювань.

Нерівно розсіяні вимірювання – вимірювання, випадкові величини яких розподілені неоднаково.

Нечітка кластеризація – кластеризація, за якої для кожного об'єкта визначається дійсне значення, що показує ступінь приналежності до кластера – функція приналежності (ФП). Ступінь приналежності визначається відстанню від об'єкта до відповідних кластерних центрів. Даний алгоритм ітераційно обчислює центри кластерів і нові ступені приналежності об'єктів.

Нульова точка (точка відліку) шкали – означає, що набір чисел, відповідних вираженості вимірюваної ознаки, має точку відліку, що позначається за 0, яка відповідає повній відсутності вимірюваної властивості.

Обґрунтованість або валідність вимірів – еквівалентність вимірювань характеристикам вимірюваного об'єкта (міра відповідності оцінок, отриманих у процесі вимірювання, уявленням про сутність властивостей досліджуваних об'єктів і їх ролі в досліджуваних процесах).

Одновиборочний t-критерій (t-test for single means) – параметричний тест, що дозволяє порівняти середнє значення у вибірці з заданим числом; застосовуваний для великих вибірок (100 і більше спостережень).

Одномірний простір ознак – простір, в якому об'єкти відображаються значеннями однієї ознаки.

Одноразовий вимір – вимір, виконаний один раз.

Однофакторний дисперсійний аналіз для незв'язаних вибірок – статистичний метод, який досліджує зміни результативної ознаки під впливом зміни умов або градацій будь-якого фактора.

Однофакторний дисперсійний аналіз для пов'язаних вибірок – статистичний метод, призначений для дослідження впливу різних умов дії фактора (градацій фактора) на ту саму вибірку. Визначається на підставі відомої залежності між цією величиною та величинами, що піддаються прямим вимірюванням.

Ознака в математиці та логіці – те саме, що і достатня умова. У менш строгих науках слово "ознака" вживається як опис фактів, що дозволяє (згідно з існуючою теорією і т.п.) зробити висновок про наявність явища, що цікавить.

Перевага – судження про об'єкт з точки зору його близькості до пропонованого ідеалу.

Повна редукція – полягає в отриманні так званих індивідуальних діагностичних ознак, якими є деякі з вихідних ознак. Первинний набір q ознак y = (y1, y2, ..., yq) замінюється набором s діагностичних ознак x = (x1, x2, ..., ys), (s <q).

Покроковий дискримінантний аналіз з виключенням – у цьому випадку всі змінні будуть спочатку включені в модель, а потім на кожному кроці будуть усуватися змінні, що вносять малий внесок у передбачення.

Покроковий дискримінантний аналіз з включенням – у покроковому аналізі дискримінантних функцій модель дискримінації будується покроково. Точніше, на кожному кроці проглядають усі змінні та знаходять ту з них, яка вносить найбільший внесок у відмінність між сумами. Ця змінна повинна бути включена в модель на даному етапі; відбувається перехід до наступного кроку.

Похибка вимірювань – різниця між отриманим під час вимірювання X' і істинним Q значеннями вимірюваної величини.

Правильність вимірювання – визначається як якість вимірювання, що відображає близькість до нуля систематичних похибок результатів.

Принцип вимірювань – фізичне явище або сукупність фізичних явищ, покладених в основу вимірювань.

Простір ознак з нульовою розмірністю – простір, в якому об'єкти не мають характеристик.

Прямі вимірювання – вид вимірювань за умовами, що визначають точність результату, за якого вимірювана величина порівнюється з мірою безпосередньо або ж за допомогою вимірювальних приладів, градуйованих у необхідних одиницях.

Радіально-базисні функції – штучні нейронні мережі, що використовують в якості активаційних функцій; такі мережі скорочено називають RBF-мережами.

Ранжований ряд – ряд, упорядкований за певним принципом (наприклад, за зростанням значень).

Ранжування спостережень – вибудовування спостережень за порядком від найменшого значення до найбільшого.

Результативні ознаки – залежні змінні, які зазнають впливу факторів.

Рекурентні нейронні мережі – сигнал з вихідних нейронів або нейронів прихованого шару частково передається назад, на входи нейронів вхідного шару (зворотний зв'язок). Окремим випадком рекурентних мереж є двоспрямовані мережі. У таких мережах між шарами існують зв'язки як в напрямку від вхідного шару до вихідного, так і в зворотному.

Репрезентант – представник вибірки, який передає найсуттєвіші особливості численного набору ознак вихідної вибірки.

Рівень фактора (градація) – конкретна реалізація фактора.

Рівно розсіяні вимірювання – незалежні, однаково розподілені випадкові величини.

Робастність – нечутливість до різних відхилень і неоднорідностей у вибірці.

Розмірність простору – кількість координатних осей простору.

Самоорганізована карта Кохонена – змагальна нейронна мережа з навчанням без учителя, що виконує завдання візуалізації і кластеризації.

СГС (сантиметр-грам-секунда) – система одиниць, в якій використовують три незалежні розмірності (довжина, маса та час); усі інші зводяться до них шляхом множення, ділення, піднесення до степеня.

Середньоквадратичне відхилення – в теорії ймовірностей і статистиці – найбільш поширений показник розсіювання значень випадкової величини щодо її математичного очікування.

Синапс – у нейронних мережах – зв'язок між формальними нейронами. Вихідний сигнал від нейрона надходить у синапс, який разом із ним передається до нейрона. Складні синапси можуть мати пам'ять.

Спільні вимірювання – вид вимірювань, які проводяться одночасно для вимірювання двох або декількох неоднойменних величин для знаходження залежностей між ними.

Спільність – сума відносних вкладів усіх загальних факторів у дисперсію показника уj.

Стандартизація – процес зведення показників, які мають різні одиниці вимірювання, до загального типу вимірювань.

Статистичне оцінювання багатовимірної випадкової величини – метод багатовимірного статистичного аналізу, що дозволяє визначити багатовимірну середню, матрицю коваріацій, ймовірнісні оцінки, робастне оцінювання.

Статичні вимірювання – вид вимірювання за характером залежності вимірюваної величини від часу, за якого вимірювана величина залишається постійною в часі.

Стимул – об'єкт з певним набором потрібних для дослідження характеристик. Іноді цим поняттям називають деяку типову характеристику об'єкта, яка не підлягає безпосередньому вимірюванню.

Стимулятор – ознака, яка надає позитивний вплив на досліджуваний об'єкт.

Стрес (функція стресу) – критерій відмінності матриці близькостей і матриці відстаней.

Стрес-формули – формули, що застосовують для оцінювання відповідних емпіричних і теоретичних рангових даних.

Структурні коефіцієнти – коефіцієнти кореляції між окремими змінними та дискримінантною функцією. Якщо щодо деякої змінної абсолютна величина коефіцієнта велика, то вся інформація про дискримінантну функцію представлена в цій змінній.

Сукупні вимірювання – вид вимірювань, які проводять одночасно для вимірювання кількох одноменних величин, за яких шукану величину визначають розв'язанням системи рівнянь, отримуваних шляхом прямих вимірювань різних сполучень цих величин.

Суматор – у нейронних мережах – блок, який підсумовує сигнали, що надходять від нейронів через синапси. У загальному випадку акумулятор може перетворювати сигнали та передавати їх нейронам або суматору також через синапси.

Таксономія – наука про способи впорядкування і класифікації об'єктів.

Теорія вимірювань – одна із складових частин прикладної статистики, необхідна для аналізу якісних даних, як теорія, що дає базу для розроблення, вивчення і застосування конкретних методів розрахунку.

Тест z (z-критерій) – параметричний тест, який використовують для виконання базових припущень про нормальність закону розподілу та подання в метричній шкалі.

Тест Колмогорова – Смірнова – непараметричний тест, який використовують для порівняння спостережуваного та теоретичного розподілу (нормального, рівномірного, пуассонівського, біноміального) для даних, поданих принаймні в порядковій шкалі.

Технічні вимірювання – вид вимірювань, в яких похибка результату визначається характеристиками засобів вимірювань.

Толерантність – міра лінійної залежності між однією змінною і набором інших змінних. Якщо у дискримінантному аналізі рівень толерантності становить менше 0,001, це означає, що лінійна залежність для даної змінної настільки висока, що її включення в дискримінантне рівняння неприпустимо.

Точність вимірювань – характеристика вимірювань, що відображає близькість їх результатів до істинного значення вимірюваної величини.

Фактори – незалежні впливові змінні.

Факторне навантаження – коефіцієнт кореляції, міра зв'язку змінної і головної компоненти.

Факторний аналіз – сукупність методів, які на основі реально існуючих зв'язків ознак (або об'єктів) дозволяють виявляти латентні (приховані) узагальнювальні характеристики організаційної структури та механізму розвитку досліджуваних явищ і процесів.

Форма кластера – розташування точок у просторі.

Формальний нейрон – у нейронних мережах – процесорний елемент, перетворювач даних, який отримує вхідні дані та перетворює їх відповідно до заданої функції і параметрів. Формальний нейрон працює за дискретним часом.

Функціонал або критерій якості – деяка міра якості класифікації, яка використовується для оцінювання отриманих результатів.

Функція активації (активаційна функція, функція збудження) – функція, що обчислює вихідний сигнал штучного нейрона.

Характерність – частина дисперсії, обумовлена специфікою змінної і помилками вимірювання.

Центроїди груп – середні значення дискримінантних функцій для кожної з двох і більше груп. Чим ближче об'єкт до центроїду групи, тим більша ймовірність того, що він належить до цієї групи.

Чітка (непересічна) кластеризація – кластеризація, в якій кожен об'єкт належить тільки до одного кластеру.

Шкала – теоретична вісь у просторі, яка є носієм значень узагальненої ознаки (фактора).

Шкала вимірювань – порядок визначення та позначення можливих значень конкретної величини або проявів якої-небудь властивості. Тип шкали задає групу допустимих перетворень шкали.

Штучні нейронні мережі (ШНМ) – математичні моделі, а також їх програмні або апаратні реалізації, побудовані за принципом організації та функціонування біологічних нейронних мереж – мереж нервових клітин живого організму.

Щільність – властивість кластера, що дозволяє розглядати кластер як скупчення точок у просторі даних, відносно щільне в порівнянні з іншими областями.

Предметний покажчик

Рекомендована література

1. Айвазян С. А. Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др.; под ред. С. А. Айвазяна. – Москва : Финансы и статистика, 1989. – 587с.

2. Айвазян С. А. Прикладная статистика: Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, А. Д. Мешалкин. – Москва : Финансы и статистика, 1983. – 471 с.

3. Афифи А. Статистический анализ / А. Афифи, С. Эйзен ; пер. с англ. – Москва : Мир, 1980. – 488 с.

4. Бахрушин В. Є. Методи аналізу даних : навч. посіб. для студентів / В. Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.

5. Боровиков В. П. Прогнозирование в системе *Statistica* в среде Windows / В. П. Боровиков, Г. И. Ивченко. – Москва : Финансы и статистика, 1997. – 268 с.

6. Боровиков В. П. Программа *Statistica* для студентов и инженеров / В. П. Боровиков. – 2-е изд. – Москва : Компьютер Пресс, 2001. – 301 с.

7. Боровиков В.П. *Statistica* Статистический анализ и обработка данных в среде *Windows* / В. П. Боровиков, И. П. Боровиков. – Москва : Инф.-изд. дом "Филинъ", 1997. – 608 с.

8. Буреева Н. Н. Многомерный статистический анализ с использованием ППП Statistica: учеб.-метод. материал по программе повышения квалификации "Применение программных средств в научных исследованиях и преподавании математики и механики". / Н. Н. Буреева. – Нижний Новгород : ННГУ, 2007. – 112 с.

9. Вуколов Э. А. Основы статистического анализа: практикум по статистическим методам и исследованию операций с использованием пакетов *Statistica* и *Exel* : учебн. пособ. / Э. А. Вуколов. – Москва : ФОРУМ; ИНФРА-М, 2004. – 464 с.

10. Гейман О. А. Нелинейность экономики и неравномерность развития регионов / О. А. Гейман / НАН Украины, Науч.-исслед. центр индустр. проблем развития. – Харьков : ИНЖЭК, 2009. – 427 с.

11. Гмурман В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – Москва : Высшая школа, 2003. – 523 с.

12. Голіков А. П. Економіко-математичне моделювання світо-господарських процесів / А. П. Голіков. – Київ : Знання, 2009. – 222 с.

13. Денисов В. И. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов : В 2 ч. / В. И. Денисов, Ю. Б. Лемешко, Е. Б. Цой. – Новосибирск : Новосиб. гос. техн. ун-т, 1993. – 346 с.

14. Дубина И. Н. Математико-статистические методы в эмпирических социально-экономических исследованиях : учеб. пособ. / И. Н. Дубина. – Москва : Финансы и статистика; ИНФРА-М, 2010. – 416 с.

15. Дубров А. М. Многомерные статистические методы : учебник / А . М. Дубров, В. С. Мхитарян, Л. И. Трошин. – Москва : Финансы и статистика, 2003. – 352 с.

16. Дэйвид Г. Порядковые статистики / Г. Дейвид. – Москва : Наука, 1979. – 336 с.

17. Єгоршин О. О. Методи багатовимірного статистичного аналізу : навч. посіб. / О. О. Єгоршин, А. М. Зосімов, В. С. Пономаренко. – Київ : ІЗМН, 1998. – 208 с.

18. Иберла К. Факторный анализ / К. Иберла. – Москва : Статистика, 1980. – 394 с.

19. Калинина В. Н. Введение в многомерный статистический анализ : учеб. пособ. / В. Н. Калинина, В. И. Соловьев. – Москва : ГУУ, 2003. – 66 с.

20. Кендалл М. Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт. – Москва : Наука, 1976. – 312 с.

21. Кендалл М. Статистические выводы и связи / М. Кендалл, А. Стьюарт. – Москва : Наука, 1973. – 900 с.

22. Клебанова Т. С. Механізм та моделі управління кризовими ситуаціями на підприємствах житлово-комунального комплексу / Т. С. Клебанова, М. О. Кизим, Ю. І. Мізік. – Харків : ВД "ІНЖЕК", 2011. – 178 с.

23. Клебанова Т. С. Нечітка логіка та нейронні мережі в управлінні підприємством / Т. С. Клебанова, Л. О. Чаговець, О. В. Панасенко. – Харків : ВД "ІНЖЕК", 2011. – 239 с.

24. Крамер Г. Математические методы статистики / Г. Крамер. – Москва : Мир, 1975. – 648 с.

25. Кремер Н. Ш. Теория вероятности и математическая статистика / Н. Ш. Кремер. – Москва : ЮНИТИ-ДАНА, 2002. – 343 с.

26. Куліков П. М. Економіко-математичне моделювання фінансового стану підприємства / П. М. Куліков. – Харків : ІНЖЕК, 2009. – 151 с.

27. Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам / Г. Куллдорф. – Москва : Наука, 1966. – 176 с.

28. Лемешко Б. Ю. Группирование наблюдений как способ получения робастных оценок / Б. Ю. Лемешко // Надежность и контроль качества. – 1997. – № 5. – С. 26–35.

29. Лемешко Б. Ю. К вопросу о робастности оценок по группированным данным / Б. Ю. Лемешко, С. Н. Постовалов // Сб. научных трудов НГТУ. – 1996. – № 2 (4). – С. 9–18.

30. Лемешко Б. Ю. Робастные методы оценивания и отбраковка аномальных измерений / Б. Ю. Лемешко // Заводская лаборатория. – 1997. – Т. 63. – № 5. – С. 43–49.

31. Лемешко Б. Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система / Б. Ю. Лемешко. – Новосибирск : Изд-во НГТУ, 1995. – 125 с.

32. Математичні методи в сучасних економічних дослідженнях / Л. М. Малярець, О. Г. Тижненко, О. О. Єгоршин та ін.; за ред. Л. М. Малярець. – Харків : Вид. ХНЕУ, 2011. – 270 с.

33. Математичні методи і моделі ринкової економіки / Т. С. Клебанова, М. О. Кизим, О. І. Черняк та ін.; за ред. Т. С. Клебанової. – Харків : ВД "ІНЖЕК", 2010. – 454 с.

34. Многомерный анализ данных методами прикладной статистики : учеб. пособ. / С. С. Барковский, В. М. Захаров, А. М. Лукашов и др.; под ред. С. С. Барковского. – Казань : Изд. КГТУ, 2010. – 126 с.

35. Многомерный статистический анализ в экономике : учеб. пособ. для вузов / Л. А. Сошникова, В. Н. Тамашевич, Г. Уебе и др.; под ред. проф. В. Н. Тамашевича. – Москва : ЮНИТИ-ДАНА, 1999. – 598 с.

36. Модели оценки и анализа сложных социально-экономических систем / под ред. В. С. Пономаренко, Т. С. Клебановой, Н. А. Кизима. – Харьков : ИД "ИНЖЭК", 2013. – 659 с.

37. Моделирование устойчивого развития регионов : монография / Н. А. Кизим, О. Ю. Полякова, В. Е. Хаустова и др.; под. ред. Н. А. Кизима. – Харьков : ИД "ИНЖЕК", 2010. – 180 с.

38. Моделі і методи соціально-економічного прогнозування : підручник / В. М. Геєць, Т. С. Клебанова, О. І. Черняк та ін.; за ред. В. М. Гейця. – Харків : ВД "ІНЖЕК", 2005. – 396 с.

39. Моделювання системних характеристик в економіці / О. М. Сінчук, Т. М. Берідзе, В. В. Кононенко та ін.; за ред. О. М. Сінчук. – Кременчук : ПП Щербатих О. В., 2009. – 211 с.

40. Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных : учеб. пособ. / А. Д. Наследов. – Санкт-Петербург : Речь, 2004. – 392 с.

41. Нейронные сети. *Statistica Neural Networks:* Методология и технологии современного анализа данных / под ред. В. П. Боровикова. – 2-е изд., перераб. и доп. – Москва : Горячая линия – Телеком, 2008. – 392 с.

42. Ниворожкина Л. И. Многомерные статистические методы в экономике : учебник / Л. И. Ниворожкина, С. В. Арженовский. – Москва : Изд.-торг. корпорация "Дашков и К°"; Ростов-на-Дону : Наука-Спектр, 2009. – 224 с.

43. Плюта В. Сравнительный многомерный анализ в экономических исследованиях / В. Плюта; пер. с польск. В. В. Иванова; научн. ред. В. М. Жуковской. – Москва : Статистика, 1980. – 151 с.

44. Практикум з навчальної дисципліни "Багатовимірний статистичний аналіз" для студ. спец. "Прикладна економіка" ден. форми навч. / Т. С. Клебанова, Л. С. Гур'янова, О. А. Сергієнко. – Харків : ХНЕУ, 2011. – 74 с.

45. Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др.; под ред. С. А. Айвазяна. – Москва : Финансы и статистика,1989. – 587 с.

46. Прогнозування соціально-економічних процесів / Т. С. Клебанова, В. А. Курзенев., В. М. Наумов та ін.; за ред. Т. С. Клебанової. – Харків : ХНЕУ ім. С. Кузнеця, 2015. – 654 с.

47. Сидоренко Е. Методы математической обработки в психологии / Е. Сидоренко. – Санкт-Петербург : Речь, 2006. – 350 с.

48. Симчера В. М. Методы многомерного анализа статистических данных : учеб. пособ. / В. М. Симчера. – Москва : Финансы и статистика, 2008. – 400 с.

49. Халафян А. А. *Statistica* 6.0: статистический анализ данных : учебник / А. А. Халафян. – 3-е изд. – Москва : ООО "Бином-Пресс", 2007. – 512 с.

50. Хьюбер П. Робастность в статистике / П. Хьюбер. – Москва : Мир, 1984. – 303 с.

51. Цисарь И. Ф. Компьютерное моделирование экономики / И. Ф. Цисарь. – Москва : ДИАЛОГ-МИФИ, 2008. – 382 с.

52. Шуленин В. П. Введение в робастную статистику / В. П. Шуленин. – Томск : Изд-во Том. ун-та, 1993. – 227 с.

53. Birch M. W. A new proof of the Pearson–Fisher theorem / M. W. Birch // Ann. Math. Statist. – 1964. – V. 35. – P. 817.

54. Eisenberger J. Systematic statistics used for data compression in space telemetry / J. Eisenberger, E.C. Posner // J. Amer. Statist. Ass. 60. – 1965. – Pp. 97–133.

55. Gupta S.S. Estimation of the parameters of the logistic distribution / S.S. Gupta, M. Gnanadesikan // Biometrika, 53, 1966. – Pp. 565–570.

56. Hampel F. R. The influence curve and its role in robust estimation / F. R. Hampel // J. Amer. Statist. Ass. – 1974. – V. 69. – № 346. – Pp. 383–393.

57. Ogawa J. Contributions to the theory of systematic statistics / J.Ogawa // I. Osaka Math. J. 3, 1951. – Pp. 175–213.

58. Saleh A.K.M.J. Asymptotic optimum quantiles for the estimation of the parameters of the negative exponential distribution / A.K.M.J. Saleh, M.M. Ali // Ann. Math. Statist. 37, 1966. – Pp. 143–151.

59. Андерсон Т. Введение в многомерный статистический аналіз / Т. Андерсон. – [Електронний ресурс]. – Режим доступу : http://www.knigka.org.ua/2007/10/26/vvedenie_v_mnogomernyjj_statisticheskijj_analiz.html.

60. Винюков И. А. Многомерные статистические методы : учеб. пособ. / И. А. Винюков. – Москва : Финуниверситет, 2014 – 192 с. [Электронный ресурс]. – Режим доступа : elib.fa.ru/fbook/vinyukov_stat_metod.pdf.

61. Годун В. М. Інформаційні системи і технології в статистиці / В. М. Годун, Н. С. Орленко, М. А. Сендзюк [Електронний ресурс] – Режим доступу : http://library.if.ua/book/80/5668.html.

62. Головне управління статистики в Харківській області [Електронний ресурс]. – Режим доступу : www.kh.ukrstat.gov.ua.

63. Державна служба статистики України. [Електронний ресурс]. – Режим доступу : www.ukrstat.gov.ua.

64. Економіко-математичні методи аналізу господарської діяльності [Електронний ресурс]. – Режим доступу : http://www.unicyb.kiev.ua/Library/ TEA/3%5B1%5D.pdf.

65. Міністерство Економіки України [Електронний ресурс]. – Режим доступу : http://www.me.gov.ua/.

66. Міністерство Фінансів України [Електронний ресурс]. – Режим доступу : www.minfin.gov.ua.

67. Многомерная классификация регионов России по динамике их развития в 2005-2011 гг. : монография / колл. авт. под науч. ред. И. А. Винюкова. – Москва : Финуниверситет, 2014 – 180 с. [Электронный ресурс]. – Режим доступа : http://elib.fa.ru/fbook/vinyukob_monography.pdf.

68. Обозревательный математический сайт [Электронный ресурс]. – Режим доступа : www.econometrics.exponenta.ru.

69. Офіційний сайт Державної податкової адміністрації України. – Режим доступу : www.sta.gov.ua.

70. Практичний досвід інформаційно-аналітичної підтримки процедур розробки і прийняття управлінських рішень [Електронний ресурс]. – Режим доступу : http://www.ecsor.com.ua/files/conf_report_2_ukr.pdf.

71. Сайт Экономического Факультета МГУ. – Режим доступа : www.econ.msu.ru.

72. Сервер Государственного комитета статистики Украины. – Режим доступу : www.ukrstat.gov.ua.

73. Сервер Национального банка Украины. – Режим доступа : www.bank.gov.ua.

74. Украинская инвестиционная газета [Электронный ресурс]. – Режим доступа : www.investgazeta.net.

75. Украинский финансовый сервер. – Режим доступа : www.ifs.kiev.ua.

76. Электронный учебник StatSoft [Электронный ресурс]. – Режим доступа : http://www.statsoft.ru/statportal/tabID__44/DesktopDefault.aspx.

НАВЧАЛЬНЕ ВИДАННЯ

Клебанова Тамара Семенівна
Гур'янова Лідія Семенівна
Чаговець Любов Олексіївна та ін.

БІЗНЕС-АНАЛІТИКА БАГАТОВИМІРНИХ ПРОЦЕСІВ

Навчальний посібник

Самостійне електронне текстове мережеве видання

Відповідальний за видання *Т. С. Клебанова*

Відповідальний редактор *М. М. Оленич*

Редактор *Н. І. Ганцевич*

Коректор *Н. І. Ганцевич*

План 2018 р. Поз. № 24-ЕНП. Обсяг 272 с.

Видавець і виготовлювач – ХНЕУ ім. С. Кузнеця, 61166, м. Харків, просп. Науки, 9-А

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру
ДК № 4853 від 20.02.2015 р.