МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Прикарпатський національний університет
імені Василя Стефаника
Кафедра статистики і вищої математики

**Осипчук М.М.**

# СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕНЬ

КОНСПЕКТ ЛЕКЦІЙ

Івано-Франківськ – 2007

# Зміст

# Передмова

Взагалі статистичні дослідження — це дослідження деякої підмножини або загальної сукупності предметів, елементів, осіб, подій, тощо з метою визначення їх кількісних характеристик, таких, як частина елементів з певною ознакою, середнє або сумарне значення деякого фактора, частота появи певної події та ін. Ці характеристики та їх параметри, як правило, слугують для розуміння цінності або особливостей даної сукупності з тим, щоб відносно неї можна було прийняти більш-менш розумні рішення. Часто дослідник може обстежити деякий процес або комплекс взаємозв'язків для того, щоб краще зрозуміти їх механізм.

Можна з упевненістю стверджувати, що практично кожне наукове дослідження пов'язане з вивченням деякої вибірки. Більш того, наші знання, міркування і вчинки в основному спираються на вибіркові дані і це твердження істинне як для повсякденного життя, так і для серйозної наукової роботи, оскільки і в життєвих ситуаціях, і в науці нам відомий лише фрагмент тієї загальної картини, яка повинна розширити наші знання у тому чи іншому питанні.

Вибіркові методи дають значну економію часу і коштів на проведення досліджень порівняно з повним обстеженням і при цьому забезпечують необхідну точність резуль-

татів, а в деяких випадках є єдиним методом отримання необхідної інформації.

В пропонованому конспекті лекцій розглянуто основні методи проведення статистичних досліджень від збору інформації до її обробки та формулювання висновків. Виклад матеріалу базується на посібнику Мамчич Т.І., Оленко А.Я., Осипчук М.М., Шпортюк В.Г. Статистичний аналіз даних з пакетом Statistica. - Дрогобич: Видавнича фірма "Відродження", 2006.

# Лекція 1

# Методи вибіркових обстежень

## 1.1 Планування статистичних обстежень

На етапі планування обстеження, дослідник повинен проаналізувати цілий ряд різноманітних проблем і знайти відповідь на такі типові питання:

1. Що є метою обстеження (дослідження)?

2. Які характеристики потрібно виміряти й обчислити?

3. Що буде елементом обстеження (адміністративний район, підприємство, сім'я чи окрема особа)?

4. Яким чином буде отримано необхідну інформацію (за допомогою наявної статистичної звітності, інтерв'ю, анкетуванням по пошті чи по телефону, або в результаті інших спостережень)?

5. Необхідний ступінь точності висновків, величина і характер ризику від помилки, на який можна піти.

6. Фінансові та інші ресурси.

7. Вартість кожної операції.

8. Кваліфікація та рівень підготовки персоналу.

9. Характер подання результатів обстеження (звіти, публікації).

10. Ступінь деталізації та секретності інформації.

11. Буде проведене суцільне чи вибіркове обстеження ?

12. Необхідна кількість елементів у вибірці (обсяг вибірки), що гарантує необхідну точність.

13. Які аналітичні методи будуть застосовані? Вид статистичного аналізу та програмного забезпечення.

14. Методи знаходження вибіркових характеристик, процедури статистичного оцінювання та перевірки гіпотез.

15. Оцінка точності результатів.

16. Побудова довірчих інтервалів.

17. Тлумачення результатів.

Цей перелік включає якісний аналіз проблематики, питання фінансового та матеріального забезпечення, менеджменту та управління персоналом. Суттєва складова частина (п. 11 — 17) стосується теоретико-ймовірнісних і статистичних методів планування та аналізу вибіркових даних. Саме ці питання будуть темою подальшого викладу. Зауважимо, що конкретні вибірки можуть бути сформовані за різними методиками. Так, для оцінки рівня витрат на харчування можна відібрати перших 20 студентів за списком усього курсу; для з'ясування політичних поглядів населення регіону дослідник може відібрати, 100 типових, на його, погляд жителів. Чи будуть результати таких обстежень відзеркалювати справжній стан справ, або інакше, чи будуть ці вибірки репрезентативними? У наведених вище прикладах позитивної відповіді на ці питання дати неможливо, оскільки властивості таких вибірок невідомі, а суб'єктивний фактор може відігравати значну роль.

Інша ситуація, коли вибірки сформовані на основі об'єктивних формалізованих правил випадкового відбору. Тіль-

ки цей тип вибірки має розроблену теорію, яка включає методи й алгоритми обробки даних, здатна описати кількісні характеристики точності результатів та дати певні рекомендації щодо планування вибіркового обстеження. При цьому для аналізу вибіркових даних залучають розвинений апарат теорії ймовірностей та математичної статистики. У наступних розділах нагадаємо базові поняття математичної статистики, необхідні для правильного розуміння основних статистичних процедур, які використовують при обробці вибіркових даних.

## 1.2 Основні поняття математичної статистики

Потрібно розрізняти такі поняття:

***Загальна сукупність.*** *Загальна сукупність* - набір елементів, властивості та характеристики котрих будуть вивчатися.

Загальною сукупністю може бути все населення країни або якогось населеного пункту, усі супермаркети міста, усі трикотажні фабрики, усі працівники певної фірми, усі студенти ВУЗу тощо.

***Генеральна сукупність.*** Далі нас будуть цікавити лише кількісні характеристики елементів загальної сукупності, наприклад, рівень доходів жителів міста, місячний товарообіг супермаркетів, стаж працівників фірми, витрати на транспорт студентів ВУЗу. Таким чином, приходимо до поняття сукупності за ознакою або генеральної сукупності.

Більш формально, при кожному конкретному дослідженні загальна сукупність із кількісного боку зображує-

ться деякою випадковою величиною.

*Генеральна сукупність* — це множина всіх значень, які може набувати дана випадкова величина.

Так, якщо на фірмі працює 120 чоловік, то генеральна сукупність, що описує їхній стаж роботи, складається зі 120 даних; генеральна сукупність даних стосовно доходів 200000-ного міста складається із 200000 даних.

У математичній статистиці розроблені методи аналізу даних, що стосуються як скінченних, так і нескінченних генеральних сукупностей.

Далі, кількість елементів у генеральній сукупності будемо позначати $N$ і називати обсягом генеральної сукупності.

Строго математично генеральну сукупність визначають як імовірнісний простір із заданою на ньому випадковою величиною. Тому можна говорити про ймовірнісний розподіл генеральної сукупності та такі параметри, як математичне сподівання, дисперсія, середня квадратична похибка, коефіцієнт варіації та ін.

***Вибірка.*** У широкому розумінні, *вибірка* — деяка частина елементів загальної сукупності, кількість елементів у ній називають обсягом вибірки.

*Проста випадкова вибірка* — це вибірка, в якій кожен елемент має рівну і незалежну від інших імовірність бути відібраним і включеним у вибірку.

Після того, як вибірка сформована, визначають (вимірюють) значення ознаки, яку вивчають. Результат дослідження зображують набором із $n$ чисел $(x_1, \ldots, x_n)$. Кожне вибіркове значення можна тлумачити як випадкову величину, а всі вибіркові дані — як випадковий вектор.

Конкретні значення отримані при обстеженні вибірки — реалізація вибірки.

Часто вибіркові значення зручно розглядати в порядку зростання

$$x_1^* \le x_2^* \le \ldots \le x_n^*,$$

$x_1^*$ — найменше, $x_n^*$ — найбільше з можливих значень.

**Означення**. *Набір* $(x_1^*, \ldots, x_n^*)$ *називають варіаційним рядом, а випадкові величини* $x_i^*$, $i = 1, \ldots, n$ *— порядковими статистиками.*

## 1.3 Типи вибірок зі скінченної загальної сукупності

Розрізняють два типи вибірок:

- Коли у вибірках не допускають дублювання (повторення) елементів незалежно від порядку їх відбору, то такі вибірки називають безповторними (вибірками без повернення). Вони відповідають схемі відбору жетонів із номерами з урни без повернення вже відібраних жетонів.

- Якщо ж допускають дублювання елементів у вибірці, то такий відбір називають повторним (з поверненням). Він відповідає ситуації, коли після кожного кроку відібраний жетон знову повертають в урну.

Весь подальший матеріал, якщо не обумовлено протилежне, стосується безповторних вибірок. Практично випадковий відбір можна здійснити, якщо є можливість упорядкувати всі елементи загальної сукупності та присвоїти їм відповідні номери, а потім відібрати елементи вибірки користуючись таблицями випадкових чисел або комп'ютерними датчиками випадкових чисел.

## 1.4 Простий та стратифікований випадковий вибір

Якщо з генеральної сукупності обсягом у $N$ одиниць випадковим чином вибирають $n$ одиниць, то такий вибір називають простим, або власне випадковим. У такому випадку вибірка — це підмножина повної множини обсягом $n < N$, отримана за деяким правилом, яке забезпечує рівні можливості бути вибраними для всіх елементів множини.

Простий вибір характеризують трьома числами:

$N$ — обсяг генеральної сукупності,

$n$ — об'єм вибірки,

$f = n/N$ — частка відбору.

На практиці найбільш розповсюдженими та важливими параметрами, які оцінюють за вибіркою, є:

- частка і кількість одиниць із певною ознакою;
- середні значення ознак, які вивчають;
- сумарні значення ознак, які вивчають;
- частка двох сумарних або середніх значень.

Саме ці величини (характеристики, параметри) відображають рівень, структуру та динаміку суспільних явищ.

При стратифікованому відборі генеральну сукупність, що складається з $N$ елементів розділяють на $L$ підсукупностей обсягом $N_1, N_2, \ldots, N_L$ , що не мають спільних елементів і $N_1 + N_2 + \ldots + N_L = N$. Такі підсукупності називають стратами (від англійського stratum). Можливі дві ситуації:

- страти визначають природним чином з аналізу проблеми, на вирішення якої спрямоване обстеження, і сукупність a priori розбита на страти ( наприклад, згідно з територіальним або адміністративним поділом, за віковими або професійними ознаками);

- дослідник сам визначає страти з метою отримати більшу точність висновків порівняно з простим випадковим відбором при фіксованому обсязі вибірки або ж фіксовану точність, але при меншому обсязі вибірки.

Після того, як страти визначені, вибірка здійснюється для кожної з них окремо. Якщо ці вибірки здійснюють за правилами простого випадкового відбору, то в цілому всю схему обстеження називають стратифікованим випадковим відбором. Такий метод вибіркового обстеження досить часто використовують, якщо:

1. Необхідно отримати висновки не тільки для всієї сукупності, але й для певних підрозділів.

2. Застосування страт зумовлено організаційними міркуваннями, коли організації та установи, що замовляють або проводять обстеження, мають районні відділення, кожне з яких забезпечує обстеження у своєму регіоні.

3. Проблеми, пов'язані з проведенням обстежень можуть дуже відрізнятися в різних частинах сукупності. Так, наприклад, при вивченні ділової активності можна виділити в окремий список великі фірми, а при обстеженні малих підприємств використати територіальний підхід.

4. Стратифікація може дати виграш у точності при оцінюванні параметрів генеральної сукупності.

Таким чином, крім загальних статистичних проблем оцінювання основних параметрів генеральної сукупності (частки, середнього, сумарного значення тощо), їх середньоквадратичних похибок, побудови довірчих інтервалів і ви-

значення необхідного обсягу вибірки при стратифікованому відборі виникають і додаткові задачі, пов'язані з оптимальним розбиттям генеральної сукупності на страти і розподілом елементів по стратах.

## 1.5 Багатоступеневий гніздовий відбір

При багатоступеневому відборі елементи, що безпосередньо досліджуються, вибирають лише на останній стадії, після декількох послідовних випадкових відборів. Таким чином виділяють елементи відбору першого ступеня, другого тощо.

Приклад. Треба дослідити особисті підсобні господарства за декількома ознаками. Потрібний відбір можна виконати в три етапи:

1. елементи відбору першого ступеня: адміністративні райони (наприклад, 50% всіх районів),

2. елементи відбору другого ступеня: села (наприклад, 20% усіх сіл району),

3. елементи відбору третього ступеня: особисті підсобні господарства (наприклад, 30% господарств із відібраних сіл).

На кожному етапі проводиться простий випадковий відбір, який характеризується своєю часткою відбору $f_1, f_2, f_3, \ldots$. У наведеному прикладі $f_1 = 0.5$, $f_2 = 0.2$, $f_3 = 0.3$. Остаточно частка відібраних для обстеження особистих підсобних господарств є $f = f_1 f_2 f_3 = 0.5 \cdot 0.2 \cdot 0.3 = 0.03$, тобто будуть обстежені лише 3% усіх особистих господарств.

Гніздовий відбір значно зменшує та спрощує обстеження, проте при цьому виникають деякі методологічні складнощі при аналізі отриманих даних.

# Лекція 2

# Оцінювання параметрів

З кількісного боку генеральна сукупність характеризується цілим рядом показників (параметрів). З погляду математичної статистики та можливості змістовного тлумачення даних основними параметрами є математичне сподівання (генеральне середнє), дисперсія, середня квадратична похибка, коефіцієнт кореляції та параметри регресії (при вивченні взаємозв'язку між декількома ознаками).

Числове значення генерального параметра можна знайти, якщо мати повну інформацію стосовно всієї генеральної сукупності. Але такої інформації, як правило немає. Вибіркове обстеження саме і спрямовано на отримання висновків щодо параметрів генеральної сукупності на основі вибірки. Це робиться шляхом обчислень за певними формулами типу $\hat{a} = f(x_1, \ldots, x_n)$, які дають змогу приблизно оцінити справжнє (проте невідоме) значення параметра, що досліджується.

**Означення**. *Оцінкою $\hat{a}$ невідомого значення параметра $a$ генеральної сукупності називають відповідну вибіркову характеристику (функцію спостережень)*

$$\hat{a} = f(x_1, \ldots, x_n),$$

*яку обчислюють за результатами вибіркових обстежень.*

Таким чином, розглядають вибіркове середнє, вибіркову дисперсію, вибіркові коефіцієнти варіації та кореляції тощо. Усі ці вибіркові характеристики слугують оцінками відповідних параметрів генеральної сукупності. На конкретних процедурах їх обчислення зупинимося пізніше.

## 2.1 Методи одержання оцінок

### 2.1.1 Емпіричні оцінки

Нехай $\zeta = (x_1, ..., x_n)$ — вибірка із генеральної сукупності з функцією розподілу $\mathbf{F}(x, \theta)$, де $\theta$ — невідомий параметр такий, що однозначно визначається розподілом. Тобто

$$\theta = \Phi(\mathbf{F}(x, \theta)),$$

де $\Phi$ — функція, визначена на множині функцій розподілу. Наприклад, параметр $\theta = \mathbf{E}\xi$ визначається щільністю $p(x) = p(x, \theta)$ так:

$$\theta = \int_{-\infty}^{+\infty} xp(x)dx.$$

Оскільки емпірична функція розподілу $\hat{\mathbf{F}}_n(x)$ є оцінкою $\mathbf{F}(x, \theta)$, тобто у певному розумінні "близька"до $\mathbf{F}(x, \theta)$, то за оцінку $\theta$ можна брати

$$\hat{\theta} = \Phi(\hat{\mathbf{F}}_n(x)).$$

Наприклад, для $\theta = \mathbf{E}\xi$, оскільки $\hat{\mathbf{F}}_n(x)$ відповідає дискретний розподіл, отримуємо

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

**Означення.** *Оцінку* $\hat{\theta}_n = \Phi(\hat{\mathbf{F}}_n(x))$ *параметра* $\theta = \Phi(\mathbf{F}(x,\theta))$ *називають емпіричним (вибірковим) значенням параметра* $\theta$.

### 2.1.2 Метод моментів

Нехай $\zeta = (x_1, ..., x_n)$ — вибірка із генеральної сукупності з функцією розподілу $\mathbf{F}(x, \theta_1, \theta_2, ..., \theta_s)$, $\theta_i \in R^s$. Потрібно отримати оцінки параметрів $\theta_1, \theta_2, ..., \theta_s$. Нехай $m_i(\theta_1, \theta_2, ..., \theta_s)$ — момент $i$-го порядку, підрахований за функцією розподілу $\mathbf{F}(x, \theta_1, \theta_2, ..., \theta_s)$.
Наприклад, для абсолютно-неперервних розподілів

$$m_i(\theta_1, \theta_2, ..., \theta_s) = \int_{-\infty}^{+\infty} x^i p(x, \theta_1, \theta_2, ..., \theta_s) dx,$$

де $p(x, \theta_1, \theta_2, ..., \theta_s)$ — щільність, яка відповідає функції розподілу
$\mathbf{F}(x, \theta_1, \theta_2, ..., \theta_s)$. Відповідні емпіричні моменти визначають так:

$$\hat{m}_i = \frac{1}{n} \sum_{k=1}^{n} x_k^i.$$

Метод моментів полягає у тому, що деяка кількість емпіричних моментів $\hat{m}_i$ прирівнюють до відповідних моментів $m_i(\theta_1, \theta_2, ..., \theta_s)$ і з цієї системи рівнянь знаходять оцінки $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$.
Приклад. За вибіркою $\zeta = (x_1, ..., x_n)$ із генеральної сукупності з розподілом $N(a, \sigma^2)$ методом моментів отримати оцінки для середнього $a$ та дисперсії $\sigma^2$.

Оскільки

$$m_1(\hat{a}, \hat{\sigma}^2) = \hat{a}, \qquad m_2(\hat{a}, \hat{\sigma}^2) = \hat{\sigma}^2 + (\hat{a})^2$$

і відповідні емпіричні моменти

$$\hat{m}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k, \qquad \hat{m}_2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2,$$

то отримуємо систему рівнянь

$$\hat{a} = \frac{1}{n}\sum_{k=1}^{n} x_k, \qquad (\hat{a})^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2.$$

Отже, шукані оцінки такі:

$$\hat{a} = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2.$$

### 2.1.3 Метод максимальної вірогідності

Нехай $\zeta = (x_1, ..., x_n)$ — вибірка обсягу $n$ із генеральної сукупності з розподілом $p(x, \theta_1, \theta_2, ..., \theta_s)$, який залежить від невідомих параметрів $(\theta_1, \theta_2, \ ..., \theta_s) \in \Theta$. Якщо розподіл абсолютно неперервний, то $p(x, \theta_1, \theta_2, ..., \theta_s)$ — його щільність, а якщо дискретний, то $p(x, \theta_1, \theta_2, ..., \theta_s)$ — ймовірність значення $x$.

**Означення.** Функцією максимальної вірогідності вибірки $\zeta = (x_1, ..., x_n)$ називають функцію

$$L_{(\theta_1,\theta_2,...,\theta_s)}(\zeta) = \prod_{i=1}^{n} p(x_i, \theta_1, \theta_2, ..., \theta_s).$$

Далі, якщо ми розглядаємо функцію вірогідності при конкретній реалізації вибірки, і нас цікавлять лише параметри $\theta_1, \theta_2, ..., \theta_s$, то будемо скорочено писати $L(\theta_1, \theta_2, ..., \theta_s)$.

**Означення.** Логарифмічною функцією вірогідності називають функцію

$$\ln L_{(\theta_1,\theta_2,...,\theta_s)}(\zeta) = \sum_{i=1}^{n} \ln p(x_i, \theta_1, \theta_2, ..., \theta_s).$$

Метод максимальної вірогідності спирається на таке інтуітивне уявлення: у експерименті в більшості випадків спостерігають те значення вектора $\zeta = (x_1, ..., x_n)$, при якому щільність близька до максимального значення.

Отже, за оцінку параметрів $\theta_1, \theta_2, ..., \theta_s$ беремо точку $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s$, у якій

$$L(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_s) = \max_{(\theta_1, \theta_2, ..., \theta_s) \in \Theta} L(\theta_1, \theta_2, ..., \theta_s). \tag{2.1}$$

Оскільки $L$ і $\ln L$ набувають максимальних значень в одних і тих же точках, то можна проробити цю процедуру не для $L$, а для $\ln L$ (в деяких випадках це зручніше).

За певних умов, оцінки, отримані методом максимальної вірогідності, конзистентні, асимптотично ефективні та асимптотично нормальні.

### 2.1.4 Метод найменших квадратів

Важливий приклад застосування методу максимальної вірогідності — метод найменших квадратів. На практиці його використовують для отримання наближених експериментальних залежностей.

Нехай деяка закономірність визначається функцією $y = y(x)$. Будемо "наближено" вважати, що

$$y = \theta_0 \varphi_0(x) + ... + \theta_s \varphi_s(x) = \sum_{j=0}^{s} \theta_j \varphi_j(x),$$

де $\varphi_i(\cdot)$, $i = 0, 1, ..., s$ — відомі функції, а $\theta_0, \theta_1, ..., \theta_s$ — невідомі параметри, які потрібно оцінити.

Нехай в точках $x_1, ..., x_n$ зроблені спостереження змінної $y = y(x)$, результати яких відповідно $y_1, ..., y_n$. Будемо

вважати, що відхилення спостережень від справжніх значень $y = y(x)$ — незалежні нормальні $N(0, \sigma^2)$ випадкові величини. Тоді логарифмічна функція вірогідності дорівнює:

$$\ln L_{(\theta_1,\theta_2,\ldots,\theta_s)}(\zeta) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=0}^{s}\theta_j\varphi_j(x_i))^2.$$

Для знаходження максимума логарифмічної функції вірогідності потрібно мінімізувати суму квадратів у правій частині (звідси і назва — метод найменших квадратів). Отже, оцінки $\hat{\theta}_1, ..., \hat{\theta}_s$ параметрів $\theta_1, ..., \theta_s$ беруть такими, щоб сума

$$\sum_{i=1}^{n}(y_i - \sum_{j=0}^{s}\theta_j\varphi_j(x_i))^2$$

була мінімальною.

## 2.2 Точкові та інтервальні оцінки

Припустимо, що дослідник отримав конкретну реалізацію вибірки $(x_1, \ldots, x_n)$ і обчислив значення $\hat{a}$ за формулою $\hat{a} = f(x_1, \ldots, x_n)$, тобто отримав єдине значення $\hat{a}$, яке є приблизним значенням параметра $a$. Такі оцінки називають точковими.

Більш інформативними стосовно точності апроксимації є так звані інтервальні оцінки, що вказують інтервал, який із фіксованою (заданою) ймовірністю $\mathcal{P}$ містить справжнє значення параметра $a$.

Ймовірність $\mathcal{P} = 1-\alpha$ називають надійністю або рівнем довіри, а сам інтервал — довірчим або надійним інтервалом. Для величини $\alpha$ вживають термін критичний рівень або похибка.

**Означення.** *Довірчим (надійним) інтервалом для параметра $a$ з рівнем надійності $\mathcal{P} = 1 - \alpha$, $(0 < \alpha < 1)$ називають випадковий інтервал $(a_H, a_B)$ такий, що*

$$\mathbf{P}(a_H \leq a \leq a_B) = \mathcal{P}. \tag{2.2}$$

Співвідношення (2.2) слід читати "ймовірність того, що справжнє значення $a$ лежить у інтервалі від $a_H$ до $a_B$, дорівнює $\mathcal{P}$".

Межі інтервалу $a_H = a_H(x_1, \ldots, x_n), a_B = a_B(x_1, \ldots, x_n)$, які знаходять за допомогою вибіркових даних, називають відповідно нижньою і верхньою довірчими межами (межами надійності).

Якщо $a_H = -\infty$, то довірчий інтервал називають лівостороннім, а при $a_B = +\infty$ маємо правосторонній довірчий інтервал.

Змістовно рівень довіри означає, що при багаторазовому повторенні однакової схеми відбору та процедури оцінювання за вибіркою сталого обсягу в середньому в $\mathcal{P} \cdot 100\%$ випадків значення параметра $a$ дійсно лежить у межах від $a_H$ до $a_B$ і лише в $\alpha \cdot 100\%$ випадках може виходити за ці межі.

Традиційно $\mathcal{P}$ вибирають рівним 0,95 ($\alpha = 0,05$) або 0,99 ($\alpha = 0,01$) і говорять про 95% або 99% довірчі інтервали.

# Лекція 3

# Перевірка статистичних гіпотез

## 3.1 Постановка проблеми, основні поняття

Одна з основних задач математичної статистики — перевірка узгодженості результатів послідовності спостережень випадкових величин з гіпотезами про розподіл цих величин.

Нехай відносно розподілу вибірки $\zeta = (\xi_1, ..., \xi_n)$ відомо, що він належить до деякого класу розподілів $\mathcal{P}$. Нехай $\mathcal{G} \subset \mathcal{P}$ — деякий підклас $\mathcal{P}$ (можливо $\mathcal{G}$ містить лише один розподіл $\mathbf{F}$). Потрібно за результатами експерименту (реалізацією вибірки $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$) зробити висновок: розподіл $\zeta$ може бути із $\mathcal{G}$, чи ні.

Статистичними гіпотезами будемо називати гіпотези про розподіли випадкових величин.

Нульова (основна) гіпотеза полягає у тому, що $\zeta$ має розподіл із підкласу $\mathcal{G}$. Позначення: $H_0$. Решту гіпотез на-

зивають альтернативними чи конкурентними відносно $H_0$. Позначення: $H_1$.

У багатьох випадках клас $\mathcal{P}$ утворений розподілами $\mathbf{P}_\theta$, які визначаються параметрами $\theta \in \Theta \subset R^s$. При таких припущеннях статистична гіпотеза полягає у тому, що параметр $\theta$ розподілу $\mathbf{P}_\theta$ належить вказаній множині $H_0 \subset \Theta$. Доповнення до $H_0$ : $H_1 = \Theta \backslash H_0$ буде тоді альтернативною гіпотезою.

Приклад. Нехай $\xi$ має геометричний розподіл. Тоді

$$\mathcal{P} = \{\mathbf{P}_\theta, \theta \in (0,1)\}, \quad \text{де } \mathbf{P}_\theta = \theta(1-\theta)^n, \quad n = 0,1,2,\ldots .$$

Нехай $\mathcal{G} = \{\mathbf{P}_{\frac{1}{2}}\}$. Отже, основна гіпотеза $H_0$ полягає у тому, що $\xi$ має такий розподіл: $\mathbf{P}_{\frac{1}{2}} = \{\left(\frac{1}{2}\right)^{n+1},\ n = 0,1,2,\ldots\}$, а конкурентна гіпотеза $H_1$ — в тому, що розподіл $\xi$ — $\mathbf{P}_\theta$, де $\theta \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$.

Якщо гіпотеза полягає у тому, що $\zeta$ має розподіл $K$, де $K$ — елемент класу $\mathcal{P}$, то кажуть, що це — проста гіпотеза. У протилежному випадку гіпотезу називають складною. У параметричному випадку $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta \subset R^s\}$ гіпотеза $H_0$ проста, якщо множина $H_0$ містить лише один елемент множини $\Theta$. У протилежному випадку гіпотеза складна.

У попередньому прикладі $H_0$ — проста гіпотеза, а $H_1$ — складна.

Знаючи лише реалізацію вибірки $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ необхідно зробити висновок: розподіл $\zeta$ може бути із $\mathcal{G}$ (гіпотезу $H_0$ приймають) чи ні (гіпотезу $H_0$ відхиляють).

Отже, необхідно множину можливих результатів $\zeta(\omega)$ (вибірковий простір) розбити на дві частини: множину результатів $\mathcal{S}$, при яких $H_0$ відхиляють і $\bar{\mathcal{S}}$, при яких $H_0$ приймають.

**Означення.** Множину $\mathcal{S}$ називають критичною множиною (областю) чи критерієм для перевірки гіпотези $H_0$,

якщо при $\zeta(\omega) \in \mathcal{S}$ гіпотезу $H_0$ відхиляють, а при $\zeta(\omega) \notin \mathcal{S}$ приймають.

Оскільки критичну область можна вибрати багатьма способами постає питання: які є числові характеристики "якості" критерія.

Гіпотезу $H_0$ ми перевіряємо так: якщо реалізація вибірки $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ потрапляє до $\mathcal{S}$, то $H_0$ відхиляємо, якщо $\zeta(\omega) \notin \mathcal{S}$, то гіпотезу $H_0$ приймаємо. При цьому у нас можливі такі помилки:

1. Гіпотеза $H_0$ істинна, але ми її відхилили, оскільки $\zeta(\omega) \in \mathcal{S}$;

2. Гіпотеза $H_0$ хибна, але ми її приймаємо, оскільки $\zeta(\omega) \notin \mathcal{S}$.

Помилку 1 називають помилкою першого роду, а помилку 2 — помилкою другого роду. На практиці звичайно із двох гіпотез за основу обирають ту, для якої помилка першого роду більш "шкідлива", ніж помилка другого роду.

Зрозуміло, що оскільки $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ — випадкова величина, то взагалі кажучи побудувати критерій $\mathcal{S}$, який би не приводив до помилок неможливо. Зате можна вибирати критерій $\mathcal{S}$ так, щоб були невеликі ймовірності помилок.

Введемо такі позначення:

$$\mathbf{P}_{H_0}(A) = \sup_{\mathbf{P} \in \mathcal{G}} \mathbf{P}(A), \qquad \mathbf{P}_{H_1}(A) = \sup_{\mathbf{P} \in \mathcal{P} \setminus \mathcal{G}} \mathbf{P}(A).$$

У параметричному випадку $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta \subset R^s\}$ :

$$\mathbf{P}_{H_0}(A) = \sup_{\theta \in H_0} \mathbf{P}_\theta(A), \qquad \mathbf{P}_{H_1}(A) = \sup_{\theta \in H_1 = \Theta \setminus H_0} \mathbf{P}_\theta(A).$$

**Означення.** Рівнем значимості $\alpha$ критерію $\mathcal{S}$ називають число:

$$\alpha = \mathbf{P}_{H_0}(\mathcal{S}).$$

Рівень значимості $\alpha$ обмежує зверху ймовірність помилки першого роду.

**Означення.** Функцією потужності критерію $\mathcal{S}$ називають функцію $\beta : \mathcal{P}\backslash\mathcal{G} \to [0,1]$, яка для довільного $\mathbf{P} \in \mathcal{P}\backslash\mathcal{G}$ визначена так:

$$\beta(\mathbf{P}) = \mathbf{P}(\mathcal{S}).$$

Ця функція при різних розподілах, які відповідають альтернативній гіпотезі, дорівнює ймовірності попадання реалізації вибірки в критичну область при справедливій альтернативній гіпотезі.

Якщо при заданому рівні значимості $\alpha$ критерій $\mathcal{S}$ можна вибрати не одним способом, то потрібно обирати той, при якому ймовірність помилки другого роду $\mathbf{P}_{H_1}(\bar{\mathcal{S}})$ найменша.

Оскільки звичайно помилка першого роду більш "шкідлива", ніж другого, то для "гарного" критерію перевірки гіпотез рівень значимості $\alpha$ має бути менший ймовірності помилки другого роду $\mathbf{P}_{H_1}(\bar{\mathcal{S}})$.

**Означення.** Нехай рівень значимості дорівнює $\alpha$. Критерій $\mathcal{S}^*$ називають рівномірно найбільш потужним, якщо для будь-якого іншого критерію $\mathcal{S}$ :

$$\mathbf{P}_{H_1}(\mathcal{S}^*) \leq \mathbf{P}_{H_1}(\mathcal{S}).$$

Працювати з вибірками і критеріями в $n$-вимірному просторі $R^n$ незручно. Тому звичайно діють так: розглядають деяку дійснозначну функцію $\varphi : R^n \to R$ (статистику). Вона відображає критичну область $\mathcal{S}$ в деяку множину $I \subset R$. Якщо при застосуванні $\varphi$ до реалізації вибірки

$\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ $\varphi(\xi_1(\omega), ..., \xi_n(\omega)) \in I$, то гіпотезу $H_0$ відхиляють, а якщо $\varphi(\xi_1(\omega), ..., \xi_n(\omega)) \in R \backslash I$, то приймають. Звичайно розглядають такі I, що $R \backslash I = (a, b)$. (Проміжок не обов'язково скінченний. Одне з чисел $a, b$ може бути $-\infty$ або $+\infty$ відповідно.)

## 3.2 Перевірка гіпотез для нормальних розподілів

### 3.2.1 Перевірка гіпотези про значення середнього

***Випадок, коли дисперсія невідома***

Нехай $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ — реалізація вибірки із генеральної сукупності з нормальним розподілом $N(a, \sigma^2)$ з невідомими параметрами $a$ та $\sigma^2$. Гіпотеза $H_0$ полягає у тому, що $a = a_0$. Конкурентна гіпотеза $H_1$ може бути або $a \neq a_0$ (двостороння альтернатива), або $a > a_0$ $(a < a_0)$ (одностороння альтернатива).

Розглянемо оцінку $\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$. Вона незміщена і конзистентна оцінка параметра $a$. Отже, відхилення $\bar{\xi}$ від $a$ в середньому менше, ніж відхилення $\bar{\xi}$ від $a_0 \neq a$. Тому критерій можна будувати так: відхиляти гіпотезу $H_0$, якщо $\bar{\xi} - a_0$ велике і приймати $H_0$, якщо $\bar{\xi} - a_0$ мале.

Для цього скористаємось тим, що випадкова величина

$$\varphi(\zeta) = \frac{\bar{\xi} - a}{s/\sqrt{n}}, \quad \text{де} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2,$$

має розподіл Стьюдента з $(n-1)$ ступенем вільності.

Нехай $t_{1-\frac{\alpha}{2},n-1}$ — квантіль рівня $1-\frac{\alpha}{2}$ розподілу Стьюдента з $(n-1)$ ступенем вільності. Будемо приймати гіпотезу $H_0$: $a=a_0$, якщо

$$\frac{|\bar{\xi}-a_0|}{s/\sqrt{n}} < t_{1-\frac{\alpha}{2},n-1}$$

і відхиляти, якщо

$$\frac{|\bar{\xi}-a_0|}{s/\sqrt{n}} \geq t_{1-\frac{\alpha}{2},n-1}.$$

Ймовірність помилки першого роду при цьому — $\alpha$. Для побудованого так критерію конкурентна гіпотеза була $a \neq a_0$. При односторонній альтернативі, наприклад $a > a_0$, гіпотезу відхиляють, якщо

$$\frac{\bar{\xi}-a_0}{s/\sqrt{n}} \geq t_{1-\alpha,n-1}.$$

Рівень значимості цього критерія — $\alpha$.

***Випадок, коли дисперсія відома***

Якщо на відміну від попереднього випадку дисперсія $\sigma^2$ відома, то для побудови критерію можна розглянути статистику $\varphi(\zeta)=\frac{\bar{\xi}-a}{\sigma/\sqrt{n}}$, яка має нормальний $N(0,1)$ розподіл. Нехай $d_{1-\frac{\alpha}{2}}$ — квантіль рівня $1-\frac{\alpha}{2}$ нормального $N(0,1)$ розподілу. Будемо приймати гіпотезу $H_0$: $a=a_0$, якщо

$$\frac{|\bar{\xi}-a_0|}{\sigma/\sqrt{n}} < d_{1-\frac{\alpha}{2}}$$

і відхиляти, якщо

$$\frac{|\bar{\xi}-a_0|}{\sigma/\sqrt{n}} \geq d_{1-\frac{\alpha}{2}}.$$

Рівень значимості такого критерію $\alpha$.

Аналогічно до попереднього при односторонній альтернативі $H_1$: $a > a_0$, критерій рівня значимості $\alpha$ :

$$\frac{\bar{\xi} - a_0}{\sigma/\sqrt{n}} \geq d_{1-\alpha}.$$

### 3.2.2 Перевірка гіпотези про рівність середніх

***Випадок, коли дисперсії невідомі***

Нехай $\zeta_1 = (\xi_1, ..., \xi_n)$ і $\zeta_2 = (\eta_1, ..., \eta_m)$ — дві незалежні вибірки з генеральних сукупностей з розподілами $N(a_1, \sigma^2)$ і $N(a_2, \sigma^2)$ відповідно. Параметри $a_1, a_2, \sigma^2$ - невідомі.

Основна гіпотеза $H_0$ полягає у тому, що $a_1 = a_2$. Конкурентна гіпотеза $H_1$: $a_1 \neq a_2$.

Розглянемо оцінку

$$\bar{\xi} - \bar{\eta} = \frac{1}{n}\sum_{i=1}^{n} \xi_i - \frac{1}{m}\sum_{j=1}^{m} \eta_j.$$

Вона є незміщеною і слушною оцінкою параметра $a_1 - a_2$. Отже, відхилення $\bar{\xi} - \bar{\eta}$ від $a_1 - a_2$ в середньому менше ніж відхилення $\bar{\xi} - \bar{\eta}$ від будь-якого іншого числа. Тому критерій для перевірки рівності середніх потрібно будувати так: відхиляти гіпотезу $H_0$, якщо $\bar{\xi} - \bar{\eta}$ значно відрізняється від 0 і приймати $H_0$, якщо значення $\bar{\xi} - \bar{\eta}$ близьке до нуля.

Нехай

$$s_\xi^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2, \qquad s_\eta^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \bar{\eta})^2,$$

$$s^2 = \frac{1}{n+m-2}[(n-1)s_\xi^2 + (m-1)s_\eta^2].$$

Для побудови критерію скористаємось тим, що при $a_1 = a_2$ статистика

$$\frac{\bar{\xi} - \bar{\eta}}{s\sqrt{\frac{n+m}{nm}}}$$

має розподіл Стьюдента з (n+m-2) ступенями вільності.

Отже, критерій такий: гіпотезу $H_0$ відхиляють, якщо

$$\frac{|\bar{\xi} - \bar{\eta}|}{s\sqrt{\frac{n+m}{nm}}} \geq t_{1-\frac{\alpha}{2},n+m-2}$$

і приймають в іншому випадку. Рівень значимості цього критерія — $\alpha$.

При односторонній альтернативній гіпотезі $H_1$: $a_1 > a_2$ гіпотезу $H_0$ відхиляють, якщо

$$\frac{\bar{\xi} - \bar{\eta}}{s\sqrt{\frac{n+m}{nm}}} \geq t_{1-\alpha,n+m-2},$$

Рівень значимості такого критерія також $\alpha$.

***Випадок, коли дисперсії відомі***

Якщо, на відміну від попереднього випадку дисперсії $\sigma_1^2$ та $\sigma_2^2$ для законів розподілу вибірок $\zeta_1$ та $\zeta_2$ відомі, то для перевірки гіпотези $H_0$ використовують статистику

$$\frac{\bar{\xi} - \bar{\eta} - (a_1 - a_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}},$$

яка має нормальний $N(0,1)$ розподіл. Тому гіпотезу $H_0$, $a_1 = a_2$ відхиляють при рівні значимості $\alpha$ і альтернатив-

ній гіпотезі $H_1$: $a_1 \neq a_2$, якщо

$$\frac{|\bar{\xi} - \bar{\eta}|}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > d_{1-\frac{\alpha}{2}}$$

і приймать в іншому випадку. При односторонній альтернативній гіпотезі $a_1 > a_2$, критерій:

$$\frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > d_{1-\alpha}.$$

### 3.2.3 Перевірка гіпотези про значення дисперсії

***Випадок, коли математичне сподівання невідоме***

Нехай $\zeta(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ — реалізація вибірки із генеральної сукупності з нормальним $N(a, \sigma^2)$ розподілом. Параметри $a$ і $\sigma^2$ — невідомі. Гіпотеза $H_0$ полягає у тому, що $\sigma^2 = \sigma_0^2$. Конкуруюча гіпотеза $H_1$ може бути або $\sigma^2 \neq \sigma_0^2$ (двостороння альтернатива) або $\sigma^2 > \sigma_0^2$ ($\sigma^2 < \sigma_0^2$) (одностороння альтернатива).

Розглянемо оцінку $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$. Як було показано раніше, вона є незміщеною і слушною оцінкою параметра $\sigma^2$. Отже, відхилення $\frac{s^2}{\sigma^2}$ від 1 в середньому менше ніж відхилення $\frac{s^2}{\sigma^2}$ від будь-якого іншого числа. Тому критерій можна будувати так: приймати гіпотезу $H_0$, якщо відношення $\frac{s^2}{\sigma_0^2}$ близьке до 1 і відхиляти у протилежному випадку.

Для побудови критерію скористаємось тим, що випадкова величина $\frac{(n-1)s^2}{\sigma^2}$ має $\chi^2$ розподіл з $(n-1)$ ступенем вільності.

Нехай $\chi^2_{\gamma,n-1}$ — квантіль рівня $\gamma$ розподілу $\chi^2$ з ($n$-1) ступенем вільності. Будемо приймати гіпотезу $H_0$: $\sigma^2 = \sigma_0^2$, якщо

$$\frac{(n-1)s^2}{\sigma_0^2} \in (\chi^2_{\frac{\alpha}{2},n-1}, \chi^2_{1-\frac{\alpha}{2},n-1})$$

і відхиляти в інших віпадках. Рівень значимості такого критерія $\alpha$. Цей критерій застосовуємо у випадку двосторонньої альтернативи. Для односторонньої альтернативи, наприклад $\sigma^2 > \sigma_0^2$, маємо критерій:

$$\frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{1-\alpha,n-1}.$$

Рівень значимості такого критерію - $\alpha$.

***Випадок, коли математичне сподівання відоме***

Якщо параметр $a$ відомий, то потрібно розглядати статистику $\frac{n\hat{\sigma}^2}{\sigma_0^2}$, де $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - a)^2$, яка має $\chi^2$ розподіл з $n$ ступенями вільності.

Тоді гіпотезу $H_0$: $\sigma^2 = \sigma_0^2$ приймаємо, якщо

$$\frac{n\hat{\sigma}^2}{\sigma_0^2} \in (\chi^2_{\frac{\alpha}{2},n}, \chi^2_{1-\frac{\alpha}{2},n})$$

і відхиляємо в інших випадках. Рівень значимості такого критерія з двосторонньою альтернативою дорівнює $\alpha$.

Для односторонньої альтернативи, наприклад, $\sigma^2 > \sigma_0^2$, при рівні значимості $\alpha$ застосовують критерій:

$$\frac{n\hat{\sigma}^2}{\sigma_0^2} \geq \chi^2_{1-\alpha,n}.$$

### 3.2.4 Перевірка гіпотези про рівність дисперсій

***Випадок, коли математичні сподівання невідомі***

Нехай $\zeta_1(\omega) = (\xi_1(\omega), ..., \xi_n(\omega))$ і $\zeta_2(\omega) = (\eta_1(\omega), ..., \eta_m(\omega))$ — реалізації вибірок із генеральних сукупностей з нормальними розподілами $N(a_1, \sigma_1^2)$ та $N(a_2, \sigma_2^2)$ відповідно. Всі параметри $a_1, \sigma_1^2, a_2, \sigma_2^2$ — невідомі.

Основна гіпотеза $H_0$ полягає у тому, що $\sigma_1^2 = \sigma_2^2$**.** Конкурентна гіпотеза $H_1$: $\sigma_1^2 \neq \sigma_2^2$. Оцінки

$$s_\xi^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2, \qquad s_\eta^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \bar{\eta})^2$$

незміщені слушні оцінки для параметрів $\sigma_1^2$ та $\sigma_2^2$ відповідно. Тому відхилення $\frac{s_\xi^2}{s_\eta^2}$ від $\frac{\sigma_1^2}{\sigma_2^2}$ в середньому менше ніж відхилення $\frac{s_\xi^2}{s_\eta^2}$ від будь-якого іншого числа.

Тому критерій потрібно будувати так: відхиляти гіпотезу $H_0$, якщо $\frac{s_\xi^2}{s_\eta^2}$ значно відрізняється від 1 і приймати $H_0$ у протилежному випадку.

Скористаємось тим, що при $\sigma_1^2 = \sigma_2^2$ статистика $\frac{s_\xi^2}{s_\eta^2}$ має розподіл Фішера з $(n-1, m-1)$ ступенями вільності. Нехай $F_{\gamma,(n-1,m-1)}$ — квантіль рівня $\gamma$ розподілу Фішера з $(n-1, m-1)$ ступенями вільності. Критерій такий: гіпотезу $H_0$ відхиляють, якщо

$$\frac{s_\xi^2}{s_\eta^2} \notin (F_{\frac{\alpha}{2},(n-1,m-1)}, F_{1-\frac{\alpha}{2},(n-1,m-1)})$$

і приймають у протилежному випадку. Рівень значимості критерія $\alpha$.

Оскільки $F_{\alpha,(n,m)} = \frac{1}{F_{1-\alpha,(m,n)}}$, то критичну область можна записати ще так:

$$\left(\frac{1}{F_{1-\frac{\alpha}{2},(m-1,n-1)}}, F_{1-\frac{\alpha}{2},(n-1,m-1)}\right).$$

Для односторонньої альтернативи $\sigma_1^2 > \sigma_2^2$ критерій з рівнем значимості $\alpha$ має вигляд

$$\frac{s_\xi^2}{s_\eta^2} \geq F_{1-\alpha,(n-1,m-1)}.$$

***Випадок, коли математичні сподівання відомі***

Якщо у попередньому випадку параметри $a_1$ та $a_2$ відомі, то використовуємо статистику

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\frac{1}{n}\sum_{i=1}^n (\xi_i - a_1)^2}{\frac{1}{m}\sum_{j=1}^m (\eta_j - a_2)^2},$$

яка має розподіл Фішера з $(n, m)$ ступенями вільності.

При двосторонній альтернативі критерій з рівнем значимості $\alpha$ такий:

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \notin \left(\frac{1}{F_{1-\frac{\alpha}{2},(m,n)}}, F_{1-\frac{\alpha}{2},(n,m)}\right).$$

А при односторонній альтернативі $\sigma_1^2 > \sigma_2^2$ :

$$\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \geq F_{1-\alpha,(n,m)}.$$

У багатьох випадках, при досить великому обсязі вибірки, запропоновані критерії використовують і для генеральних сукупностей, які не мають нормального розподілу. Це грунтується на тому факті, що за центральною граничною теоремою, $\bar{\xi}$ та $\bar{\eta}$ розподілені асимптотично нормально.

# Лекція 4

# Тести про вигляд розподілу

## 4.1 Двовибірковий критерій Колмогорова-Смірнова

Нехай $x_1$, $x_2$,..., $x_n$ — вибірка з неперервно розподіленої генеральної сукупності з функцією розподілу $F(x)$, а $y_1$, $y_2$,..., $y_m$ — вибірка з неперервно розподіленої генеральної сукупності з функцією розподілу $G(x)$. Припустимо, що розподіли обох генеральних сукупностей однакові $(H_0)$ : $F(x) = G(x)$.

Критерієм перевірки гіпотези є статистика

$$D_{mn} = \sup_{-\infty<x<+\infty} |F_m(x) - G_n(x)|,$$

де $F_m(x)$ і $G_n(x)$ — емпіричні функції розподілу обох вибірок.

При достатньо великих $n$ і $m$ статистика

$$D_B = \sqrt{\frac{mn}{m+n}} D_{mn}$$

має розподіл Колмогорова незалежно від розподілів розглянутих генеральних сукупностей. Цей факт і використовують для перевірки гіпотези $H_0$. Якщо $D_B < D_\alpha$, де $D_\alpha$ — квантиль порядку $\alpha$ розподілу Колмогорова, то нульова гіпотеза не суперечить досліджуваним вибіркам (ймовірність помилки $\alpha$).

## 4.2 Критерій $\chi^2$ та його застосування

### 4.2.1 Перевірка гіпотези про вид розподілу

Нехай в результаті експерименту отримали вибірку $\zeta = (\xi_1, ..., \xi_n)$ із генеральної сукупності з невідомим розподілом $\mathbf{F}$. $\mathbf{G}$ - заданий розподіл. Потрібно перевірити гіпотезу $H_0$: $\mathbf{F} = \mathbf{G}$.

Ідея побудови критерію для перевірки гіпотези $H_0$, грунтується на тому, що емпіричний розподіл $\hat{\mathbf{F}}_n$, отриманий за вибіркою $\zeta$ мало відрізняється від справжнього розподілу $\mathbf{F}$. Тому, якщо гіпотеза $H_0$ справедлива, то відхилення $\hat{\mathbf{F}}_n$ від $\mathbf{G}$ мале, інакше - велике.

Міру відхилення $\hat{\mathbf{F}}_n$ від $\mathbf{G}$ будують так: розбивають область значень випадкової величини на скінченну кількість множин $\triangle_i$, $i = 1, 2, ..., r$, які не перетинаються, і за міру відхилення беруть

$$\hat{\chi}_n^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} = \sum_{i=1}^{r} \frac{n}{p_i} \left( \frac{\nu_i}{n} - p_i \right)^2,$$

де $p_i = \mathbf{P}\{\xi_k \in \triangle_i\}$ обчислюють за гіпотетичним розподілом $\mathbf{G}$, а $\nu_i$ - число елементів вибірки, які попали у мно-

жину $\triangle_i$. Множини $\triangle_i$ вибирають так, щоб всі $p_i > 0$.

Якщо справедлива гіпотеза $H_0$: $\mathbf{F} = \mathbf{G}$, то частоти $\frac{\nu_i}{n}$ є слушними і незміщеними оцінками $p_i$ і тому відхилення $\hat{\chi}_n^2$ у цьому випадку мінімальне.

Критерій перевірки гіпотези $H_0$ будують на основі того, що у випадку справедливості $H_0$ розподіл випадкової величини $\hat{\chi}_n^2$, при $n \to \infty$ збігається до розподілу $\chi^2$ з $r-1$ ступенем вільності. Тому при досить великих $n$ за розподіл $\hat{\chi}_n^2$ беруть розподіл $\chi^2$ з $r-1$ ступенем вільності.

Критерій $\chi^2$ з рівнем значимості $\alpha$ полягає у тому, що гіпотезу $H_0$ відхиляють при

$$\hat{\chi}_n^2 > \chi_{1-\alpha,r-1}^2$$

і приймають в іншому випадку.

### 4.2.2 Перевірка гіпотези про вид розподілу, який залежить від невідомих параметрів

Нехай $\zeta = (\xi_1, ..., \xi_n)$ — вибірка із генеральної сукупності з невідомим розподілом $\mathbf{F}$. Гіпотеза $H_0$ полягає у тому, що $\mathbf{F} = \mathbf{G}(\theta)$, $\theta = (\theta_1, ..., \theta_m)$, де розподіл $\mathbf{G}$ визначається параметрами $\theta_1, ..., \theta_m$, які невідомі. Наше завдання, як і у попередньому пункті, відхилити чи ні гіпотезу $H_0$.

У цьому випадку діють так: за методом максимальної вірогідності отримують оцінки $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m$ параметрів і як гіпотетичний розглядають розподіл $\mathbf{G}(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)$. Тоді розподіл відхилення

$$\hat{\chi}_n^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m))^2}{np_i(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)}$$

збігається до розподілу $\chi^2$ з $r-1-m$ ступенем вільності.

У цьому випадку критерій $\chi^2$ з рівнем значимості $\alpha$ полягає у відхиленні гіпотези $H_0$ при

$$\hat{\chi}_n^2 > \chi^2_{1-\alpha,r-1-m}$$

Зауваження. Критерій $\chi^2$ використовує той факт, що розподіл випадкової величини $\frac{(\nu_i - np_i)}{\sqrt{np_i}}$ близький до нормального $N(0,1)$. Тому для всіх множин $\triangle_i$ повинна виконуватись умова $np_i > 10$. Якщо для деяких множин $\triangle_i$ ця умова не виконується, то їх потрібно об'єднати з сусідніми.

### 4.2.3 Перевірка гіпотези про однорідність

Нехай в результаті $k$ серій незалежних випробувань отримали результати $\xi_i = (\xi_{i_1}, \xi_{i_2}, ..., \xi_{i_{n_i}}),\ i = 1, 2, ..., k$. Чи можна вважати ці результати, отриманими спостереженнями над однією і тією ж випадковою величиною, тобто вважати, що закон розподілу від серії до серії не змінюється? Якщо це так, то кажуть що статистичні дані однорідні.

Якщо $\mathbf{F}_i(x)$ — функція розподілу спостережень $i$-ї серії, то ми повинні прийняти чи відхилити гіпотезу однорідності $H_0$ :

$$\mathbf{F}_1(x) \equiv \mathbf{F}_2(x) \equiv ... \equiv \mathbf{F}_k(x).$$

Для побудови критерію, так як і в розділі 4.2.1 розіб'ємо область значень випадкових величин на скінченну кількість множин $\triangle_i,\ i = 1, 2, ..., r$, які не перетинаються. Нехай $\nu_{ij}$ — кількість результатів в $j$-й серії випробувань, які попали в множину $\triangle_i$. Тоді кількість результатів $j$-ої серії $n_j = \sum_{i=1}^r \nu_{ij}$, а загальна кількість спостережень $n = \sum_{j=1}^k n_j = \sum_{j=1}^k \sum_{i=1}^r \nu_{ij}$.

Візьмемо за міру відхилення величину:

$$\hat{\chi}_n^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{n_j\nu_{i\cdot}} - 1\right),$$

де $\nu_{i\cdot} = \sum_{j=1}^{k}\nu_{ij}$.

При $n \to \infty$ величина $\hat{\chi}_n^2$ має граничний розподіл $\chi^2$ з $(r-1)(k-1)$ ступенями вільності.

Отже, гіпотезу $H_0$ відхиляємо, якщо $\hat{\chi}_n^2 > \chi^2_{1-\alpha,(r-1)(k-1)}$ при рівні значимості $\alpha$.

### 4.2.4 Перевірка гіпотези про незалежність випадкових величин

Нехай $\xi$ та $\eta$ — дві дискретні випадкові величини, які можуть набувати значення $x_1, x_2, ..., x_k$ та $y_1, y_2, ..., y_l$ відповідно.

За результатами $n$ спостережень випадкового вектора $\zeta = (\xi, \eta)$ потрібно перевірити гіпотезу $H_0$ : випадкові величини $\xi$ та $\eta$ - незалежні, тобто

$$\mathbf{P}\{\xi = x_i, \eta = y_j\} = \mathbf{P}\{\xi = x_i\}\mathbf{P}\{\eta = y_j\} = p_i q_j,$$

$$1 \le i \le k, \quad 1 \le j \le l.$$

Позначимо $\nu_{ij}$ — кількість спостережень $\zeta$, результатами яких є $(x_i, y_j)$. Тоді результати наших $n$ спостережень можна подати у вигляді таблиці спряженості ознак:

| $\xi\backslash\eta$ | $y_1$ | $y_2$ | ... | $y_l$ | Сума |
|---|---|---|---|---|---|
| $x_1$ | $\nu_{11}$ | $\nu_{12}$ | ... | $\nu_{1l}$ | $\nu_{1\cdot}$ |
| $x_2$ | $\nu_{21}$ | $\nu_{22}$ | ... | $\nu_{2l}$ | $\nu_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_k$ | $\nu_{k1}$ | $\nu_{k2}$ | ... | $\nu_{kl}$ | $\nu_{k\cdot}$ |
| Сума | $\nu_{\cdot 1}$ | $\nu_{\cdot 2}$ | ... | $\nu_{\cdot l}$ | $n$ |

де $\nu_{\cdot j} = \sum_{i=1}^{k} \nu_{ij},\ j = 1, 2, ..., l;\ \nu_{i\cdot} = \sum_{j=1}^{l} \nu_{ij},\ i = 1, 2, ..., k.$

Для перевірки гіпотези $H_0$ використовують критерій $\chi^2$ про вигляд розподілу, що залежить від невідомих параметрів. У нашому випадку невідомі параметри — $p_i$ та $q_j$.

Весь вибірковий простір розбивають на множини, кожна з яких складається лише з однієї точки $(x_i, y_j)$. Міру відхилення емпіричного розподілу $\frac{\nu_{ij}}{n}$ від гіпотетичного $p_i q_j$ визначають так:

$$\hat{\chi}_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\nu_{ij} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j},$$

де $\hat{p}_i$ та $\hat{q}_j$ — оцінки максимальної вірогідності для $p_i$ та $q_j : \hat{p}_i = \frac{\nu_{i\cdot}}{n}$, $\hat{q}_j = \frac{\nu_{\cdot j}}{n}$. Тому

$$\hat{\chi}_n^2 = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\nu_{ij}^2}{\nu_{\cdot j} \nu_{i\cdot}} - 1. \tag{4.1}$$

Оскільки $\sum_{i=1}^{k} p_i = 1$ і $\sum_{j=1}^{l} q_j = 1$, то кількість параметрів, які оцінюють, дорівнює $k - 1 + l - 1 = k + l - 2$. Вибірковий простір розбито на $k \cdot l$ частин. Тому $\hat{\chi}_n^2$ має граничний розподіл $\chi^2$ з $l \cdot k - 1 - (k + l - 2) = (k - 1)(l - 1)$ ступенями вільності.

Отже, гіпотезу $H_0$ відхиляють при рівні значимості $\alpha$, якщо

$$\hat{\chi}_n^2 > \chi_{1-\alpha,(k-1)(l-1)}^2.$$

Зауваження. Якщо $\xi$ та $\eta$ неперервні випадкові величини, то область значення кожної з них розбивають на скінченну кількість проміжків, які відіграють роль $x_i$ та $y_j$.

# Лекція 5

# Непараметричні критерії

Розглянуті раніше критерії перевірки статистичних гіпотез грунтувались на певних припущеннях про розподіли у моделях, які вивчались (у більшості випадків розподіл визначався своїми параметрами). При невиконанні цих припущень (у випадку інших розподілів) критерії, як правило, непридатні. На практиці, при обробці результатів спостережень, розподіл генеральної сукупності буває невідомий, так що застосування методів, розглянутих раніше, може давати помилки. У таких випадках застосовують методи, які не залежать від розподілу генеральної сукупності. Їх називають непараметричними методами.

Непараметричні методи в основному використовують не самі числові значення елементів вибірки, а структурні властивості вибірки (наприклад, відношення порядку між її елементами).

У зв'язку з цим, частина інформації, яка міститься у вибірці, втрачається. Тому, наприклад, потужність непараметричних критеріїв менша ніж у аналогічних параметричних критеріїв. Проте, непараметричні методи застосовують при більш загальних припущеннях та використо-

вують більш прості обчислення.

## 5.1 Критерій знаків

Критерій знаків застосовують для перевірки гіпотези $H_0$ про те, що вибірки $(\xi_1, \xi_2, ..., \xi_n)$ та $(\eta_1, \eta_2, ..., \eta_n)$ із однієї генеральної сукупності, тобто про те, що функції розподілу $\mathbf{F}_\xi(x)$ та $\mathbf{F}_\eta(y)$ двох генеральних сукупностей однакові: $\mathbf{F}_\xi(x) \equiv \mathbf{F}_\eta(x)$ (генеральні сукупності однорідні).

Будемо вважати, що розподіли $\mathbf{F}_\xi$ та $\mathbf{F}_\eta$ абсолютно неперервні, але нам не відомі. Якщо наші вибірки отримані із однорідних генеральних сукупностей, то

$$\mathbf{P}\{\xi_i - \eta_i > 0\} = \mathbf{P}\{\xi_i - \eta_i < 0\} = \frac{1}{2}, \quad i = 1, 2, ..., k.$$

Тут $k$ — кількість ненульових різниць $\xi_i - \eta_i,\ k \leq n$.

Статистика критерію знаків — кількість знаків "+" чи "−" у послідовності знаків різниць $\xi_i - \eta_i,\ i = 1, 2, ..., k$. Далі, для визначеності, будемо брати знак "+".

За умови справедливості гіпотези $H_0$, кількість знаків "+"має біноміальний розподіл з параметрами $p = \frac{1}{2}$ та $k$. Отже, ми отримали задачу перевірки гіпотези $H_0$: $p = \frac{1}{2}$ при альтернативній гіпотезі $H_1$. $H_1$ може бути, як одностороння: $p > \frac{1}{2}$ ($\mathbf{P}\{\xi_i - \eta_i > 0\} > \frac{1}{2}$) чи $p < \frac{1}{2}$ ($\mathbf{P}\{\xi_i - \eta_i < 0\} > \frac{1}{2}$) так і двостороння: $p \neq \frac{1}{2}$ ($\mathbf{P}\{\xi_i - \eta_i > 0\} \neq \frac{1}{2}$).

Нехай $r$ — кількість знаків "+", а $\alpha$ — рівень значимості критерію.

Тоді гіпотезу $H_0$ відхиляють, якщо

$\sum_{i=r}^{k} C_k^i \left(\frac{1}{2}\right)^k \leq \alpha$ (при альтернативі $p > \frac{1}{2}$);

$\sum_{i=0}^{r} C_k^i \left(\frac{1}{2}\right)^k \leq \alpha$ (при альтернативі $p < \frac{1}{2}$);

$\sum_{i=r}^{k} C_k^i \left(\frac{1}{2}\right)^k \leq \frac{\alpha}{2}$ чи $\sum_{i=0}^{r} C_k^i \left(\frac{1}{2}\right)^k \leq \frac{\alpha}{2}$ (при альтернативі $p \neq \frac{1}{2}$).

У багатьох випадках гіпотезу $H_0$ перевіряють, використовуючи статистику Фішера. Гіпотезу $H_0$ відхиляють, якщо

$\hat{F}_1 = \frac{r}{k-r+1} \geq F_{1-\alpha,(2(k-r+1),2r)}$ (альтернатива $p > \frac{1}{2}$);

$\hat{F}_2 = \frac{k-r}{r+1} \geq F_{1-\alpha,(2(r+1),2(l-r))}$ (альтернатива $p < \frac{1}{2}$);

$\hat{F}_1 \geq F_{1-\frac{\alpha}{2},(2(k-r+1),2r)}$ чи $\hat{F}_2 \geq F_{1-\frac{\alpha}{2},(2(r+1),2(l-r))}$ (альтернатива $p \neq \frac{1}{2}$).

## 5.2 Критерій Вілкоксона

Нехай $\bar{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ та $\bar{\eta} = (\eta_1, \eta_2, ..., \eta_m)$ — реалізації незалежних вибірок з неперервних розподілів $\mathbf{F}$ та $\mathbf{G}$ відповідно. Про розподіли $\mathbf{F}$ та $\mathbf{G}$ відомо, що

$$\mathbf{G}(x) = \mathbf{F}(x - \theta),$$

де $\theta$ — невідомий параметр.

Гіпотеза $H_0$ полягає у тому, що $\theta = 0$ ($\mathbf{F} \equiv \mathbf{G}$, тобто генеральні сукупності однорідні).

Розмістимо вибірки $\xi_1, \xi_2, ..., \xi_n$ і $\eta_1, \eta_2, ..., \eta_m$ у спільний варіаційний ряд. Для кожного $\xi_i$ $(\eta_j)$ визначимо його ранг, як номер місця, на якому стоїть $\xi_i$ $(\eta_j)$ у спільному варіаційному ряді. Якщо деякі вибіркові значення збігаються, то їм приписують ранг, який дорівнює середньому арифметичному відповідних місць.

Статистику Вілкоксона $W$ визначають, як суму рангів вибіркових значень вибірки меншого обсягу.

Нехай $n \leq m$, $r_1, r_2, ..., r_n$ — ранги $\xi_1, ..., \xi_n$. Тоді:

$$W = r_1 + r_2 + ... + r_n.$$

Величина $W$ описує міру "змішаності" значень вибірок. Якщо $W$ велике (більшість вибіркових значень $\xi_i$ розміщені праворуч значень $\eta_j$) чи мале (більшість значень $\xi_i$ розміщені ліворуч значень $\eta_j$), то "змішаність" незначна, інакше "гарна".

Мінімально можливе значення статистики $W$ :

$$1+2+...+n=\frac{n(n+1)}{2},$$

максимально можливе:

$$(m+1)+(m+2)+...+(m+n)=\frac{(2m+n+1)n}{2},$$

"середнє":

$$\frac{1}{2}\left[\frac{n(n+1)}{2}+\frac{(2m+n+1)n}{2}\right]=\frac{n(n+m+1)}{2}.$$

Якщо гіпотеза $H_0 : \theta = 0$ ($\mathbf{F} \equiv \mathbf{G}$) справедлива, то $\bar{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ і $\bar{\eta} = (\eta_1, \eta_2, ..., \eta_m)$ — незалежні вибірки з одного розподілу і змішаність "гарна". $W$ тоді близьке до середнього значення $\frac{n(n+m+1)}{2}$. У іншому випадку $W$ істотно відхиляється від $\frac{n(n+m+1)}{2}$.

При заданому рівні значимості $\alpha$ межі $W_{\alpha,n,m}$ та $n(n+m+1)-W_{\alpha,n,m}$, що відділяють значення $W$, які "мало відрізняються" від $\frac{n(n+m+1)}{2}$ і ті які відрізняються істотньо, знаходять за відповідними таблицями.

Отже, критерій Вілкоксона полягає у відхиленні гіпотези $H_0$, якщо

$W < W_{\alpha,n,m}$ (при односторонній альтернативі $\theta > 0$);

$W > n(n+m+1) - W_{\alpha,n,m}$ (при односторонній альтернативі $\theta < 0$);

$W \notin [W_{\frac{\alpha}{2},n,m}; n(n+m+1) - W_{\frac{\alpha}{2},n,m}]$ (при альтернативі $\theta \neq 0$).

Оскільки при $n, m \to +\infty$ $W$ асимптотично нормальна з середнім $\frac{n(n+m+1)}{2}$ і дисперсією $\frac{nm(n+m+1)}{12}$, то можна використовувати наближені значення:

$$W_{\alpha,n,m} \approx \frac{1}{2}n(n+m+1) + d_\alpha\sqrt{\frac{1}{12}nm(n+m+1)},$$

$$n(n+m+1) - W_{\alpha,n,m} \approx \frac{1}{2}n(n+m+1) - d_\alpha\sqrt{\frac{1}{12}nm(n+m+1)},$$

де $d_\alpha$ — квантіль розподілу $N(0,1)$.

## 5.3 Критерій Манна і Уітні

Критерій застосовують для порівняння двох незалежних вибірок обсягу $n_1$ та $n_2$. Перевіряють гіпотезу $H_0$, яка стверджує, що вибірки одержані з однорідних генеральних сукупностей.

Статистику критерію $W$ визначають так. Розмістимо $n_1 + n_2$ значень об'єднаної вибірки в порядку зростання. Кожному елементу одержаного варіаційного ряду покладемо у відповідність його порядковий номер — ранг. Якщо кілька елементів ряду однакові, то кожному з них присвоюють ранг, що дорівнює середньому арифметичному їх номерів.

Нехай $R_1$ — сума рангів елементів першої вибірки, $R_2$ — сума рангів елементів другої вибірки. Обчислимо значення

$$w_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

$$w_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2.$$

Вибіркове значення $w_{\text{В}}$ статистики критерію є менше з чисел $w_1$ та $w_2$ ($W = \min(w_1, w_2)$). В статистичних таблицях наводяться ймовірності $p = P(W < x/H_0)$ при умові, що гіпотеза $H_0$ правильна для вибірок обсягу $n_1$ і $n_2$ ($n_1 \geq n_2$). При двосторонній альтернативній гіпотезі гіпотезу $H_0$ відхиляють, якщо $p \leq \alpha/2$.

Якщо обсяг кожної з вибірок більший, ніж 8, то перевірку гіпотези $H_0$ можна проводити з допомогою статистики

$$Z = \frac{W - \frac{1}{2}n_1 n_2}{\sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}},$$

що має (при умові, що гіпотеза $H_0$ правильна) приблизно стандартний нормальний розподіл $N(0,1)$. В цьому випадку гіпотезу $H_0$ відхиляють на рівні значимості $\alpha$, якщо вибіркове значення $Z_{\text{В}}$ статистики $Z$ задовольняє нерівність (при двосторонній альтернативній гіпотезі)

$$|Z_{\text{В}}| > u_{1-\frac{\alpha}{2}}.$$

## 5.4 Рангова кореляція

Нехай $\bar{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ та $\bar{\eta} = (\eta_1, \eta_2, ..., \eta_n)$ — реалізації вибірок з неперервних розподілів. Кожному значенню $\xi_i$ поставимо у відповідність його ранг $\xi_i'$ у варіаційному ряді. Аналогічно отримуємо і ранг $\eta_i'$.

Вибірковим значенням рангового коефіцієнта кореляції Спірмена $\rho_s$ називають величину

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(\xi_i' - \eta_i')^2}{n(n^2 - 1)}. \tag{5.1}$$

Ранговий коефіцієнт кореляції $\rho_s$, як і звичайний коефіцієнт кореляції, характеризує залежність випадкових величин. Коефіцієнт $r_s$ — непараметрична міра зв'язку.

Гіпотеза $H_0 : \rho = 0$ при альтернативній гіпотезі $H_1 : \rho_s \neq 0$ перевіряють за допомогою статистики

$$T_{n-2} = |r_s|\sqrt{\frac{n-2}{1-r_s^2}}.$$

Якщо гіпотеза $H_0$ правильна, то статистика $T_{n-2}$ має розподіл Стьюдента з $n-2$ ступенями вільності. Оже, при заданому рівні значимості $\alpha$, гіпотезу $H_0$ відхиляють, якщо

$$T_{n-2} > t_{1-\frac{\alpha}{2},n-2},$$

тобто між випадковими величинами є рангова кореляційна залежність.

# Лекція 6

# Аналіз категоризованих даних

Величини виміряні в номінальній шкалі будемо називати категоріальними. Категоріальні величини використовують тільки для якісної класифікації. Це означає, що ці змінні можуть бути виміряні тільки в термінах належності до деяких суттєво різних класів, при цьому неможливо визначити якусь кількісну характеристику класу чи навіть впорядкувати їх. Наприклад, описуючи певну спільність людей, можна розглядати такі характеристики, як стать, колір очей, ставлення до куріння і т.п. Очевидно, що ці змінні мають дещо інший тип, ніж такі, як вік, ріст, маса тіла. Крім того, завжди можна перейти від виміру у більш багатій шкалі до менш багатої. Так неперервні величини можна штучно перетворити в категоріальні, тобто категоризувати їх. Зробити це можна, поділивши множину значень неперервної величини на кілька частин, що не перетинаються, та надавши категоризованій змінній значення, що якимось чином ідентифікує множину, в яку потрапила дана величина. Категоризовані дані часто задають у вигляді частот спостережень, що попали в певні категорії чи класи. В цьому випадку для описання категоризованих да-

них важливу роль відіграє мода — значення з найбільшою частотою.

## 6.1 Гіпотези про розподіл частот

Адаптуємо тест розглянутий в § 4.2.1 до категоризованих даних. Нехай $X$ — деяка категоріальна змінна, яка може набувати значень, що належать до $k$ категорій. Розглянемо гіпотезу $H_0$, яка полягає в тому, що частоти $\nu_i$, з якими окремі категорії зустрічаються серед значень величини $X$, дорівнюють заданим числам $n_i$. Альтернативою до нульової гіпотези будемо вважати гіпотезу, яка полягає в тому, що хоча б одна з частот не збігається з заданим для неї числом.

Розглянемо вибіркові частоти $\widehat{\nu}_i$ та використаємо статистику — міру відхилення їх від теоретичних частот

$$\chi^2 = \sum_{i=1}^{k} \frac{(\widehat{\nu}_i - n_i)^2}{n_i}.$$

Малі значення статистики $\chi^2$ свідчать про несуперечливість нульової гіпотези та статистичних даних. Рівень значимості (ймовірність помилитися при цьому) визначають з умови, що випадкова величина з розподілом $\chi^2(k-1)$ більша за одержане вибіркове значення статистики.

## 6.2 Гіпотези про незалежність ознак

Адаптуємо тест розглянутий в § 4.2.4 до категоризованих даних. Розглянемо дві категоріальні змінні з $k$ та $l$ рівнями ознак, відповідно. Нехай проведено $n$ експериментів, в

яких пара значень ознак з номерами $(i, j)$ зустрічалась $n_{ij}$ раз ($i = \overline{1,k}$, $j = \overline{1,l}$). Нехай $n_{i\cdot}$ — кількість експериментів, в яких було одержано $i$-тий рівень першої ознаки, а $n_{\cdot j}$ — кількість експериментів, в яких було одержано $j$-тий рівень другої ознаки. Обчислимо для кожної пари рівнів ознак число $\widetilde{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$. При умові, що розглянені ознаки незалежні ці числа можна розглядати як очікувані частоти, з якими відповідна пара значень ознак повинна була зустрітися у вибірці. Наступну статистику можна розглядати як міру відмінності між реальними та прогнозованими частотами. Статистика

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \widetilde{n}_{ij})^2}{\widetilde{n}_{ij}}$$

має розподіл $\chi^2$ з $(k-1)(l-1)$ ступенями вільності (при умові, що всі $\widetilde{n}_{ij} \geq 4$).

Отже, гіпотезу $H_0$ про незалежність розглянутих ознак приймають (не суперечить вибірковим даним) на рівні значимості $\alpha$, якщо $\chi^2_{\text{в}} < \chi^2_{1-\alpha}((k-1)(l-1))$, де $\chi^2_{1-\alpha}((k-1)(l-1))$ — квантиль порядку $1-\alpha$ розподілу $\chi^2$ з $(k-1)(l-1)$ ступенями вільності. В іншому випадку гіпотезу $H_0$ потрібно відхилити як таку, що не узгоджується з наявними даними.

## 6.3 Оцінка залежності дворівневих даних

Нехай незалежно проведено дві серії, що містять $n_1$ та $n_2$ випробувань, відповідно. В першій серії подія $A$ відбулася $n_{11}$ раз, а в другій — $n_{21}$ раз. Потрібно перевірити гіпотезу про те, що ймовірність появи події $A$ в обох серіях одна і та

ж, тобто $H_0 : p_1 = p_2$. Результати обох серій можна подати у вигляді таблиці спряженості ознак розміру $2 \times 2$ : Тут

| Серія | Подія | | Сума |
|---|---|---|---|
| | $A$ | $\overline{A}$ | |
| 1 | $n_{11}$ | $n_{12}$ | $n_{1\cdot}$ |
| 2 | $n_{21}$ | $n_{22}$ | $n_{2\cdot}$ |
| Сума | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $n$ |

$n_{12} = n_1 - n_{11}$, $n_{22} = n_2 - n_{21}$, $n_{1\cdot} = n_{11} + n_{12}$, $n_{2\cdot} = n_{21} + n_{22}$, $n_{\cdot 1} = n_{11} + n_{21}$, $n_{\cdot 2} = n_{12} + n_{22}$.

Позначимо через $h_1 = \frac{n_{11}}{n_{1\cdot}}$, $h_2 = \frac{n_{21}}{n_{2\cdot}}$, $h = \frac{n_{\cdot 1}}{n}$. При великих значеннях $n$ та при умові, що найменша з величин $\frac{n_{\cdot i} n_{j\cdot}}{n}$, $i, j = 1, 2$, буде більшою, ніж 5, як статистику для перевірки гіпотези $H_0$ можна використати

$$Z = \frac{h_1 - h_2}{\widetilde{\sigma}_{h_1 - h_2}},$$

де $\widetilde{\sigma}^2_{h_1 - h_2}$ — оцінка дисперсії різниці випадкових величин $h_1$ та $h_2$ обчислена за формулою

$$\widetilde{\sigma}^2_{h_1 - h_2} = h(1 - h)\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right).$$

Якщо гіпотеза $H_0$ правильна, то статистика $Z$ має майже нормальний розподіл $N(0, 1)$. Критична область критерію при рівні значимості $\alpha$ визначається нерівностями

$z_{\text{в}} > u_{1-\alpha}$ при альтернативній гіпотезі $H_1 : p_1 > p_2$,

$z_{\text{в}} < u_{\alpha}$ при альтернативній гіпотезі $H_1 : p_1 < p_2$,

$|z_{\text{в}}| > u_{1-\alpha/2}$ при альтернативній гіпотезі $H_1 : p_1 \neq p_2$.

У випадку, коли результати спостережень такі, що умова $\frac{n_{\cdot i} n_{j\cdot}}{n} > 5$ не виконується для всіх пар індексів, то для перевірки гіпотези $H_0$ використовують критерій $\chi^2$. Гіпотеза $H_0$ еквівалентна гіпотезі про те, що обидві вибірки одержані з однієї генеральної сукупності. Статистика для перевірки гіпотези має вигляд

$$\chi^2_{\text{В}} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.$$

Критичну область на рівні значимості $\alpha$ визначають нерівністю $\chi^2_{\text{В}} \geq \chi^2_{1-\alpha}(1)$, де $\chi^2_{1-\alpha}(1)$ — квантиль порядку $1-\alpha$ розподілу $\chi^2$ з одним ступенем вільності.

Критерій $\chi^2$ можна використовувати при умові, що всі значення $\frac{n_{\cdot i} n_{j\cdot}}{n} > 3$ і $n > 20$. Для малих $n$ при обчисленні $\chi^2_{\text{В}}$ потрібно $n$ замінити на $n-1$; при цьому повинно бути $n_{1\cdot} > 5$ $n_{2\cdot} > \frac{n_{1\cdot}}{3}$.

# Лекція 7

# Дисперсійний аналіз

Дисперсійний аналіз широко використовують як в соціальних, так і технічних дослідженнях. Застосовують його в тих випадках, коли є необхідність з'ясувати вплив різних факторів на значення деякої величини. Причому, фактори переважно мають якісний характер і можуть мати скінчену кількість різних рівнів.

Суть цього методу досліджень полягає в тому, що загальну дисперсію досліджуваної ознаки розбивають на окремі частини, кожна з яких характеризує вплив на ознаку певного конкретного чинника. Велика частина дисперсії, викликана впливом одного з факторів, в загальній дисперсії свідчить про статистично значимий зв'язок між фактором та досліджуваною ознакою.

## 7.1 Однофакторний дисперсійний аналіз

Нехай потрібно вивчити вплив одного фактора на значення деякої величини (досліджуваної ознаки). Припустимо,

що фактор має $k$ рівнів. Розділимо результати експерименту на $k$ груп, згідно з різними рівнями дії фактора.

Нехай під впливом $i$-того рівня фактора одержано $n_i$ значень $x_{ij}$ величини $X$. Будемо вважати, що значення досліджуваної величини можуть бути задані в такому вигляді:

$$x_{ij} = a_i + \varepsilon_{ij},$$

де $a_i$ — вплив даного фактора (невипадкові величини), $\varepsilon_{ij}$ — результат впливу неврахованих факторів. Вважатимемо, що величини $\varepsilon_{ij}$ є реалізаціями центрованої нормально розподіленої випадкової величини з дисперсією $\sigma^2$, тобто $\varepsilon \sim N(0, \sigma^2)$. Якщо фактор не має впливу на величину $X$, то величини $a_i$ рівні між собою.

Таким чином ми маємо $k$ незалежних вибірок (груп), одержаних з $k$ нормально розподілених генеральних сукупностей, які мають, взагалі кажучи, різні математичні сподівання $a_1$, $a_2$..., $a_k$ та однакові дисперсії $\sigma^2$.

Перевіримо гіпотезу про рівність середніх $H_0 : a_1 = a_2 = ... = a_k$. Розглянемо випадок $k > 2$. Коли $k = 2$ простіше використати критерії розглянуті раніше.

Нехай $\overline{x}_i$ — вибіркове середнє $i$-тої вибірки,

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij};$$

$\overline{x}$ — вибіркове середнє об'єднаної вибірки,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{x}_i,$$

де $n$ — загальна кількість спостережень.

Загальна сума квадратів відхилень спостережень від загального середнього значення $\overline{x}$ може бути подана у вигляді

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2$$

Позначивши загальну суму квадратів

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2,$$

суму квадратів відхилень вибіркових середніх $\overline{x}_i$ від загального середнього $\overline{x}$

$$Q_1=\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2,$$

суму квадратів відхилень в середині груп відносно середнього значення в кожній групі

$$Q_2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2,$$

одержимо рівність

$$Q=Q_1+Q_2.$$

Це основна рівність дисперсійного аналізу.

Якщо справджується гіпотеза $H_0$, то статистики $\frac{Q_1}{\sigma^2}$ і $\frac{Q_2}{\sigma^2}$ незалежні та мають розподіли $\chi^2$ з $k-1$ та $n-k$ ступенями вільності. Тому статистики

$$S_1^2=\frac{Q_1}{k-1}\quad\text{і}\quad S_2^2=\frac{Q_2}{n-k}$$

є незміщеними оцінками невідомої дисперсії $\sigma^2$. Оцінка $S_1^2$ характеризує розсіювання групових середніх, а оцінка $S_2^2$ — розсіювання всередині груп, яке обумовлене випадковими варіаціями результатів спостережень (впливом неврахованих факторів). Значну перевагу величини $S_1^2$ над $S_2^2$ можна пояснити відмінністю середніх в групах. Цей факт можна використати для перевірки гіпотези $H_0$.

Розглянемо відношення $\frac{S_1^2}{S_2^2} = F$. Статистика $F$ має розподіл Фішера з $k-1$ та $n-k$ ступенями вільності. Гіпотеза $H_0$ не суперечить (при рівні значимості $\alpha$) результатам спостережень, якщо вибіркове значення $F_{\text{В}}$ статистики $F$ менше, ніж квантиль $F_{1-\alpha}(k-1, n-k)$ порядку $1-\alpha$ розподілу Фішера з $k-1$ та $n-k$ ступенями вільності. Значення $F_{1-\alpha}(k-1, n-k)$ можна знайти з таблиць. Якщо ж $F_{\text{В}} \geq F_{1-\alpha}(k-1, n-k)$, то гіпотезу $H_0$ відхиляють і потрібно вважати, що серед середніх $a_1$, $a_2$..., $a_k$ є хоча б два різні. При цьому в $\alpha \cdot 100\%$ випадків буде допущено помилку (відхилено правильну гіпотезу).

## 7.2 Двофакторний дисперсійний аналіз

Нехай необхідно визначити вплив двох факторів $A$ і $B$ на певну ознаку $X$. Для цього потрібно, щоб значення ознаки були одержані при всіх різних рівнях факторів $A$ і $B$ та при їх одночасному впливі на ознаку $X$. Припустимо, що одержано $n$ значень досліджуваної ознаки при кожному з $p$ рівнів фактора $A$ і кожному з $q$ рівнів фактора $B$. Позначимо ці зачення через $x_{ijk}$ $(i = \overline{1,p}, j = \overline{1,q}, k = \overline{1,n})$.

Введемо такі характеристики:

1. Середні значення

$$\overline{x}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n} x_{ijk}$$

середні значення ознаки при впливі кожної пари рівнів обох факторів;

$$\overline{y}_i = \frac{1}{nq} \sum_{j=1}^{q} \sum_{k=1}^{n} x_{ijk}$$

середні значення ознаки при кожному рівні фактора $A$;

$$\overline{z}_j = \frac{1}{np} \sum_{i=1}^{p} \sum_{k=1}^{n} x_{ijk}$$

середні значення ознаки при кожному рівні фактора $B$;

$$\overline{x} = \frac{1}{npq} \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n} x_{ijk}$$

загальне середнє значення ознаки $X$.

2. Суми квадратів відхилень та виправлені дисперсії

$$Q_1 = np \sum_{i=1}^{p} (\overline{y_i} - \overline{x})^2, \quad S_1^2 = \frac{Q_1}{p-1}$$

зумовлені впливом фактора $A$ на ознаку $X$;

$$Q_2 = nq \sum_{j=1}^{q} (\overline{z_j} - \overline{x})^2, \quad S_2^2 = \frac{Q_2}{q-1}$$

зумовлені впливом фактора $B$ на ознаку $X$;

$$Q_3 = \sum_{i=1}^{p} \sum_{j=1}^{q} (\overline{x}_{ij} - \overline{y}_i - \overline{z}_j + \overline{x})^2, \quad S_3^2 = \frac{Q_3}{(p-1)(q-1)}$$

зумовлені впливом на ознаку $X$ обох факторів $A$ і $B$;

$$Q_4 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \overline{x}_{ij})^2, \quad S_4^2 = \frac{Q_4}{pq(n-1)}$$

зумовлені впливом на ознаку інших неврахованих факторів.

Статистики

$$F_A = \frac{S_1^2}{S_4^2}, \quad F_B = \frac{S_2^2}{S_4^2}, \quad F_{AB} = \frac{S_3^2}{S_4^2}$$

мають розподіли Фішера з $p-1$ та $pq(n-1)$, $q-1$ та $pq(n-1)$, $(p-1)(q-1)$ та $pq(n-1)$ ступенями вільності, відповідно.

Порівнявши вибіркові значення розглянутих статистик з відповідними критичними значеннями (квантилями розподілу Фішера), зможемо зробити висновок про гіпотезу $H_0$. А саме, вибравши деякий рівень значимості $\alpha$, будемо стверджувати, що відсутність впливу фактора $A$ на значення ознаки $X$ не підтверджується статистичними даними, якщо $F_A^* \geq F_{1-\alpha}(p-1, pq(n-1))$ ($F_A^*$ — вибіркове значення статистики $F_A^*$). При цьому ймовірність припуститися помилки дорівнює $\alpha$. Аналогічні висновки можна зробити і для впливу фактора $B$ та обох факторів разом.

# Лекція 8

# Кластерний аналіз

Класифікація об'єктів та явищ зовнішнього світу є однією з властивостей людського розуму. Кожне слово в мові означає певний клас предметів, які чимось відрізняються від інших. У більшості випадків таке розрізнення відбувається на інтуітивному рівні. Проте, коли ми маємо справу з новими явищами чи намагаємося згрупувати вже відомі об'єкти в нові класи за якимось ознаками, то в цьому разі не можна повністю покладатись на інтуіцію. Для завдань точної класифікації потрібний певний науковий апарат і методологія, які й дає кластерний аналіз.

Формально ми маємо справу з $n$ об'єктами (індивідами, ознаками, характеристиками, явищами тощо) кожен з яких описується множиною з $p$ його характеристик. Якщо позначити значення $i$-ї характеристики для $k$-го об'єкта $x_{ki}$, то ми матимемо матрицю

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Потрібно об'єкти, які відповідають рядкам матриці, певним чином розбити на групи (кластери). За допомогою кластерного аналізу визначають кількість кластерів та які об'єкти в який кластер мають потрапити.

Перше питання, яке постає при кластеризації деякої сукупності об'єктів — як вимірювати їх подібність чи відмінність між собою. Тобто, для кожної пари векторів характеристик $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$ об'єктів $i$ та $j$ ми повинні розглянути функцію подібності (відмінності) $f(x_i, x_j)$. Значення $s_{ij} = f(x_i, x_j)$ будемо називати коефіцієнтом подібності (відмінності). Розглянемо деякі важливі коефіцієнти, які часто зустрічаються на практиці.

## 8.1 Коефіцієнти відмінності

У багатьох випадках зручно вимірювати відмінність у їх характеристиках. Ось декілька коефіцієнтів відмінності, які зустрічаються найчастіше:

1. $\sum_{k=1}^{p}(x_{ik} - x_{jk})^2$;
2. $\sum_{k=1}^{p}|x_{ik} - x_{jk}|$
3. $\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{x_{ik} - x_{jk}}$;
4. $\frac{\sum_{k=1}^{p} x_{ik}x_{jk}}{(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2)^{\frac{1}{2}}}$

Можливо, один з найчастіше вживаних коефіцієнтів — це перший. З математики добре відомо — це не що інше, як квадрат евклідової відстані між двома точками. Евклідову відстань (перший із коефіцієнтів відмінності) звичайно застосовують у тому разі, коли вважають, що різні характеристики індивіда некорельовані між собою. За наявно-

сті кореляції, яка визначається коваріаційною матрицею $S$, віддалі між векторами $x_i$ та $x_j$ визначають формулою

$$d_{ij} = (x_i - x_j)S^{-1}(x_i - x_j)'.$$

## 8.2 Міжгрупові відстані

До цього моменту ми весь час порівнювали між собою два об'єкти: знаходили їх подібність, відмінність, уводили відстань між ними. У кластерному аналізі досить часто доводиться розглядати подібні характеристики не між окремими об'єктами, а між деякими групами. Наведемо декілька прикладів міжгрупових характеристик.

Одними з найбільш простих і часто вживаних міжгрупових характеристик є ті, які обчислюють за допомогою характеристик окремих пар об'єктів груп. Так, якщо ми виберемо найменшу з відстаней між парами об'єктів груп, то отримуємо характеристику, яку називають відстанню найближчих сусідів груп. Якщо ж візьмемо найбільшу з відстаней між парами об'єктів, то отримаємо відстань найвіддаленіших сусідів груп. Ці характеристики часто застосовують в кластерному аналізі й про них ми будемо говорити детальніше далі.

Досить часто застосовують певні середні показники групових відстаней. Один із таких методів полягає в обчисленні арифметичного середнього характеристик усіх пар об'єктів (де перший з однієї групи, а другий — з іншої). Інший полягає в тому, що для кожної групи обчислюють вектор, компоненти якого є середніми для відповідних характеристик по групі

$$\bar{x}_A = (\bar{x}_{A,1}; \bar{x}_{A,2}; ...; \bar{x}_{A,p}),$$

$$\bar{x}_B = (\bar{x}_{B,1}; \bar{x}_{B,2}; ...; \bar{x}_{B,p}).$$

Відстань між групами $A$ та $B$ тоді обчислюють так:

$$d_{AB} = \sqrt{\sum_{i=1}^{p} (\bar{x}_{A,i} - \bar{x}_{B,i})^2}.$$

Якщо характеристики об'єктів у групах залежні та $S$ — коваріаційна матриця міжгрупових середніх, то тоді часто застосовують таку характеристику відстані:

$$D^2_{AB} = (\bar{x}_A - \bar{x}_B) S^{-1} (\bar{x}_A - \bar{x}_B)'.$$

Рідше використовують коефіцієнт подібності між групами

$$S_{AB} = \cos\left[\frac{1}{n_A n_B} \sum_{i \in A} \cos^{-1} s_{ij}\right],$$

де $n_A,\ n_B$ — кількість об'єктів у групах $A$ та $B$ відповідно, $s_{ij}$ — коефіцієнт подібності об'єктів $i$ та $j$.

## 8.3 Кластеризація об'єктів

Ієрархічна кластерна техніка полягає в тому, що будують деяку послідовність кластерних розбиттів множини, таку, що, з одного боку, розбиття складається з кластерів, кожен з яких містить тільки один елемент множини, а з іншого боку, маємо лише один кластер, що містить усю множину. Відповідно до напряму, в якому будують ієрархію, розглядають або агломеративні або подрібнювальні методи.

При кожному методі постає питання, який із кроків ланцюжка агломерацій чи подрібнень вважати оптимальним. Кластери, що отримані на оптимальному кроці й становлять потрібне розбиття.

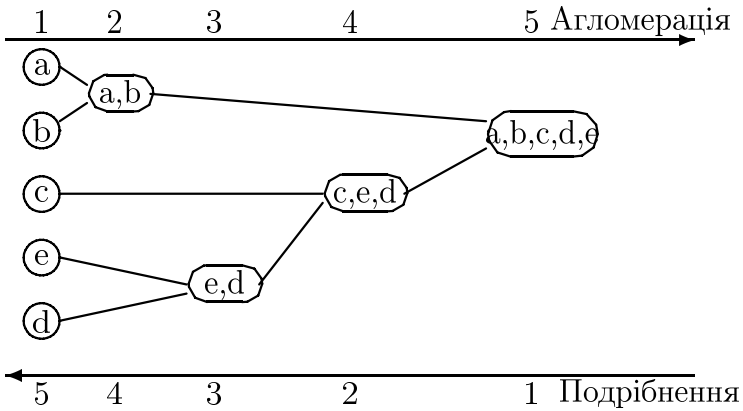


Рисунок 8.1: Дендрограма.

Графічно процес ієрархічної класифікації зображають так званими дендрограмами, які характеризують зростання чи подрібнення, що відбувається на кожному кроці. Приклад такої дендрограми для множини з п'яти об'єктів подано на рис. 8.1.

Залежно від напрямку, в якому ми розглядаємо створення кластерів, маємо або процес агломерації, або процес подрібнення.

### 8.3.1 Агломеративні методи

Нехай ми маємо множину з $n$ об'єктів. Метод полягає в утворенні послідовності $P_n, P_{n-1}, ..., P_1$ розбиттів об'єктів. Перше розбиття $P_n$ складається з $n$ кластерів, кожен з яких містить лише один об'єкт. Останнє розбиття складається з єдиного кластера з усіма об'єктами. Правило утворення проміжних розбиттів описується таким алгоритмом:

Нехай при деякому розбитті $P_k$ маємо кластери $C_1$,

$C_2, ..., C_l$. Тоді:

1. Знаходимо серед них пару найближчих кластерів $C_i$ та $C_j$. Об'єднуємо $C_i$ та $C_j$ у новий кластер, а старі клатери $C_i$ та $C_j$ знищуємо.

2. Якщо кількість кластерів дорівнює 1, процес зупиняють. Інакше повертаємось до кроку 1.

Є багато різних методів визначення подібності, відмінності чи відстані між кластерами. Тому й процес агломеризації може відбуватись по-різному.

### Метод найближчих сусідів

При такому методі відстань між двома кластерами визначають як найменшу відстань у парах, де один елемент з одного кластера, а другий – з іншого (див. рис. 8.2).

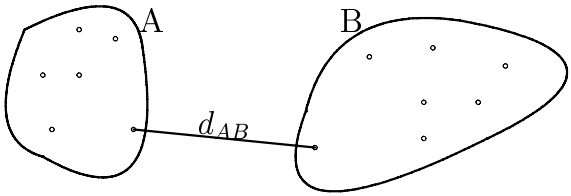


Рисунок 8.2: Найближчі сусіди.

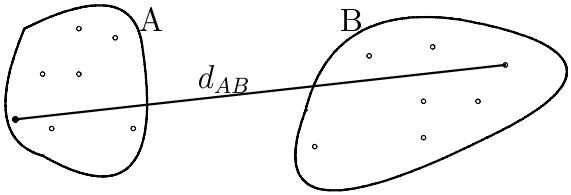


Рисунок 8.3: Найвіддаленіші сусіди.

### Метод найвіддаленіших сусідів

У цьому методі відстань між кластерами визначають, як найбільшу серед пар, де один елемент із першого кластера, а інший — з другого (див. рис. 8.3).

### Метод середніх групових відстаней

У цьому методі відстань між двома кластерами визначають як середнє арифметичне відстаней усіх пар індивідів, по одному з кожного кластера.

### Кластеризація за центрами

Цей метод полягає в тому, що для кожної групи знаходять середнє арифметичне характеристик її елементів. Відстань між отриманими векторами вважають відстанню між групами.

### Метод Уорда

Метод полягає в тому, що при переході від одного розбиття до іншого об'єднують такі два кластери, що відбуває-

ться мінімальне збільшення загальної втрати інформації. За втрату інформації для однієї групи беруть звичайно середньоквадратичне відхилення, а для кількох груп — суму всіх групових відхилень.

### 8.3.2 Вибір кількості кластерів

На практиці нас часто не цікавить побудова повної ієрархічної кластеризаційної послідовності. Нам просто потрібно вибрати одне чи два з розбиттів і вказати, яке найкраще підходить до реальної ситуації.

Зрозуміло, що в багатьох випадках вибір найкращого розбиття обумовлений не лише математичними властивостями об'єктів, але й природою конкретної прикладної задачі. Проте, можна вказати декілька корисних загальних рекомендацій.

Одна з них така: якщо будувати дендрограму, указуючи відстані, на яких відбувається утворення нових кластерів, то часто великі зміни будуть свідчити про правильний вибір кластерів.

Наприклад, розглянемо дендрограму на рис. 8.4.

Тут спостерігається великий розрив при переході від розбиття на два кластери до об'єднання всіх об'єктів у один кластер. Природно спробувати розглянути як оптимальне розбиття індивідів на два кластери.

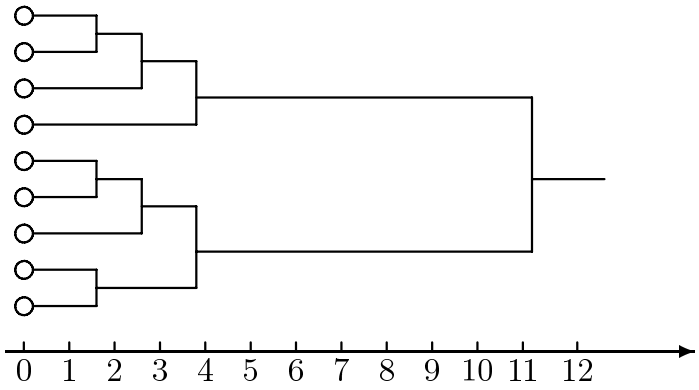


Рисунок 8.4: Дендрограма з відстанями.

# Лекція 9

# Дискримінантний аналіз

## 9.1 Загальні положення

Методи дискримінантного аналізу є статистичним апаратом для вивчення відмінностей між групами об'єктів, кожен з яких зображено багатовимірним вектором. Цей метод використовують у психології для розробки тестів, для передбачення успішності; в соціології – для вивчення поведінки електорату; для класифікації дитячої поведінки та в інших ситуаціях.

Задача дискримінантного аналізу полягає у наступному: якщо набір об'єктів (описаних багатьма показниками — багатовимірними векторами) уже поділено на групи, то потрібно встановити "правило" віднесення до однієї з відомих груп нового об'єкта. В цьому відмінність дискримінантного та кластерного аналізу, який було розглянуто раніше. В останньому відсутній наперед визначений поділ на групи, таке розбиття саме шукають за характеристиками набору об'єктів.

Метод є корисним для прогнозування соціальних явищ, він дає змогу вивчати відмінності між двома та більше

групами за кількома змінними одночасно.

"Дискримінантний аналіз" — це загальний термін, який об'єднує методи інтеграції міжгрупових відмінностей та методи класифікації спостережень за групами. При тлумаченні отримуємо відповідь на питання: чи можливо відрізнити одну групу від іншої використовуючи набір змінних; які з цих змінних найбільш інформативні. Методи, пов'язані з класифікацією, пов'язані з отриманням набору функцій (можливо, однієї), які забезпечують можливість віднести кожен об'єкт до тієї чи іншої групи. Такі функції називають *дискримінантними.*

Змінні (показники, характеристики), які враховують при віднесенні об'єкта до групи, називають дискримінантними змінними. Ці змінні повинні вимірюватись за інтервальною шкалою, чи шкалою відношень. Як правило, кількість об'єктів переважає кількість дискримінантних змінних хоча б на два.

***Обмеження для дискримінантних змінних:***

- жодна змінна не повинна бути лінійною комбінацією інших змінних;
- недопустимо, щоб для будь-якої пари змінних коефіцієнт кореляції дорівнював 1 чи $-1$;
- часто використовують припущення про збіг коваріаційних матриць різних груп;
- закон розподілу для кожної групи багатовимірний, нормальний (це дозволяє отримати значення ймовірностей належності до груп).

## 9.2 Функції втрат

Бажано так підбирати класифікуючі (дискримінантні) функції, щоб ймовірність неправильної класифікації була якомога меншою.

Позначимо через $c(j|i)$ функцію втрат, яка визначає вартість втрат від віднесення об'єкта $i$-го класу до $j$-го класу (очевидно, $c(j|i) = 0$). Через $m(j|i)$ позначимо кількість таких неправильних віднесень. Тоді сумарні втрати при класифікації n об'єктів та k класах можна записати, як

$$C_n = \sum_{i=1}^{k} \sum_{j=1}^{k} c(j|i) m(j|i).$$

Якщо останню рівність поділити на кількість класифікованих об'єктів n, то отримаємо норму втрат при заданому n. Перейдемо до границі при $n \to \infty$, отримуємо:

$$C = \lim_{n\to\infty} \frac{C_n}{n} = \lim_{n\to\infty} \sum_{i=1}^{k} \sum_{j=1}^{k} c(j|i) \frac{m(j|i)}{n_i(n)} \cdot \frac{n_i(n)}{n} =$$

$$= \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} c(j|i) p(j|i)$$

(використовуємо збіжність за ймовірністю).

Тут $n_i(n)$ — частота $i$-го класу; $\pi_i$ — ймовірність вибору об'єкта $i$-го класу із загальної сукупності (так звана апріорна ймовірність, або питома вага $i$-го класу); $p(j|i)$ — ймовірність віднести об'єкт класу $i$ до класу $j$. Якщо вважати, що втрати $c(j|i)$ однакові для всіх $i, j = \overline{1,k}$, $(c(j|i) = c_0 = const)$, то

$$C = C_0 \left(1 - \sum_{i=1}^{k} \pi_i p(j|i)\right).$$

Величина $1 - \sum\limits_{i=1}^{k} \pi_i p(j|i)$ визначає ймовірність неправильної класифікації.

## 9.3 Процедура класифікації

Задача полягає у віднесенні кожного з n класифікованих об'єктів (які являють собою m-вимірний вектор ознак) до одного з k класів, які не перетинаються. Задані k класів представлено k вибірками, які називають навчаючими.

Отже, $\overline{X_i} = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$ — це $i$-тий об'єкт, який потрібно класифікувати.

$z_1, z_2, ..., z_k$ — навчаючі вибірки, кожна з яких є непорожнім набором $m$-вимірних векторів (об'єктів), про які точно відомо, що вони належать до одного класу.

Розглядаємо набір об'єктів, які мають бути класифіковані, як вибірку з генеральної сукупності з щільністю ймовірності $f(x) = \sum\limits_{j=1}^{k} \pi_j f_j(x)$, де

$\pi_i$ — апріорна ймовірність появи елемента з класу $j$;

$f_j(x)$ — щільність розподілу в $j$-му класі.

Дискримінантна функція (класифікатор, вирішальна процедура) $\delta(x)$ може набувати тільки цілі додатні значення 1, 2, ..., k, причому ті об'єкти $X$, при яких вона набуває значення $j$, зараховуємо до $j$-го класу.

Таким чином, вся множина можливих векторів $X$ ділиться на k підмножин, що попарно не перетинаються $S =$

$(s_1, s_2, ..., s_k)$. Таким чином і вирішальна процедура може бути задана розбиттям.

Процедуру класифікації (дискримінантна функція) називають оптимальною (байєсівською), якщо вона забезпечує мінімум втрат, порівняно з іншими процедурами класифікації.

Виявляється, що оптимальна процедура класифікації $S^{(\text{ОПТ})}$, при якій втрати будуть оптимальними визначається так:

$$S^{(\text{ОПТ})} = (s_1^{(\text{ОПТ})}, s_2^{(\text{ОПТ})}, \ldots, s_k^{(\text{ОПТ})});$$

$$S_j^{(\text{ОПТ})} = \left\{ x : \sum_{i=\overline{1,k}, i \neq j} \pi_i f_i(x) c(j|i) = \right.$$

$$\left. = \min_{1 \leq l \leq k} \sum_{i=\overline{1,k}, i \neq l} \pi_i f_i(x) c(l|i) \right\},$$

тобто, спостереження $X_r$ $(r = 1, 2, ..., m)$, буде віднесене до класу $j$ тоді, коли середні питомі втрати від внесення його саме до цього класу виявляться мінімальними, порівняно з аналогічними втратами при віднесенні його до будь-якого іншого класу.

При однакових втратах $c(j|i)$ правило набуває простого виду. Спостереження $X_r$ буде віднесене до класу j тоді, коли

$$\pi_j f_j(X_r) = \max_{1 \leq l \leq k} \pi_l f_l(X_r).$$

Для того, щоб скористатись наведеними виразами невідомі ймовірності $\pi_j$ замінюють їх статистичними оцінками, побудованими на основі навчаючих вибірок:

$$\hat{\pi}_j = \frac{n_j}{n},$$

де $n_j$ — обсяг $j$-ї навчаючої вибірки; $n$ — загальний обсяг усіх вибірок.

Але можливі випадки, коли оцінки $\pi_j$ знаходять з інших міркувань, базуючись на закономірностях конкретної предметної області.

Якщо всі класи задають однаковий закон розподілу, але з відмінними параметрами, тобто два класи відрізняються лише величиною параметра, то такий вид класифікації називають *параметричним* дискримінантним аналізом. У цьому випадку в якості оцінки невідомих функцій $f_j(x, v)$ використовують функції $f_j(x, \hat{v}_j)$, де $v$ — параметр; $\hat{v}_j$ — статистична оцінка невідомого параметра $v$, обчислена за $j$-ю навчаючою вибіркою.

*Непараметричний* дискримінантний аналіз не передбачає знання функцій $f_j(x), j = \overline{1,k}$. Тут використовують непараметричні оцінки для функцій.

Канонічна дискримінантна функція є лінійною комбінацією дискримінантних змінних і задовольняє певні умови:

$$f_{ij} = u_0 + u_1 x_{1ij} + u_2 x_{2ij} + ... + u_m x_{mij}, \qquad (9.1)$$

де $f_{ij}$ — значення канонічної дискримінантної функції для $i$-го об'єкта в групі k; $u_i$ — коефіцієнти.

Коефіцієнти $u_i$ для першої функції вибирають так, щоб її середні значення для різних класів якомога більше відрізнялись один від одного. При виборі коефіцієнтів для другої функції використовують те ж правило з додатковою вимогою, щоб значення другої функції були некорельованими зі значеннями першої. Аналогічно третя функція має

бути некорельована з першими двома. Максимальна кількість дискримінантних функцій, які можна отримати таким способом дорівнює $\min(k-1, m)$.

## 9.4 Геометричне тлумачення

Кожен об'єкт (спостереження) можна трактувати як точку $m$-вимірного евклідового простору, коли координати точок є значеннями відповідних показників для заданого об'єкта. Якщо класи дійсно відрізняються за цими показниками, то вони утворять виражені згустки точок. Для кожного класу можна обчислити геометричні центри, які ще називають центроїдами. Центроїди характеризують класи, є їх "типовими представниками". Щоб вивчати взаємне розташування центроїдів, достатньо обмежитись розмірністю на одиницю меншою від кількості класів. Отже, задачу класифікації тепер можна розглядати в $(k-1)$-вимірному просторі, натягнутому на центроїди.

Початок координат поміщають в точку нульових значень показників. Першу вісь направляють так, щоб середні значення класів розділялись у більшій мірі, ніж для інших напрямків. Другу вісь направляють теж з умовою максимального розрізнення класів з додатковою умовою ортогональності до першої осі. Аналогічно будують наступні осі.

Фактично вираз (9.1) задає перетворення m-вимірного простору дискримінантних змінних в $q$-вимірний простір канонічних дискримінантних функцій. Кожній осі відповідає своє співвідношення виду (9.1). Для даного спостереження $f_{ij}$ тлумачать як координату об'єкта в просторі канонічних дискримінантних функцій.

У випадку, коли кількість дискримінантних змінних $m$

менше від кількості класів, максимальна кількість функцій $q$ дорівнює $m$. Тоді вже не відбувається перетворення з простору з більшою розмірністю в простір з меншою розмірністю, проводиться тільки заміна координат.

## 9.5 Параметричний дискримінантний аналіз. Випадок нормального розподілу класів

Нехай клас задано $m$-вимірним нормальним розподілом з вектором середніх значень $a_j$ та коваріаційною матрицею $\sum$. Зауважимо, що коваріаційна матриця спільна для всіх класів.

Обчислюють оцінки $\hat{a}_j = (\hat{a}_j^1, \hat{a}_j^2, ...., \hat{a}_j^m)$ та $\hat{\sum} = (\hat{\sigma}_{lp})$, $l = \overline{1,m}$, $p = \overline{1,k}$ за вибірками. Використання методу максимальної вірогідності дає вигляд оцінок

$$\hat{a}_j^l = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}^{(l)}, \quad l = \overline{1,m}, \quad j = \overline{1,k}; \tag{9.2}$$

$$\hat{\sigma}_{lp} = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji}^{(l)} - \hat{a}_j^{(l)})(x_{ji}^{(p)} - \hat{a}_j^{(p)}), \tag{9.3}$$

$$l, p = \overline{1,k}, \quad n = \sum_{j=1}^{k} n_j.$$

Класифікуюче правило зараховує спостереження $X$ до $j$-ї групи, якщо

$$\left[X - \frac{1}{2}(\hat{a}_{j0} + \hat{a}_j)\right]^T \hat{\sum}^{-1} (\hat{a}_{j0} - \hat{a}_j) \geq \ln \frac{\pi_j}{\pi_{j0}} \tag{9.4}$$

для всіх $j = \overline{1,k}$.

Правило, задане співвідношенням (9.4), має місце для випадку однакових значень втрат.

Для двох груп ($k = 2$) і однакових апріорних ймовірностей ($\pi_1{=}\pi_2 = 0,5$) правило класифікації наступне: *спостереження $X$ зараховують до першого класу тоді і тільки тоді, коли*

$$\left[X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2)\right]^T \hat{\sum}^{-1} (\hat{a}_1 - \hat{a}_2) \geq 0, \qquad (9.5)$$

*а до другого класу — у всіх інших випадках.*

Для одновимірного випадку ($m = 1$) нормальних спостережень і двох груп, деяке спостереження $X$ зараховують до першої групи, якщо

$$(X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2))(\hat{a}_1 - \hat{a}_2) \geq 0. \qquad (9.6)$$

## 9.6 Нелінійний дискримінантний аналіз

На практиці зустрічаються випадки, коли лінійні дискримінантні функції не достатні для проведення класифікації. Наприклад, у випадку двох дискримінантних ознак та двох груп, навчаючі вибірки розташовані, як на рисунку 9.1.

У цьому випадку ніяка лінійна дискримінантна функція виду $a_0 + a_1x_1 + a_2x_2$ (задає пряму лінію) не придатна для того, щоб задати розподіл простору $S = (s_1, s_2)$, що відповідає розмежуванню груп. Проте можна задати в якості вирішального правила таке: об'єкти, які знаходя-

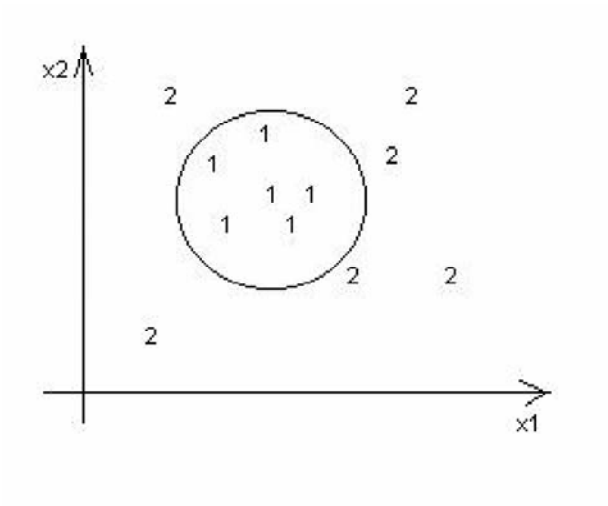


Рисунок 9.1:

ться всередині кола, зараховують до першої групи, а об'єкти, які знаходяться зовні кола — до другої.

Описана ситуація досить часто зустрічається в наукових дослідженнях, коли дані мають багатовимірний нормальний розподіл.

Розглянемо один із методів нелінійного дискримінантного аналізу — метод найближчих сусідів. Нехай потрібно віднести спостереження $X = (x_1, x_2)$ до однієї з 3 груп, див.рис. 9.2.

Для цього обчислюють всі відстані від об'єкта $X$ до елементів навчаючих вибірок. Серед усіх елементів вибирають $k$ таких, які знаходяться найближче до об'єкта $X$. Серед відібраних визначають представників якого класу найбільше. До цього класу і зараховують спостереження $X$. Величину $k$ встановлюють до початку виконання процедури з таких міркувань: при надто маленьких $k$ велика

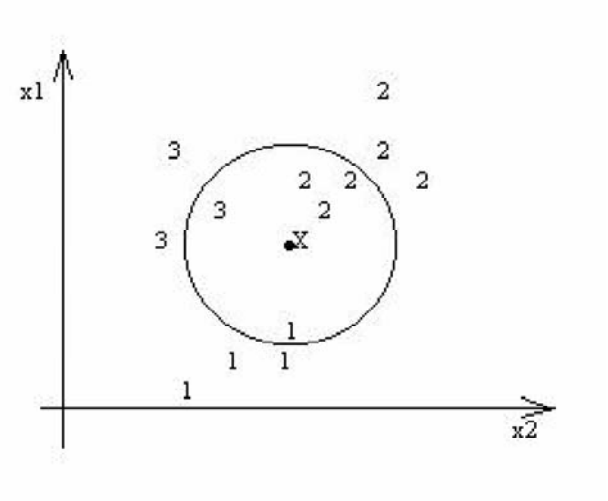


Рисунок 9.2:

ймовірність помилки за рахунок малої кількості відібраних елементів; а при надто великих $k$ на класифікацію можуть впливати елементи, що знаходяться дуже далеко і не мають ніякого відношення до цього спостереження. Рекомендована величина $k \approx \ln(n)$, де $n$ — загальний обсяг навчаючих вибірок.

В якості відстаней між об'єктами використовують евклідову відстань $d(X^i, X^k) = \sqrt{\sum\limits_{j=1}^{m} (x_j^i - x_j^k)^2}$, відстань $d = \sum\limits_{j=1}^{m} \mid x_j^i - x_j^k \mid$, або відстань Махалонобіса $d = \sqrt{\sum\limits_{j=1}^{m} \frac{(x_j^i - x_j^k)^2}{D_j}}$, де $D_j$ — вибіркова дисперсія ознаки. Останню відстань використовують для врахування різниці у розподілі окремих показників (компонент).

# Лекція 10

# Канонічний аналіз

Метод канонічної кореляції призначений для аналізу залежностей між двома наборами змінних. Тоді як обчислення попарних кореляцій між змінними дозволяє встановити залежності між окремими парами змінних, цей метод дозволяє виявляти залежність між двома наборами в цілому. Наприклад, дослідник у галузі освіти може оцінити залежність між навичками з трьох навчальних дисциплін та оцінками з п'яти шкільних предметів. Соціолог може дослідити залежність між прогнозами соціальних змін, які друкують у трьох виданнях, та реальними змінами, які відображаються п'ятьма статистичними показниками. Медик може вивчати залежність між різними несприятливими факторами та появою групи симптомів захворювання. У всіх випадках ми маємо дві множини змінних, і метою дослідження є виявлення взаємозв'язку між цими множинами

***Власні значення.*** Для обчислення канонічних коренів знаходять власні значення матриці попарних кореляцій. Ці значення виражають частку дисперсії, яка пояснюється кореляцією між відповідними канонічними змінни-

ми. Частку обчислюють відносно дисперсії канонічних змінних, тобто зважених сум за двома наборами змінних. Знаходять стільки власних значень, яка найменша кількість змінних у двох наборах даних.

***Послідовне обчислення власних значень.*** В результаті виконання процедури послідовно обчислюються власні значення. Спочатку обчислюють ваги, котрі максимізують кореляцію між зваженими сумами по двох множинах змінних і знаходиться відповідне їм значення першого кореня. Далі обчислюють наступну пару канонічних змінних, які мають максимальну кореляцію і не корелюють з попередніми парами з наступним обчисленням значення канонічного кореня.

***Канонічні кореляції.*** Корені квадратні з отриманих власних значень можна трактувати як коефіцієнти кореляції. Ці корені стосуються канонічних змінних, тому їх називають канонічними кореляціями. Відповідно до власних значень послідовно добуті канонічні кореляції утворюють спадну послідовність. Не тільки найбільші кореляції, але й наступні допускають змістовне тлумачення.

***Значущість коренів.*** Канонічні корені оцінюють один за одним у порядку спадання величини. Для подальшого аналізу залишають лише значущі корені. Існують різні думки дослідників стосовно послідовної перевірки, проте найчастіше перевірка значимості проводиться саме так.

***Канонічні ваги.*** Після встановлення значимості коренів виникає проблема їх тлумачення. Оскільки кожен канонічний корінь являє собою дві зважені суми, що відповідають двом наборам даних, то ці ваги трактують аналогічно до часткових кореляцій. Ці ваги називають канонічними вагами. Канонічні ваги тлумачать аналогічно до вагових коефіцієнтів факторів. Вважають, що чим біль-

ша вага за абсолютним значенням, тим більший внесок кожної змінної в значення канонічної змінної. Розгляд канонічних ваг дозволяє з'ясувати, як конкретні змінні в кожній множині впливають на зважену суму, тобто канонічну змінну. Канонічні ваги можуть використовуватись для обчислення значень канонічних змінних. Для цього достатньо скласти вхідні змінні з відповідними ваговими коефіцієнтами.

***Факторна структура.*** Іншим способом тлумачення канонічних коренів є розгляд звичайних кореляцій між канонічними змінними (або факторами) і змінними з кожної множини. Ці кореляції також називають канонічними навантаженнями факторів. Вважають, що змінні, сильно корельовані з канонічною змінною у великій мірі "пояснюються"нею. Цей підхід трактування канонічних змінних схожий на метод факторного аналізу.

***Факторна структура та канонічні ваги.*** Канонічні значення відповідають унікальному внеску кожної змінної у зважену суму або канонічну змінну. Навантаження канонічних факторів відображають повну кореляцію між відповідними змінними та зваженою сумою. Можливі такі ситуації, що при близьких до нуля канонічних вагах відповідні навантаження змінних дуже великі, або навпаки, при великих канонічних вагах навантаження малі. Звісно такі випадки важко тлумачити. Проте ця ситуація може виникати при наявності двох дуже пов'язаних, майже дублюючих змінних. При обчисленні ваг для зважених сум по кожній множині до цієї суми буде включено тільки одну з цих двох змінних. Якщо більша вага буде приписана одній із змінних, то внесок іншої змінної можна вважати несуттєвим. При цьому звичайні кореляції між існуючими сумарними значеннями двох канонічних

змінних (тобто навантаження факторів), то вони можуть виявитись суттєвими в обох факторів.

***Дисперсія.*** Коефіцієнти канонічної кореляції відповідають кореляції між зваженими сумами по двох множинах даних. Вони не відображають інформації про те, яку частину мінливості (дисперсії) кожен канонічний корінь пояснює в змінних.

Інформацію про частку дисперсії можна отримати з навантажень канонічних факторів. Ці навантаження являють собою кореляції між канонічними змінними та початковими змінними у відповідній множині. Піднесені до квадрату кореляції будуть відображати частку дисперсії, що пояснюється кожною змінною. Для кожного кореня можна обчислити середнє значення цих часток. При цьому отримують середню частку мінливості поясненої в цій множині на основі відповідної змінної.

***Надлишковість.*** Канонічна кореляція при піднесенні до квадрату дає частку дисперсії, загальної для сум по кожній множині (канонічній змінній). Якщо помножити цю частку на частку добутої дисперсії, то отримують міру надлишковості множини змінних, тобто величину, яка відображає, наскільки надлишкова одна множина змінних, якщо задана інша множина. Так можна обчислювати надлишковість першої множини змінних при заданій другій множині, а також надлишковість другої множини змінних при заданій першій множині. Для отримання загального коефіцієнта надлишковості додають надлишковості по всіх (значущих) коренях.

***Практична значущість.*** При великих розмірах вибірки канонічні кореляції невеликого розміру, наприклад, 0,3, можуть виявитись статистично значущими. Для обчислення надлишковості цей коефіцієнт підносять до квадра-

ту. Отримуємо незначну величину, яка свідчить про незначну частку мінливості змінних. Це слід враховувати при з'ясуванні того, наскільки реальна мінливість в одній множині змінних пояснюється другою множиною.

***Припущення.*** Наведемо ряд припущень, врахування яких важливе для отримання достовірних результатів.

Застосування критеріїв для перевірки значимості канонічної кореляції базується на припущенні, що змінні у вибірці мають багатовимірний нормальний розподіл.

Рекомендують використовувати достатньо великі вибірки для отримання достовірних оцінок навантажень канонічних факторів. Деякі автори рекомендують забезпечити в 20, а то і в 40 – 60 разів більше спостережень, ніж кількість досліджуваних змінних. Хоча, як показує практика, при значних кореляціях між даними навіть малі розміри вибірки ( наприклад, $n = 50$) дозволяють у більшості випадків виявити ці кореляції.

***Викиди.*** Наявність викидів може здійснити значний вплив на величину коефіцієнтів кореляції. При збільшенні обсягу вибірки вплив невеликої кількості викидів нівелюється. Рекомендують перед проведенням процедури виявити значні викиди, наприклад, за допомогою діаграми розсіяння.

***Погано обумовлені матриці.*** Вимагають, щоб змінні в обох множинах не були цілком надлишковими. Наприклад, при включенні однієї і тієї ж змінної двічі в одну з множин отримується надлишковість, при якій незрозуміло, яку ж вагу приписати цій змінній. Крім того, при надлишковості спостерігається сильна корельованість між спостереженими змінними, тоді проблематичним є обчислення відповідної оберненої матриці, що цілком порушує процедуру обчислення канонічної кореляції. Такі кореля-

ційні матриці називають погано обумовленими.

***Використання зважених сум.*** Замість розгляду звичайних сум по множинах корисно розглядати зважені суми, щоб ваги, приписані окремим доданкам, відповідали реальній структурі змінних.

# Лекція 11

# Факторний аналіз

Метод факторного аналізу присвячений дослідженню структури зв'язків між змінними. Нехай емпіричні дані подано у вигляді прямокутної матриці розмірами $m \times n$:

$$\begin{pmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nm} \end{pmatrix}$$

Це числові показники, які відповідають n об'єктам (рядки) та m досліджуваним показникам (стовпці). Змінними назвемо m вектор-стовпців. Якщо спостерігаємо взаємну кореляцію між певними змінними, то логічно припустити, що існує деякий фактор, який впливає одночасно на всі ці змінні. Іншими словами, трактуємо фактор, як "причину" одночасних змін групи показників.

Суть даного методу зводиться до того, що зміни відносно великої кількості досліджуваних ознак пояснюють впливом меншої кількості факторів, які безпосередньо не вимірюють і є "прихованими", "латентними". Кількість факторів суттєво менша від кількості змінних $(f_1, f_2, ..., f_k)$,

$k < m$. Фактори є загальними для всіх змінних. Виділяють фактори, які, як правило, є некорельованими.

Залежність змінних від факторів може бути лінійною або іншого виду. Вважають, що кожна із змінних $X_i$, $i = \overline{1,m}$, крім того, що залежить від факторів $f_1, f_2, ..., f_k$, залежить також від деякої випадкової (специфічної для даної змінної) компоненти $u^{(i)}$. Компонента $u^{(i)}$ містить ту частину інформації про змінну $X_i$, яка не пояснюється впливом факторів $f_1, f_2, ..., f_k$.

Метою застосування даного методу є виділення та змістовне тлумачення латентних загальних факторів, кількість яких суттєво менша від кількості спостережених змінних, одночасно прагнучи мінімізувати залежність змінних від своїх специфічних компонент. Хотілось би виділити кілька факторів, які б досить повно описували модель.

У такій постановці задача факторного аналізу полягає у пониженні розмірності моделі: замість великої кількості показників модель описують невеликою кількістю факторів.

## 11.1 Лінійна модель

Як і раніше розглядаємо прямокутну матрицю емпіричних спостережень:

$$\begin{pmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nm} \end{pmatrix}$$

Тут об'єктам відповідають рядки, а досліджуваним показникам (змінним) — стовпці. $X_i$, $i = \overline{1,m}$ — змінні; $x_j = (x_{j1}, x_{j2}, ..., x_{jm})$ — вектори, що відповідають об'єктам.

Вважаємо, що змінні $X_1, X_2, ..., X_m$ — центровані (від кожного значення, яке спостерігають, відняли середнє, що означає перенесення початку координат в "центр" набору даних). Вважаємо, що змінні $X_1, X_2, ..., X_m$ залежать від факторів $f_1, f_2, ..., f_k$ лінійно:

$$X_i = q_{i1}f_1 + q_{i2}f_2 + ... + q_{ik}f_k + u_i, \quad i = \overline{1,m}. \qquad (11.1)$$

$u_i$ — специфічна стохастична компонента, яка визначає ту частку змінної $X_i$, яка не пояснюється дією загальних факторів. Припускаємо, що вектор $\overline{U} = (u_i)$, $i = \overline{1,m}$ має $m$-вимірний нормальний розподіл з нульовим середнім, його компоненти є попарно незалежними. Коваріаційна матриця $V = E(vv^T)$ має діагональний вигляд з діагональними елементами $v_{ii} = Du_i,\ i = \overline{1,m}$. Для кожного об'єкта (спостереження)

$$X_j = QF_j + U_j. \qquad (11.2)$$

(11.2) можна записати як:

$$X = QF + V, \qquad (11.3)$$

де $X = (X_1, X_2, ..., X_m)^T$, $F = (f_1, f_2, ..., f_k)$, $V = (u_1, u_2, ..., u_m)$, $Q = (q_{ij})$, $i = \overline{1,m}$, $j = \overline{1,k}$.

Вектор $F$ переважно вважають випадковим вектором, що має $k$-вимірний нормальний розподіл з нульовим середнім, хоча часам трактують і як вектор невідомих детермінованих параметрів.

У факторній моделі (11.1) вектор спостережень $X$ — нормально розподілений $m$-вимірний (як лінійна комбінація двох нормально розподілених випадкових векторів $F$ та $U$).

З рівностей (11.2) і (11.3) отримуємо:

$$\mathbf{E}X_i = 0, \quad \begin{cases} \sigma_{ii} = \sum\limits_{l=1}^{k} q_{il} + v_{ii} \\ \sigma_{ij} = \sum\limits_{l=1}^{k} q_{il} q_{jl} \end{cases}, \quad i,j = \overline{1,m} \qquad (11.4)$$

або, в матричній формі

$$\mathbf{E}X = 0, \quad \sum QQ^{\mathrm{T}} + V.$$

Тут $\mathbf{E}$ означає математичне сподівання, $\sum = (\sigma_{ij})_{i,j=\overline{1,m}}$ — коваріаційна матриця.

(11.2) і (11.3) формально збігаються з аналітичними залежностями множинної регресії. Принципова відмінність факторного аналізу полягає в тому, що тут фактори є латентними, вони явно не спостерігаються, не визначаються чисельно, як це відбувається у множинній регресії.

## 11.2 Існування та однозначність моделі

Виявляється, що не для будь-якого набору вихідних показників $X = (X_1, X_2, ..., X_m)$ можна вказати задану кількість загальних факторів $f_1,\ f_2, ..., f_k$, які б пояснювали наявну кореляцію між показниками. Не кожна коваріаційна матриця $\sum$ допускає зображення у вигляді (11.4), а отже, не кожен вектор спостережень допускає тлумачення моделі факторного аналізу. Крім того, якщо при заданих кількостях $m$ та $k$ і заданій коваріаційній матриці $\sum$ можлива побудова моделі факторного аналізу, визначення самих факторів $F = (f_1, f_2, ..., f_k)$ (а відповідно і матриці $Q = (q_{ij})$) не єдине.

Окремі випадки апріорних співвідношень, за яких модель однозначно ідентифікується, такі:

1. Розв'язок $(Q, V)$ системи (11.4) належить класу таких матриць $Q$ та $V$, для яких матриця $Q^T V Q$ має діагональний вигляд, причому діагональні елементи її різні і впорядковані за спаданням.

2. Розв'язок $(Q, V)$ такий, що $Q^T Q$ — діагональна матриця, причому всі діагональні елементи різні і впорядковані за спаданням.

3. Розв'язок $(Q, V)$ шукають серед матриць $Q$, які для наперед заданої матриці $B = (b_{ij})_{i=\overline{1,m}, j=\overline{1,k}}$ задовольняють умову $B^T Q = D$ (всі наддіагональні елементи матриці $D$ нульові).

Остання умова доречна у випадках, коли відома деяка апріорна інформація про відсутність зв'язку певної кількості показників від загальних факторів.

## 11.3 Алгоритм методу

Одне з перших питань, які виникають при побудові факторної моделі — це кількість факторів. Тобто, яка найменша кількість факторів дозволяє пояснити кореляції між показниками, які спостерігають. Встановлюють цю кількість за допомогою статистичного критерію для перевірки значимості розбіжності між певною моделлю та набором даних. При відсутності апріорних даних звертаються до однофакторної моделі з наступною перевіркою значимості розбіжності. Якщо розбіжність статистично значуща, то

оцінюють модель з ще одним додатковим фактором і знову застосовують критерій. І так процес продовжують до тих пір, доки розбіжність буде визнана незначною, тобто розбіжність буде пояснюватись випадковістю вибірки.

Самі фактори конструюють за коваріаційною (кореляційною) матрицею. Основна математична ідея грунтується на відшуканні власних чисел та власних векторів редукованої коваріаційної (кореляційної) матриці. Редукованою кореляційною матрицею називають кореляційну (коваріаційну) матрицю із загальностями на головній діагоналі (в якості загальностей використовують квадрати відповідних множинних коефіцієнтів кореляції). У нашому випадку для визначення власних чисел та векторів потрібно розв'язати матричне рівняння $RW = \lambda W$, де $R$ — редукована кореляційна матриця, $W$ — шуканий власний вектор, $\lambda$ — шукане власне число. Сума власних чисел дорівнює кількості змінних, а добуток дорівнює детермінанту кореляційної матриці. Крім того, перше (найбільше) власне число являє собою величину дисперсії, яка відповідає певній осі $m$-вимірного простору, друге та наступні власні числа відповідають дисперсії вздовж інших осей цього простору. В якості факторів можливо розглядати ці знайдені вектори. Якщо поділити перше власне число на m (кількість змінних), то отримаємо частку дисперсії, що відповідає даному напрямку (першому фактору). Аналогічно знаходимо відповідну частку для інших факторів. Фактори розглядаємо у порядку спадання власних чисел (а отже і частки дисперсії).

При знаходженні факторів використовують метод найменших квадратів, який тут полягає у мінімізації залишкової кореляції після виділення визначеної кількості факторів та оцінки міри відповідності (сума квадратів відхи-

лень) коефіцієнтів кореляції, які обчислені та спостерігають.

Алгоритм в загальних рисах такий:

**На першому кроці** припускають, що кількість факторів – k (можна розпочати з $k = 1$). Для встановлення величини k використовують також критерії, які будуть розглянуті пізніше.

**На другому кроці** оцінюють загальності. Для кожної змінної в якості такої оцінки використовують квадрат множинного коефіцієнта кореляції між відповідною змінною та сукупністю усіх інших змінних. Також може використовуватись найбільший за абсолютною величиною коефіцієнт кореляції у відповідному рядку змінної кореляційної матриці.

**На третьому кроці** виділяють $k$ факторів, для яких обчислені коефіцієнти кореляції як найкраще (в сенсі мінімізації суми квадратів відхилень) наближають спостережені кореляції.

**На четвертому кроці** знову проводять оцінку загальностей, причому використовують матрицю факторного відображення, отриману на попередньому етапі.

Процес повторюють, доки покращення стане неможливим. Описаний алгоритм відомий під назвою "Метод головних факторів з ітераціями по загальностях".

Може використовуватись також метод мінімальних залишків Хармана, який є теж ітераційний. В цьому методі критерієм зупинки слугує критерій хі-квадрат.

Метод максимальної вірогідності теж спрямований на відшукання факторної моделі, яка б якнайкраще пояснювала спостережені кореляції. Тут вважають, що розподіл змінних багатовимірний нормальний. Задача зводиться до оцінки значень факторних навантажень генеральної сукупності, за яких при заданих припущеннях функція вірогідності для розподілу елементів кореляційної матриці максимальна. Метод фукціонує в припущенні, що дані, які спостерігають, — це вибірка з генеральної сукупності, яка точно відповідає $k$-факторній моделі.

Може також використовуватись критерій знаходження факторних навантажень, при яких загальні фактори і змінні, які спостерігають, знаходяться в канонічній кореляції, тобто коефіцієнт кореляції між ними максимальний.

Інший критерій — визначення факторних навантажень, за яких детермінант матриці залишкових кореляцій максимальний.

Для реалізації названих критеріїв, як правило, використовують ітераційні схеми.

Всі варіанти методу максимальної вірогідності зводяться до розв'язку характеристичного рівняння $\det(R'' - \lambda \mathrm{I}) = 0$, де $R'' = U^{-1}R'U^{-1}$, $R'$ – редукована кореляційна матриця. На відміну від методу найменших квадратів в обчислювану на кожному кроці оцінку загальностей з більшою вагою входять кореляції із змінними, що мають меншу специфічність $(u_i)$.

## 11.4 Критерії визначення кількості факторів

1. З методами максимальної вірогідності та найменших квадратів найчастіше використовують критерій хі-квадрат. Як показує досвід, це дає верхню оцінку кількості факторів. Тому після відповідних обертань деякі другорядні фактори (за величиною частки їх дисперсій) варто усунути.

2. Критерії, які базуються на власних числах. Залишають фактори з власними числами, більшими 1 (Кайзер). При цьому використовують кореляційну матрицю. Хоча цей критерій носить евристичний характер він був перевірений на модельних даних. Крім того, вважають (Харман), що потрібно припинити виділення спільних факторів, коли сума власних чисел перевищить суму оцінок загальностей.

3. Критерій, який грунтується на величині частки описаної дисперсії. Критерій визначається часткою дисперсії останнього фактора (фактори розташовують за спаданням частки дисперсії). Наприклад, це може бути 1%, 5% чи 10%.

4. Критерій відсіювання Каттелла. Розглядають графічне зображення власних чисел кореляційної матриці. Будують ламану з координатами $(k_i, \lambda_k)$, $k = \overline{1, m}$, де $\lambda_k$ — власні числа кореляційної матриці, впорядковані за спаданням. Виділення закінчують на факторі, після якого досліджувана залежність наближається до прямої, майже горизонтальної лінії (див. рис. 11.1).

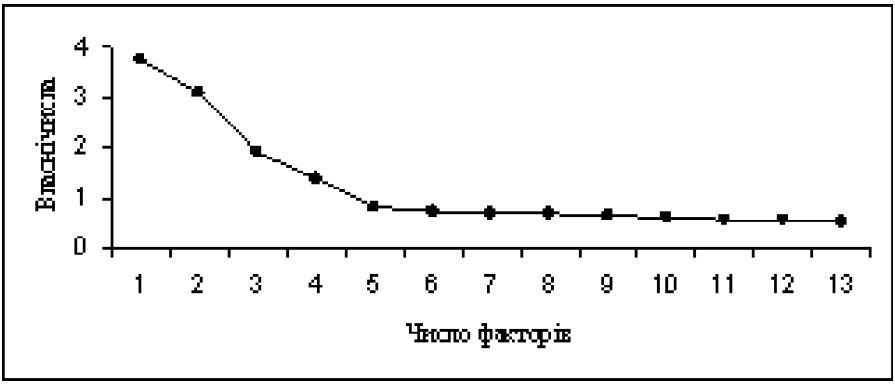


Рисунок 11.1:

5. Критерій інтерпретовності та інваріантності. Пропонують використовувати таку кількість факторів, що узгоджується з усіма наведеними критеріями. А остаточне рішення повинне грунтуватися на змістовному тлумаченні в предметній області.

## 11.5 Методи обертання

Застосовуючи описані методи, отримують набір ортогональних (незалежних) факторів, впорядкованих за спаданням їхнього внеску в загальну модель. Ортогональність та впорядкованість є штучними обмеженнями, привнесеними в модель для однозначності розв'язку.

У типовому випадку змінні будуть мати високі факторні навантаження більше, ніж по одному фактору. Факторні навантаження, що стосуються одного фактора, набуватимуть різних знаків. Хотілось би позбутись цих недоліків, що утруднюють змістовне тлумачення моделі.

Математично розв'язок $(Q, V)$ системи (11.4) можна шукати лише з точністю до обертання системи координат. То ж виникає питання: чи не можна досягти прозорого тлумачення для іншого набору факторів, отриманого обертанням. Метою всіх обертань є отримання найбільш простої факторної структури.

Зауважимо, поняття простоти неоднозначне. Існує кілька підходів до цього поняття. В одному з них основною вимогою до простої структури є наявність хоча б одного нульового елемента в кожному рядку матриці факторних навантажень.

Також бажано, щоб кожен стовпчик матриці факторних навантажень мав не менше нулів, ніж факторів.

У кожного із стовпців будь-якої пари стовпців має бути кілька нулів в тих позиціях, де в іншому стовпці вони ненульові, це гарантує можливість розрізнити вторинні осі.

Якщо кількість загальних факторів перевищує 4, і в кожній парі стовпців деяка кількість нульових навантажень в одних і тих же рядках, то це дає можливість поділити змінні на групи, що не перетинаються.

Для кожної пари стовпців матриці факторних навантажень має бути якомога менше значних навантажень, що відповідають одним і тим же рядкам. Тоді буде забезпечена мінімізація факторної складності змінних.

Існує 3 різні підходи до проблеми обертання:

**Перший підхід** — графічний. Якщо у $k$-вимірному просторі факторів спостерігають яскраво виражені групи змінних, скупчення, то має зміст провести нові осі через ці скупчення.

**Другий підхід** використовує аналітичні методи. Проводять ортогональне або косокутне обертання згідно

до певного критерію.

**Третій підхід** передбачає знаходження такого розв'язку (матриці факторних навантажень), який є найближчим до заданої матриці. Задана матриця враховує вимоги до факторної структури.

### 11.5.1 Методи ортогонального обертання

Якщо факторна складність змінної більше від одиниці (змінна має значні факторні навантаження більше, ніж для одного фактора), то варіація квадрата всіх факторних навантажень для цієї змінної характеризує складність моделі для цієї змінної. При фіксованій кількості факторів і заданих загальностях, дисперсія квадратів факторних навантажень максимальна, якщо загальність є одним з цих квадратів навантажень, а всі інші квадрати — нулі (що означає залежність змінної лише від одного фактора).

Критерій *квартимакс* $Q$ спрямований на обертання осей з метою максимізації дисперсії квадратів факторних навантажень. Мірою простоти тут є величина $Q = \sum_{i=1}^{m} \sum_{j=1}^{k} q_{ij}$.

Критерій *варімакс* $V$ має на меті спрощення опису факторів. В ньому максимізують дисперсію квадратів навантажень фактора

$$V = \left[ \sum_{j=1}^{k} m \sum_{i=1}^{m} q_{ij}^4 - \sum_{j=1}^{k} (\sum_{i=1}^{m} q_{ij}^2)^2 \right] / n^2.$$

Практика використання критеріїв квартимакс і варімакс показує, що останній дає кращу роздільність факторів.

Можна отримати узагальнений критерій $\alpha Q+\beta V = M$, де $\alpha$ та $\beta$ — вагові навантаження, або ж

$$\sum_{j=1}^{k}\sum_{i=1}^{m} q_{ij}^4 - \gamma \sum_{j=1}^{k} (\sum_{i=1}^{m} q_{ij}^2)^2/n = M,$$

де $\gamma = \beta/(\alpha+\beta)$.

При $\gamma = 0$ отримують критерій *квартимакс* $Q$; при $\gamma = 1$ — *варімакс* $V$. Критерій, який отримують при $\gamma = k/2$ називають *еквімакс*, при $\gamma = 0,5$ — *біквартимакс*.

### 11.5.2 Методи косокутного обертання

Якщо відмовитись від вимоги незалежності факторів (ортогональності), то це збільшує кількість можливих розв'язків і, очевидно, дає більше можливостей для знаходження простої структури факторної моделі.

Методи пошуку не обов'язково ортогональних факторів шляхом косокутного обертання називають *облімін*. Серед цих методів найбільш поширений *квартимін*, який схожий до ортогонального методу *квартимакс*, тільки не висувається вимога ортогональності факторів.

## 11.6 Вибіркова адекватність факторної моделі

Для вирішення питання адекватності факторної моделі по відношенню до заданого набору змінних використовується спеціальний критерій — "міру вибіркової адекватності" ($MBA$) :

$$MBA = \frac{\sum\limits_{j \neq k} \sum r_{jk}}{\sum\limits_{j \neq k} \sum r_{jk}^2 + \sum\limits_{j \neq k} \sum g_{ik}^2},$$

де $r_{ij}$ — коефіцієнти кореляції, які спостерігають, $g_{ij}$ — елементи матриці $Q = SR^{-1}S$ , тут $R$ — кореляційна матриця, $S = (\mathrm{diag}R^{-1})^{1/2}$.

Коефіцієнт $MBA$ може набувати значення від 0 до 1. Критерій набуває значення 1 тоді і тільки тоді, коли кожна змінна може бути повністю виражена через інші. Якщо $MBA \geq 0{,}9$, то це відмінний рівень адекватності, якщо $MBA \geq 0{,}8$ — хороший, $MBA \geq 0{,}7$ — середній, $MBA \geq 0{,}6$ — посередній, $MBA \leq 0{,}5$ — неприйнятний.

Величина $MBA$ збільшується при збільшенні кількості змінних, зменшенні кількості загальних факторів, збільшенні обсягу спостережень, збільшенні середнього значення коефіцієнтів кореляції.

Для факторної моделі, отриманої методом максимальної вірогідності, розглядається коефіцієнт надійності $\rho$. На практиці частіше використовують його асимптотичне наближення

$$\rho \approx 1 - \frac{E_1 - 1}{E_2 - 1}, \qquad E_1 = \sum_{i \neq j} \sum (r_{ij})^2 / df_k;$$

$$E_2 = \sum_{i \neq j} \sum (r_{ij})^2 / \left[1/2n(n-1)\right],$$

де $r_{ij}$ — частинні коефіцієнти кореляції без впливу факторів, $df_k$ — кількість ступенів вільності, $df_k = 1/2[(n-r)^2 - (n+r)]$.

# Рекомендована література


[1] Айвазян С.А., Мхитарян В.С. *Теория вероятностей и прикладная статистика.* –М.:ЮНИТИ-ДАНА, 2001.

[2] Айвазян С.А., Енюков И.С., Мешкалин Л.Д. *Прикладная статистика: Основы моделирования и первичная обработка данных.* Справочное издание под ред. Айвазяна С.А. –М.: Финансы и статистика, 1983.

[3] Айвазян С.А., Енюков И.С., Мешкалин Л.Д. *Прикладная статистика: Исследование зависимостей.* Справочное издание под ред. Айвазяна С.А. –М.: Финансы и статистика, 1985.

[4] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешкалин Л.Д. *Прикладная статистика: Классификация и снижение размерности.* Справочное издание под ред. Айвазяна С.А. –М.: Финансы и статистика, 1989.

[5] Алексахин С.В. и др. *Прикладной статистический анализ данных. Теория. Компьютерная обработка. Области применения.* В 2-х кн. –М.: ПРИОР, 2002.

[6] Андерсен Т. *Введение в многомерный статистический анализ.* –М.: Физматгиз, 1963.

[7] Афифи А., Эйзен С. *Статистический анализ. Подход с использованием ЭВМ.* –М.: Мир, 1982.

[8] Дж.-О. Ким, Ч.У.Мьюллер, У.Р.Клекка и др. *Факторный, дискриминантный и кластерный анализ.* – М.:Финансы и статистика, 1989.

[9] Крамер Г. *Математические методы статистики.* – М.: Мир, 1975.

[10] Турчин В.М. *Математична статистика в прикладах і задачах.* –К.: НМК В, 1993.

[11] Тюрин Ю.Н., Макаров А.А. *Статистический анализ данных на компьютере.* –М.: Инфра, 1998.