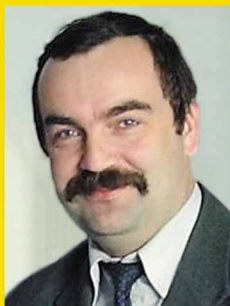




Мамчич Тетяна Іванівна — завідувач кафедри вищої математики та інформатики Волинського державного університету імені Лесі Українки. Закінчила механіко-математичний факультет Київського університету імені Т. Шевченка. Кандидат фізико-математичних наук за спеціальністю “Теорія ймовірностей та математична статистика”, доцент. Керівник лабораторії з проблем прикладної статистики. Автор більше 40 наукових публікацій.



Оленко Андрій Якович — доцент кафедри теорії ймовірностей та математичної статистики Київського національного університету імені Т. Шевченка та кафедри математики національного університету “Києво-Могилянська Академія”. Закінчив механіко-математичний факультет Київського університету імені Т. Шевченка. Кандидат фізико-математичних наук за спеціальністю “Теорія ймовірностей та математична статистика”. Учасник міжнародних проектів із статистичних застосувань у фінансах, страхуванні, комунікаційних системах, вивченні рибних запасів. Автор більше 90 публікацій, в тому числі монографії і ряду навчальних посібників.



Осипчук Михайло Михайлович — доцент кафедри вищої математики Івано-Франківського національного технічного університету нафти і газу. Закінчив механіко-математичний факультет Київського університету імені Т. Шевченка. Кандидат фізико-математичних наук за спеціальністю “Теорія ймовірностей та математична статистика”. Автор більше 20 наукових публікацій присвячених, зокрема, статистичним методам обробки даних в медицині, застосуванням математичних методів до задач видобування нафти і газу. Автор ряду навчальних посібників.



Шпортюк Володимир Григорович — доцент кафедри фінансів факультету економічних наук Національного університету “Києво-Могилянська Академія”. Закінчив механіко-математичний факультет Київського університету імені Т. Шевченка. Кандидат фізико-математичних наук за спеціальністю “Теорія ймовірностей та математична статистика”. Автор 30 наукових публікацій.

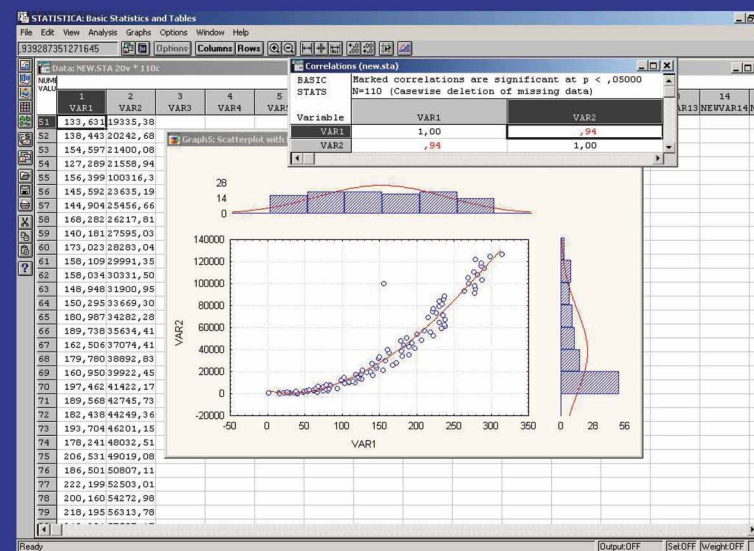
Автори будуть вдячні за поради, що допоможуть удосконалити посібник.
Контактна електронна адреса: olenk@univ.kiev.ua



СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ПАКЕТОМ STATISTICA

Т. І. Мамчин, А. Я. Оленко
М. М. Осипчук, В. Г. Шпортюк

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ПАКЕТОМ STATISTICA



Т. І. Мамчин
А. Я. Оленко
М. М. Осипчук
В. Г. Шпортюк



Національний Університет



"Києво-Могилянська Академія"

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ З ПАКЕТОМ STATISTICA

Т. І. Мамчич

*Волинський Державний Університет ім. Л. Українки,
(Луцьк)*

А. Я. Оленко

*Київський Національний Університет ім. Т. Шевченка,
Університет "Києво-Могилянська Академія",
(Київ)*

М. М. Осипчук

*Івано-Франківський Технічний Університет Нафти і Газу,
(Івано-Франківськ)*

В. Г. Шпортюк

*Університет "Києво-Могилянська Академія",
(Київ)*



Видавнича фірма "Відродження"
2006

УДК 519.25 (075)

ББК 22.17я7

М 22

Мамчич Т.І., Оленко А.Я., Осипчук М.М., Шпортюк В.Г.

М22 Статистичний аналіз даних з пакетом STATISTICA. Навчально-методичний посібник. – Дрогобич: Видавнича фірма "Відродження", 2006. – 208 с., з іл.

ISBN 966-538-161-X

У посібнику розглянуто найчастіше вживані методи прикладного статистичного аналізу даних. Кожен розділ посібника відповідає певній групі методів і містить опис прикладної задачі, статистичних засобів її дослідження, необхідне теоретичне обґрунтування та приклади застосування до реальних даних за допомогою пакету STATISTICA. Посібник розрахований на студентів кваліфікаційного рівня "магістр" з соціальних наук, "спеціаліст" з природничих наук, може бути корисним для аспірантів, викладачів та науковців, які застосовують статистичні методи аналізу. Бібліографія – 30 назв.

УДК 519.25 (075)

ББК 22.17я7

Рецензент: доктор фіз.-мат. наук, проф. *Мішура Ю. С.*

Рекомендовано до друку Вченою радою Національного університету "Киево-Могилянська Академія" 25 червня 2005 р., протокол № 6(7)

Funding for this book was given in part by the Curriculum Resource Center ("CRC") at the Central European University ("CEU"), whose programs are partially funded by the Higher Education Support Program ("HESP"). The opinions expressed herein do not necessarily express the views of CEU

Typesetting: this book was created by the authors using L^AT_EX

ISBN 966-538-161-X

© Т. Мамчич, 2006.

© А. Оленко, 2006.

© М. Осипчук, 2006.

© В. Шпортюк, 2006.

Передмова

Можна з упевненістю стверджувати, що практично кожне наукове дослідження пов'язане з вивченням результатів обстежень, спостережень, експериментів чи вимірювань. Дослідження даних займають вагоме місце як у соціальних, так і в природничих науках. В аналізі емпіричних даних одним із основних інструментів є статистичні методи. У навчальних планах підготовки фахівців із соціології передбачено курси “Кількісні методи в соціологічних дослідженнях”, “Обробка, аналіз та узагальнення соціологічної інформації”, “Соціальна статистика”, фахівців з економіки – “Статистика”, “Прикладний аналіз даних”. Відповідні дисципліни включено і в навчальні плани підготовки політологів: “Політична та соціальна статистика”, “Логіка наукових досліджень” та інші. Для спеціальностей з комп'ютерних наук, фізики та математики викладаються аналогічні дисципліни, але з поглибленим математичним обґрунтуванням методів аналізу.

Цей посібник присвячено найчастіше вживаним методам прикладної статистики. У доборі матеріалу враховано специфіку соціальних наук, яка полягає у великій кількості нечислових показників (порядкових та номінальних). Часто проблемою є також той факт, що емпіричні розподіли не узгоджуються з класичними і параметричні методи не можуть бути застосовані. Тому посібник містить широкий спектр непараметричних методів.

Оскільки видання має навчальне призначення, то велика увага приділена детальному описові алгоритмів розв'язання конкретних задач статистичними методами. Задачі дібрано на основі реальних даних, які виникають у соціальних науках.

Важливим аспектом застосування статистичних методів є їх комп'ютерна реалізація. Сучасна статистична обробка даних практично неможлива без відповідних комп'ютерних програм, таких, наприклад, як пакети STATISTICA, SPSS, SAS, S-Plus. Тому найбільш ефективно навчання студентів має включати поряд з вивченням теоретичних засад і методів також паралельне оволодіння навичками застосування статистичного комп'ютерного забезпечення. Такі підходи фактично відсутні в більшос-

4 ПЕРЕДМОВА

ті вітчизняних посібників, які видані раніше і не відповідають сучасним вимогам підготовки висококваліфікованих спеціалістів. Саме на розв'язання цих проблем і спрямоване пропоноване видання.

У посібнику описано застосування статистичних методів з допомогою пакету STATISTICA. Зауважимо, що цей пакет має стандартний інтерфейс та використовує традиційну термінологію, тому можна легко адаптувати практичну частину і для інших комп'ютерних програм.

Посібник має шістнадцять розділів, кожен з яких відповідає певній темі і містить опис прикладної задачі, статистичних методів її дослідження, необхідні теоретичні обґрунтування, означення, формули, методичні настанови до їх застосовування. Після теоретичного опису кожного методу продемонстровано приклади детального дослідження емпіричних даних за допомогою пакету STATISTICA. Оскільки посібник призначений для студентів нематематичних спеціальностей, то частину матеріалу, яка становить лише теоретичний інтерес, опущено.

Для глибшого оволодіння теоретичним і практичним матеріалом читачам слід звернутися до використаних при підготовці посібника книг, які наведені в списку літератури.

Посібник підготований за часткової підтримки Центрального Європейського Університету (Будапешт). Він розрахований на студентів кваліфікаційного рівня “магістр” із соціальних наук, “спеціаліст” із природничих наук. Може бути також корисним для аспірантів, викладачів та науковців, які використовують статистичні методи аналізу.

Розділи 1–5, 10 та 11 посібника уклали А.Я. Оленко, розділи 6–8 – А.Я. Оленко та М.М. Осипчук разом, розділ 9 – В.Г. Шпортюк, розділи 12 та 13 – М.М. Осипчук. Т.І. Мамчич уклала розділи 14–16.

Автори

Розділ 1

Методи вибірових обстежень

1.1 Планування статистичних обстежень

На етапі планування обстеження, дослідник має проаналізувати цілу низку 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. У який спосіб буде отримано потрібну інформацію (за допомогою наявної статистичної звітності, інтерв'ю, анкетуванням по пошті чи по телефону чи 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 вибірове обстеження?
12. Необхідна кількість елементів у вибірці (обсяг вибірки), що гарантує потрібну точність.
13. Які аналітичні методи будуть застосовані? Вид статистичного аналізу та програмного забезпечення.
14. Методи знаходження вибірових характеристик, процедури статистичного оцінювання та перевірки гіпотез.
15. Оцінка точності результатів.
16. Побудова довірчих інтервалів.
17. Тлумачення результатів.

6 РОЗДІЛ 1

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

1.2 Основні поняття математичної статистики

Потрібно розрізняти такі поняття:

Загальна сукупність. *Загальна сукупність* – набір елементів, властивості та характеристики яких будуть вивчатися.

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

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

Більш формально, у кожному конкретному дослідженні загальна сукупність із кількісного боку зображається деякою випадковою величи-

ною.

Генеральна сукупність – це множина всіх значень, яких може набувати дана випадкова величина.

Так, якщо на фірмі працює 120 осіб, то генеральна сукупність, що описує їхній стаж роботи, складається з 120 даних; генеральна сукупність даних стосовно доходів 200-тисячного міста складається із 200000 даних.

У математичній статистиці розроблені методи аналізу даних, що стосуються як скінченних, так і нескінченних генеральних сукупностей. Характерною особливістю застосування математико-статистичних методів у соціології є аналіз даних, що відповідають скінченним генеральним сукупностям. Проте часто, коли мають справу з дуже великими генеральними сукупностями (населенням усієї країни або великих міст тощо), відповідні генеральні сукупності можна вважати нескінченними і застосовувати до них асимптотичні методи статистики.

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, \dots, x_n) . Оскільки нас цікавлять лише кількісні характеристики елементів, то, формально, вибірка – це набір даних (x_1, \dots, x_n) , отриманих у результаті обстеження частини загальної сукупності.

Приклад. Нехай у групі $N = 10$ студентів. Вивчають їх зріст, при цьому вирішено відібрати для обстеження чотирьох студентів, тобто взяти вибірку обсягом $n = 4$. Це можна зробити $C_{10}^4 = 210$ способами. Нехай відібрано студентів, чий прізвища за списком мають номери 1, 8, 17, 9. Дані про їх зріст (180, 152, 164, 170). Проте, якщо ми візьмемо іншу вибірку з чотирьох студентів (номери 4, 5, 13, 19 за списком), то отримаємо інші вибіркові дані (175, 160, 191, 156). Таким чином, конкретні вибіркові дані різні для кожної з 210 можливих вибірок. Оскільки за-

8 РОЗДІЛ 1

здаlegідь невідомо, які саме елементи будуть відібрані, то невідомо, які саме значення (x_1, \dots, x_n) будуть отримані в результаті обстеження.

Наведені міркування показують, що кожне вибіркоче значення можна тлумачити як випадкову величину, а всі вибіркочі дані – як випадковий вектор.

Конкретні значення, отримані при обстеженні вибірки, – реалізація вибірки.

Часто вибіркочі значення зручно розглядати в порядку зростання

$$x_1^* \leq x_2^* \leq \dots \leq x_n^*,$$

x_1^* – найменше, x_n^* – найбільше з можливих значень.

Означення. *Набір (x_1^*, \dots, x_n^*) називають варіаційним рядом, а випадкові величини x_i^* , $i = 1, \dots, n$ – 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 – обсяг 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.

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

– страти визначають природним чином з аналізу проблеми, на розв'язання якої спрямоване обстеження, і сукупність а 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 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. Потрібно отримати висновки не тільки для всієї сукупності, але й для певних підрозділів.

10 РОЗДІЛ 1

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 вираш у точності при оцінюванні параметрів генеральної сукупності.

Отже, крім загальних статистичних проблем оцінювання основних параметрів генеральної сукупності (частки, середнього, сумарного значення тощо), їх середньоквадратичних похибок, побудови довірчих інтервалів і визначення необхідного обсягу вибірки при стратифікованому відборі виникають і додаткові задачі, пов'язані з оптимальним розбиттям генеральної сукупності на страти і розподілом елементів у ці страти.

1.5 Багатоступеневий гніздовий відбір

При багатоступеневому відборі елементи, що безпосередньо досліджують, вибирають лише на останній стадії, після декількох послідовних випадкових відборів. У такий спосіб 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, \dots . У наведеному прикладі $f_1 = 0.5, f_2 = 0.2, f_3 = 0.3$. Остаточна частка відібраних для обстеження

особистих підсобних господарств $\epsilon f = f_1 f_2 f_3 = 0.5 \cdot 0.2 \cdot 0.3 = 0.03$, тобто будуть обстежені лише 3% всіх особистих господарств.

Гніздовий відбір значно зменшує та спрощує обстеження, проте при цьому виникають деякі методологічні складнощі в аналізі отриманих даних. Основи статистичних методів планування та аналізу даних при гніздовому відборі викладені в монографіях Шварца [21] і Кокрена [15].

Більш складні гібридні плани обстежень наведені Кокреном і Джессеном [11]. У. Кокрен також приділив значну увагу аналізу помилок, що виникають при обстеженнях, а Р. Джессен, крім іншого, наводить детальне порівняння ефективності різних методів і планів обстежень, застосовуючи методи оптимізації, а також звертає увагу на неформальне змістовне тлумачення даних.

1.6 Виконання в пакеті STATISTICA

У пакеті STATISTICA дані, отримані в результаті експерименту, зручно зберігати у вигляді таблиці, стовпці якої звичайно відповідають показникам, які спостерігають (випадковим величинам), а рядки – окремим спостереженням. Така форма дає зручне візуальне зображення даних для попереднього аналізу (описової статистики) і подальшої більш детальної обробки (див. рис. 1.1.)

Заповнюємо таблицю звичайним для Windows-програм чином. Якщо дані знаходяться в іншому форматі в зовнішніх файлах, то можна встановити динамічні зв'язки. Тоді кожна зміна даних у пакеті STATISTICA приводитиме до зміни даних у зовнішньому файлі, і навпаки.

Для роботи з великими масивами даних використовують менеджер мегафайлів – кількість стовпців у таблицях, з якими він оперує, може досягати 32000.

Досить часто для порівняння реальних даних з теоретичними потрібно згенерувати на комп'ютері вибірку з певного теоретичного розподілу. Покажемо на прикладі, як це можна зробити. Принципи роботи з таблицями даних, які буде продемонстровано в прикладі, застосовують і в усіх інших випадках у роботі з реальними даними.

Генерація вибірки

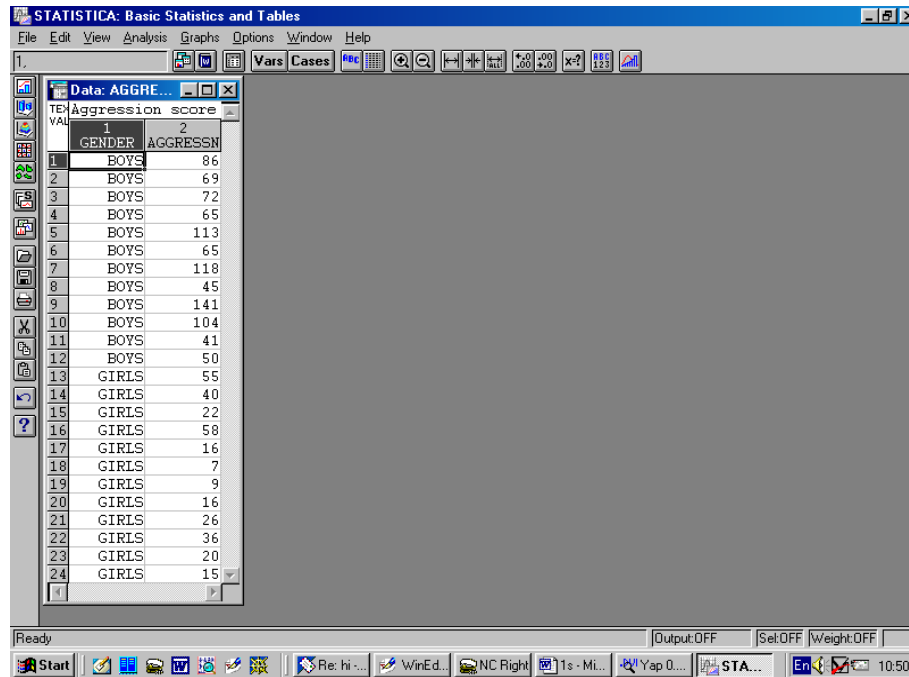
Згенеруємо, наприклад, вибірку обсягом $n = 50$ із показникового розподілу з середнім значенням 5.

Створимо новий файл:

File – New Data,

вкажемо ім'я файла у вікні *File Name : new* (наприклад) – *OK*.

12 РОЗДІЛ 1



The screenshot shows the STATISTICA software window titled "STATISTICA: Basic Statistics and Tables". The main window displays a data table with the following content:

1	2
GENDER	AGGRESSN
1	BOYS 86
2	BOYS 69
3	BOYS 72
4	BOYS 65
5	BOYS 113
6	BOYS 65
7	BOYS 118
8	BOYS 45
9	BOYS 141
10	BOYS 104
11	BOYS 41
12	BOYS 50
13	GIRLS 55
14	GIRLS 40
15	GIRLS 22
16	GIRLS 58
17	GIRLS 16
18	GIRLS 7
19	GIRLS 9
20	GIRLS 16
21	GIRLS 26
22	GIRLS 36
23	GIRLS 20
24	GIRLS 15

Рис. 1.1. Зображення даних у вигляді таблиці

На екрані таблиця, в заголовках якої вказано назви та розміри:
 $10v * 10c$ – (10 змінних-стовпців (variables) по 10 спостережень-рядків (cases)).

Змінимо таблицю до розмірів 1×50 :

кнопка *Vars* (на екрані) – *Delete*;

у вікні *Delete Variables*: вкажемо які змінні-стовпчики вилучити:

From variable: var 2, To variable: var 10 – *OK*.

Кнопка *Cases* – *Add* (додавання) – вікно *Add Cases*: вкажемо, скільки рядків додати і куди:

Number of Cases to Add: 40, Insert after Case: 1 (наприклад) – *OK*.

Згенеруємо вибірку:

виділимо стовпчик – змінну *Var1* (клацанням кнопкою миші по заголовку) – натиснемо праву клавішу миші – в меню, що відкрилося, виберемо *Variable specs* (специфікації змінної) – у вікні *Variable 1* введемо:

Name: x (наприклад),

у нижньому полі *Long name* вводять вираз, який визначає вид розподілу генеральної сукупності (вираз має починатися знаком “дорівнює”).

Це можна зробити:

або безпосередньо набором на клавіатурі,

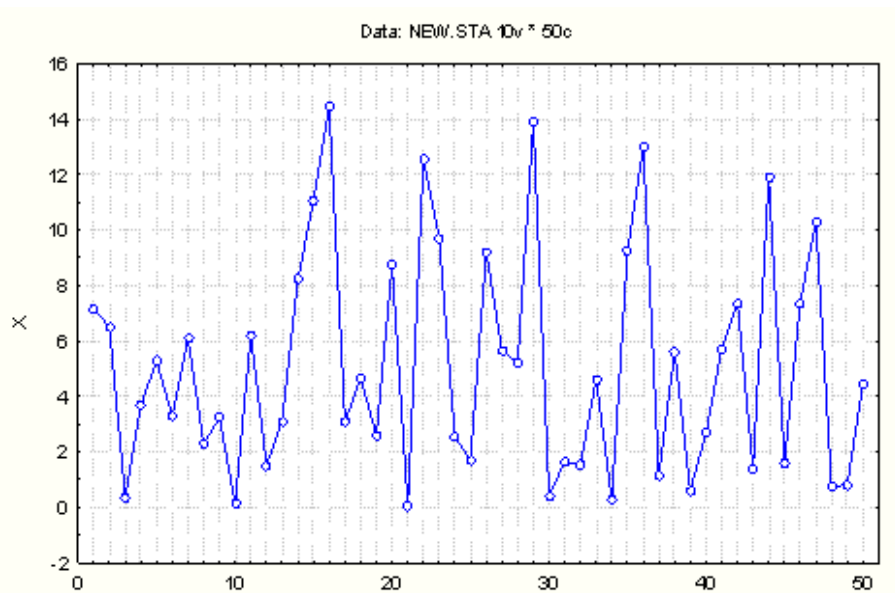


Рис. 1.2. Спостереження, розподілені за показниковим законом з середнім 5 ($n = 50$)

або за допомогою клавіші *Functions*, обираючи в меню *Category* та *Name* потрібну функцію та вставляючи клавішею *Insert*.

Щоб задати закон розподілу, потрібно ввести, наприклад,
 $=Rnd(2)$ для рівномірного розподілу на відріжку $[0, 2]$ ($R[0, 2]$),
 $=Vnormal(Rnd(1); 2; 0.5)$ для $N(2, \sigma^2 = 0.5^2)$,
 $=VExpon(Rnd(1); 0.2)$ для $E(5)$ з середнім $1/0.2 = 5$; (для прикладу ми вибрали значення параметра $\lambda = 0.2$).

Така форма запису визначається способом генерації: за допомогою функції, оберненої (буква *V*) до функції розподілу та генератора випадкових чисел $R[0, 1]$ ($Rnd(1)$).

Роздрукуємо вибірку командою *Print* меню *File*.

Зобразимо вибірку графічно:

Graphs – Custom Graphs (настроюємо графіки) – *2D graphs* – у вікні, що відкривалося, все можна лишити за замовчуванням – *OK*. Отримуємо графік (рис. 1.2).

Побудова варіаційного ряду

Перший спосіб:

виділимо потрібну змінну (стовпчик) – натискаємо праву клавішу миші

– *Graphs*;

– виберемо *Quiq Stats Graphs* (швидкі статистики та графіки);

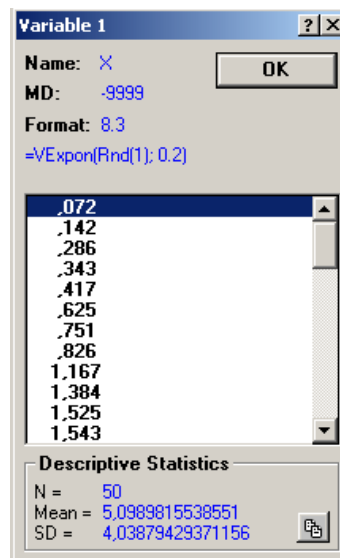


Рис. 1.3. Частина варіаційного ряду

– *Values / Stats of Vars* (значення та статистики);
спостерігаємо варіаційний ряд та вибіркове середнє (*mean*) та стандартне відхилення (*SD*), див. рис. 1.3.

Другий спосіб:

увійдемо до модуля *Data Menagement* (подвійне клацання лівою клавішею миші на чистому полі та вибір модуля у вікні *Module Switcher*; якщо модуль уже завантажений, то *Alt+Tab* до появи модуля):

Analysis;

Sort;

вставляємо ім'я змінної та напрямок сортування:

Ascen (за зростанням) або *Desc* (за спаданням) – *OK*.

Спостерігаємо варіаційний ряд.

Емпірична функція розподілу

Перший спосіб побудови:

Graphs – Stats 2D Graphs – Histogram – у вікні, що з'явилося, встановимо:

Graph Type : *Regular, Cumulative Counts* (накопичені частоти),

Fit Type: *Exponential* (для нашого прикладу) або *off* (без підбору),

Variables: *x*.

Спостерігаємо графік функції емпіричного розподілу (рис. 1.4). Графік можна відредагувати: змінити колір лінії, точки, фон, шкали, написи. Для цього потрібно навести стрілку на відповідний елемент і двічі

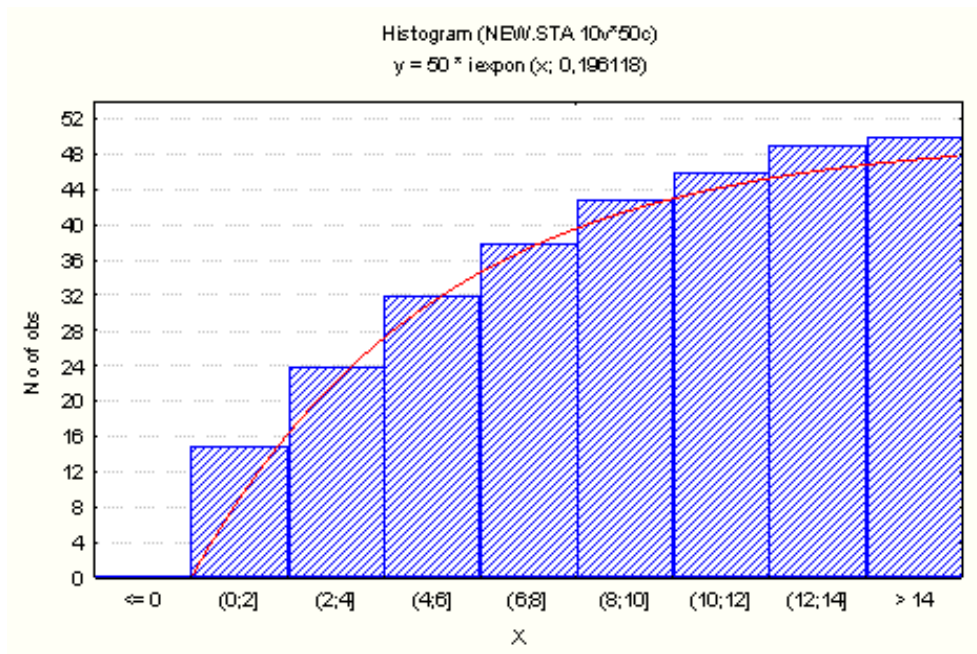


Рис. 1.4. Емпірична функція розподілу

клацнути лівою клавішею миші.

Графік можна роздрукувати або зберегти у файлі.

Другий спосіб побудови:

– впорядкуємо за зростанням нашу вибірку (див. **Побудова варіаційного ряду**);

– утворимо нову змінну F для значень функції:

клавіша Var – Add (див. **Генерація вибірки**);

виділимо нову змінну $NEWVAR$ – права клавіша миші

– $Variable Specs$ – $Name: F$ – $Long name: = V0/50$

(оператор $V0$ створює масив цілих чисел);

і побудуємо графік:

$Graphs$ – $Custom Graphs$ – $2D Graph$

у новому вікні встановимо: в полі $X : x$, в полі $Y : F$,

$Step Plot$ (східці, але не $Line Plot$ – лінії) – OK .

Спостерігаємо функцію емпіричного розподілу.

Побудова гістограми частот

$Graphs$ – $Stats 2D Graphs$ – $Histograms$;

у вікно, що з'явилося, вставимо: ім'я змінної,

$Graph Type: Regular$,

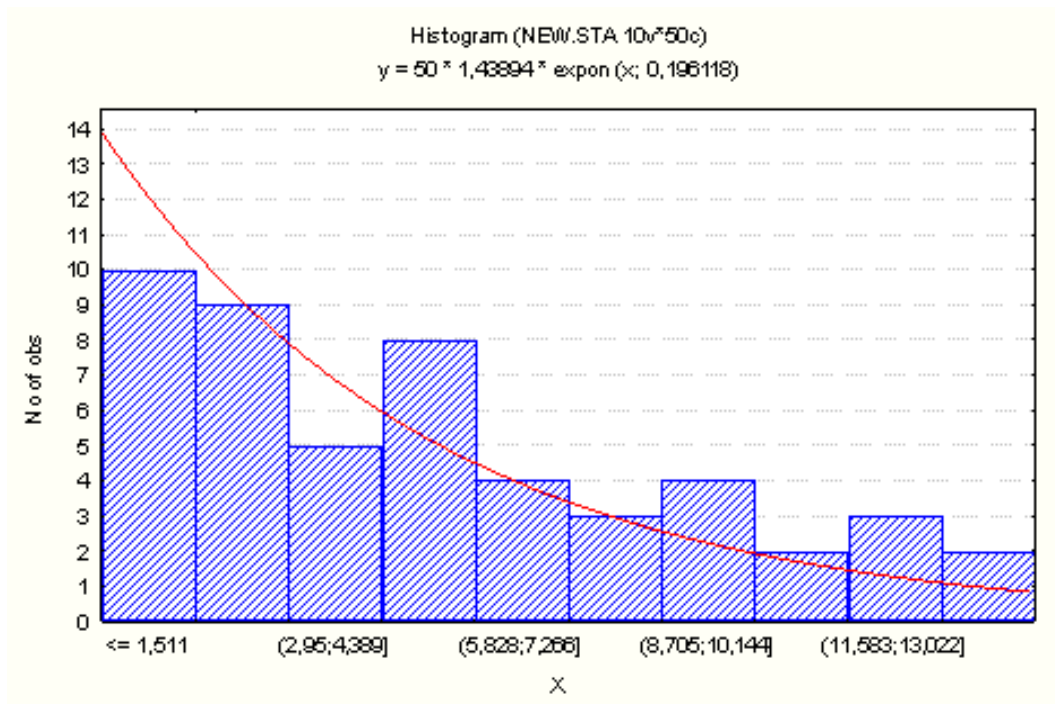


Рис. 1.5. Гістограма

Fit Type – off (без підбору), або той теоретичний тип розподілу, з яким ми хочемо порівняти емпіричний розподіл, кількість інтервалів групування:

Integer Mode – Auto (автоматичний вибір кількості інтервалів) – ОК. Отримуємо гістограму (рис. 1.5). Відредагуємо графік, якщо потрібно.

Проведемо порівняльний аналіз:

На графіку натиснемо кнопку *Next*.

У вікні, що з'явилося, виберемо кількість інтервалів групування *Categories*: змінимо автоматично вибрану кількість інтервалів, наприклад, 10 – ОК.

Порівнюємо отриману гістограму з попередньою. Для цього ввійдемо до *Windows – Tile vertically*. Спостерігаємо обидві гістограми.

Розділ 2

Оцінювання параметрів

З кількісного боку генеральна сукупність характеризується цілою низкою 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, здебільшого немає. Вибіркове обстеження якраз і спрямоване на отримання висновків щодо параметрів генеральної сукупності на основі вибірки. Це робиться шляхом обчислень за певними формулами типу $\hat{a} = f(x_1, \dots, x_n)$, 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)*

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

яку обчислюють за результатами вибірових обстежень.

У такий спосіб розглядають вибіркове середнє, вибіркову дисперсію, вибіркові коефіцієнти варіації та кореляції тощо. Усі ці вибіркові характеристики слугують оцінками відповідних параметрів генеральної сукупності. Конкретні процедури їх обчислення розглянемо пізніше.

2.1 Властивості оцінок

Оскільки оцінка \hat{a} є функцією від випадкових величин, то для широкого класу функцій $f(\cdot)$, $\hat{a} = f(x_1, \dots, x_n)$ буде знову випадковою величиною,

яка характеризується своїм розподілом, математичним сподіванням, дисперсією. Саме в цих термінах формують вимоги до точності оцінок.

2.1.1 Незміщеність

Означення. Оцінку \hat{a} параметра a називають незміщеною, якщо її математичне сподівання $\mathbf{E}\hat{a}$ дорівнює справжньому значенню параметра, який оцінюють, тобто

$$\mathbf{E}\hat{a} = a.$$

Іншими словами, $\mathbf{E}(\hat{a} - a) = 0$. Наочно, це означає, що при багаторазовому використанні оцінки \hat{a} для апроксимації a , середнє значення похибки $\hat{a} - a$ дорівнює нулеві. Незміщеність вказує на відсутність систематичної похибки.

2.1.2 Конзистентність

Означення. Оцінку $\hat{a} = \hat{a}_n = f(x_1, \dots, x_n)$ називають конзистентною (від англ. *consistent*), якщо для кожного $\varepsilon > 0$

$$P\{|\hat{a}_n - a| > \varepsilon\} \rightarrow 0 \text{ при } n \rightarrow N,$$

тобто при зростанні обсягу вибірки оцінка \hat{a}_n збігається до a за ймовірністю.

Оцінку \hat{a}_n називають *строго конзистентною*, якщо має місце збіжність з ймовірністю 1.

Змістовно конзистентність означає, що з ростом обсягу вибірки n точність оцінки зростає.

2.1.3 Ефективність

У математичній статистиці кількісно міру похибки при заміні a на незміщену оцінку \hat{a} характеризують за допомогою дисперсії оцінки

$$\sigma_{\hat{a}}^2 = \mathbf{D}\hat{a} = \mathbf{E}|\hat{a} - a|^2.$$

Зрозуміло, що варто користуватися тими оцінками, які мають меншу дисперсію. Часто для оцінювання того ж самого параметра генеральної сукупності можна запропонувати декілька оцінок. *Ефективною* називають таку оцінку, яка має мінімально можливу дисперсію.

Ефективність є дуже бажаною властивістю оцінки, проте не завжди вдається її отримати.

2.2 Методи одержання оцінок

2.2.1 Емпіричні оцінки

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

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

де Φ – 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 $\mathbf{F}(x, \theta)$, то за оцінку θ 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 θ .

2.2.2 Метод моментів

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

Наприклад, для абсолютно-неперервних розподілів

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

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

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

20 РОЗДІЛ 2

Метод моментів полягає у тому, що деяка кількість емпіричних моментів \hat{m}_i прирівнюють до відповідних моментів $m_i(\theta_1, \theta_2, \dots, \theta_s)$ і з цієї системи рівнянь знаходять оцінки $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s$.

Приклад. За вибіркою $\zeta = (x_1, \dots, x_n)$ із генеральної сукупності з розподілом $N(a, \sigma^2)$ методом моментів отримати оцінки для середнього a та дисперсії σ^2 .

Оскільки

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

і відповідні емпіричні моменти

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

то отримуємо систему рівнянь

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

Отже, шукані оцінки такі:

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

2.2.3 Метод максимальної вірогідності

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

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

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

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

Означення. Логарифмічною функцією вірогідності називають функцію

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

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

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

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

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

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

2.2.4 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

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

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

Нехай в точках x_1, \dots, x_n зроблені спостереження змінної $y = y(x)$, результати яких відповідно y_1, \dots, 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:

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

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

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

була мінімальною.

2.3 Точкові та інтервальні оцінки

Припустимо, що дослідник отримав конкретну реалізацію вибірки (x_1, \dots, x_n) і обчислив значення \hat{a} за формулою $\hat{a} = f(x_1, \dots, x_n)$, 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 α 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*

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

Співвідношення (2.2) слід читати “ймовірність того, що справжнє значення a лежить в інтервалі від a_H до a_B , дорівнює \mathcal{P} ”. Межі інтервалу $a_H = a_H(x_1, \dots, x_n)$, $a_B = a_B(x_1, \dots, x_n)$, які знаходять за допомогою вибіркової дані, 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.

2.4 Квантилі

Нехай $\mathcal{P} \in (0, 1)$, $\mathbf{F}(x)$ – деяка функція розподілу.

Означення. *Квантиль рівня \mathcal{P} – розв’язок $d_{\mathcal{P}}$ рівняння*

$$\mathbf{F}(x) = \mathcal{P}.$$

Іншими словами, для випадкової величини ξ з розподілом $\mathbf{F}(x)$,

$$\mathbf{P}(\xi < d_{\mathcal{P}}) = \mathcal{P}.$$

Якщо $F(x)$ строго монотонна і неперервна, то d_p визначено однозначно.

2.5 Виконання в пакеті STATISTICA

Статистичне порівняння оцінок

Не завжди вдається аналітично обчислити дисперсію оцінки. Як експериментально визначити, якою з оцінок користатися? За однією вибіркою не можна судити про розкид значень оцінки, оскільки значення лише одне; необхідно мати кілька вибірок, наприклад, $k = 20$, (чи хоча б 5 – 10), оцінити розкид значень для кожної оцінки і віддати перевагу тій оцінці (тому способу оцінювання), для якої розкид найменший. Якщо ж вибірка лише одна, то можна (якщо n досить велике) розбити її випадковим чином на кілька вибірок, і за ними порівнювати якість оцінок.

Для прикладу, сформуємо $k = 20$ вибірок обсягу $n = 10$ і визначимо значення оцінок a_1, a_2, a_3 на кожній вибірці.

Запустимо пакет *STATISTICA for Windows*, вибравши в меню *Basic Statistic/Tables and Banners*.

Створення таблиці потрібних розмірів

З пункту меню *File* виберемо команду *New Data*; вкажемо ім'я файлу для збереження майбутньої інформації, наприклад *ESTIM – ОК*.

Бачимо на екрані таблицю 10×10 , де кожен стовпець відведено під змінну (назву якої винесено в заголовок стовпця).

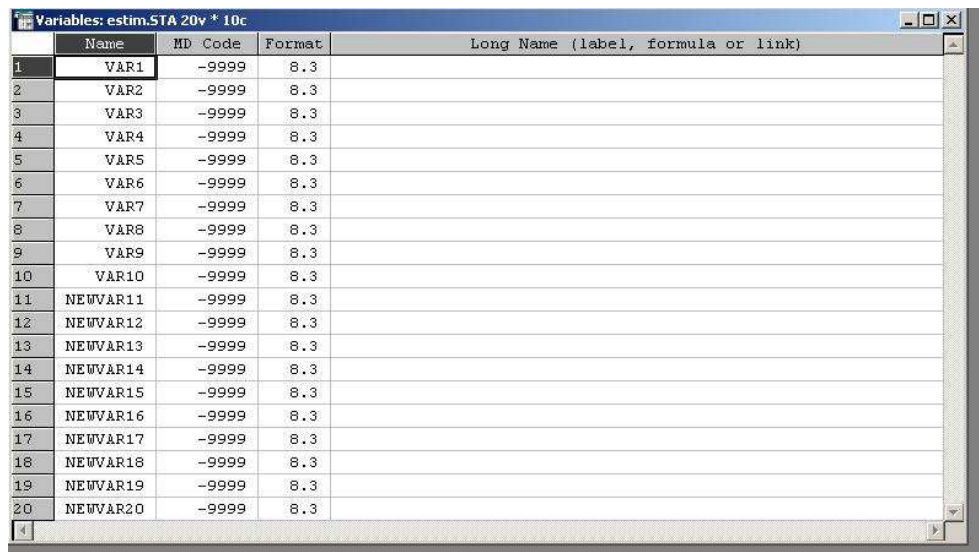
Перетворимо цю таблицю до розмірів 20×10 (20 вибірок по 10 спостережень): кнопка *Vars* (змінні), чи через меню *Edite-Variables*, і у новому меню виберемо команду *Add* (додати).

На екрані з'явиться запит про кількість додаткових змінних (стовпців) і про те, куди їх помістити.

Додамо 10 змінних, проставивши 10 у поле *Number... to add* (набором на клавіатурі чи кнопками праворуч від поля; у поле *Insert after* вкажемо ім'я змінної *Var10*, після якої будуть вставлені в матрицю нові стовпці; потім *ОК*).

Тепер можна переконатися (переглядом таблиці), що в ній 20 стовпців; крім того, розміри таблиці (у даному випадку, $20v * 10c$) завжди зазначено у її заголовку. Кількість рядків не змінюємо: 10. Зауважимо, що коли кількість рядків (*case*) чи стовпців (*variable*) у таблиці перевищує необхідну, можна таблицю не зменшувати.

24 РОЗДІЛ 2



	Name	MD Code	Format	Long Name (label, formula or link)
1	VAR1	-9999	8.3	
2	VAR2	-9999	8.3	
3	VAR3	-9999	8.3	
4	VAR4	-9999	8.3	
5	VAR5	-9999	8.3	
6	VAR6	-9999	8.3	
7	VAR7	-9999	8.3	
8	VAR8	-9999	8.3	
9	VAR9	-9999	8.3	
10	VAR10	-9999	8.3	
11	NEWVAR11	-9999	8.3	
12	NEWVAR12	-9999	8.3	
13	NEWVAR13	-9999	8.3	
14	NEWVAR14	-9999	8.3	
15	NEWVAR15	-9999	8.3	
16	NEWVAR16	-9999	8.3	
17	NEWVAR17	-9999	8.3	
18	NEWVAR18	-9999	8.3	
19	NEWVAR19	-9999	8.3	
20	NEWVAR20	-9999	8.3	

Рис. 2.1. Специфікація змінних

Генерація вибірок

Послідовність дій:

клавіша *Vars – All specs* (специфікація усіх) – з’являється вікно-таблиця, у першому стовпці якої є назви змінних (*var1, var2, ..., var20*), а в четвертому (*Long Name*) – функція розрахунку, див. рис. 2.1;

виділимо першу клітинку цього стовпця і введемо

$=\text{Rnd}(10)$ – генерація випадкових чисел, рівномірно розподілених на відрізьку $[0;10]$.

Скопіюємо цей запис у буфер обміну: *Edit – Copy* (чи кнопкою *Copy*), а потім перенесемо її в інші клітинки (з 2 по 20). Закриємо вікно.

Виконаємо зроблені призначення:

кнопка $x = ?$ – *All variables – OK*.

Збережемо 2 – 3 перші вибірки-стовпця для того, щоб надалі роздрукувати:

виділяємо 2 – 3 стовпці – *File – Export Data* – формат *Text – File name: Samples* (наприклад).

Роздрукувати їх можна і відразу:

File – Print – Variables (вказати, які саме).

Визначення значень оцінок a_1, a_2 і a_3 за 20 вибірками

Визначимо статистики, за якими обчислюватимемо оцінки: виділимо всю матрицю, потім проведемо три такі операції (див. рис. 2.2):

Edit – Block Stats/Columns – Sums (Max's, Medians)

Можемо це зробити інакше: права клавіша миші – *Block Stats/Columns* (блок статистик по стовпчиках) – *Sums, (Max's, Medians)*.

У нижній частині таблиці з'являються три рядки з потрібними статистиками. Для нових рядків уведемо зручніші позначення:

кнопка *Cases – Names*.

Транспонуємо нашу матрицю, що тепер має розмір $20v \times 13c$, у матрицю $13v \times 20c$ (щоб робити дії зі стовпцями):

Edit – Transpose – Data File.

Додамо в матрицю 3 стовпці (з 14 по 16 для значень оцінок)

Vars – Add – Number of vars: 3 – after: 13

і введемо формулу для обчислення значення оцінки a_1 : виділимо 1-й новий стовпець *Newvar1*, натиснемо

Vars – Current Specs (специфікація)

*–Name: A1 – Long name: = 2/10 * V11.*

Аналогічно введемо формули для обчислення значень оцінок a_2 і a_3 :

для a_2 : $= 11/10 * V12$,

для a_3 : $= 2 * V13$.

Отримані результати (стовпці $a1, a2, a3$) значень трьох оцінок на 20 вибірках збережемо, щоб надалі роздрукувати:

виділимо $a1, a2, a3$ – *File – Export Data* – формат *Text* – вкажемо, куди зберегти і з яким ім'ям.

Характеристики розсіювання оцінок

Виділимо стовпці $a1, a2, a3$ – *Edit – Block Stats/Column – SD's* (стандартне відхилення), потім аналогічно: *Min's, Max's*.

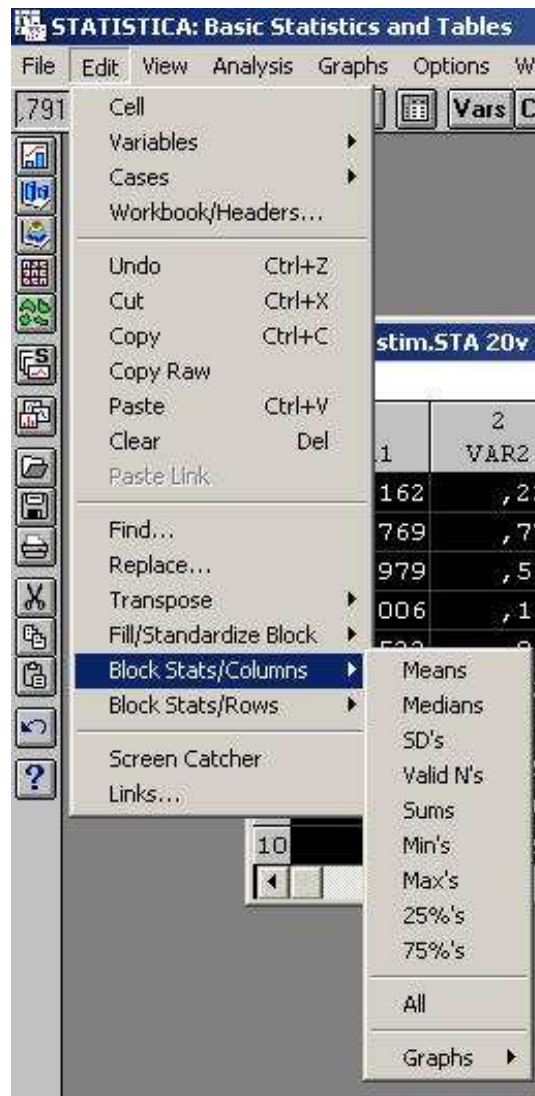


Рис. 2.2. Знаходження оцінок основних параметрів

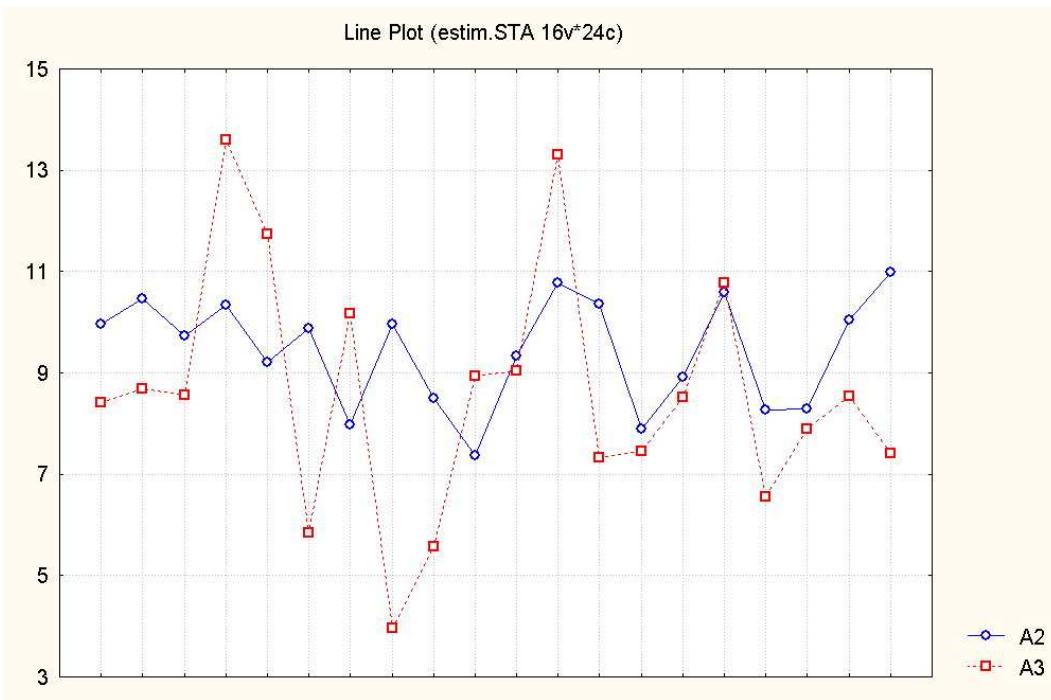


Рис. 2.3. Порівняння оцінок

Порівняння розмахів w і стандартних відхилень S_a для трьох оцінок показує, що оцінка $\hat{a}_2(X_1, \dots, X_n)$ найточніша, а оцінка $\hat{a}_3(X_1, \dots, X_n)$ найменш точна.

Порівняння оцінок a_2 і a_3 графічно

Graphs – Stats 2D Graphs – Line Plots (Variables)

у вікні *2D Line Plots: Variables: A2 – A3*,

Graphs Type : Multiple, Fit (підбір розподілу) : off (виключити),

Cases: 1–20; – ОК.

З графіка (рис. 2.3) видно, що значення оцінок розташовані в околі 10 і що оцінка a_2 має розсіювання менше, ніж a_3 . Роздрукуємо цей графік: *File – Print Graphs*.

Оцінювання за вибірками обсягу $n=40$ і $n=160$

Повторимо попередні процедури для $n = 40$ і $n = 160$.

Порівняння $S_a(n)$ трьох оцінок графічно для значень $n = 10, 40, 160$:

– утворимо 4 нові змінні довжини 3:

n : зі значеннями 10, 40, 160; $Sa1, Sa2, Sa3$: зі значеннями стандартного відхилення для трьох оцінок.

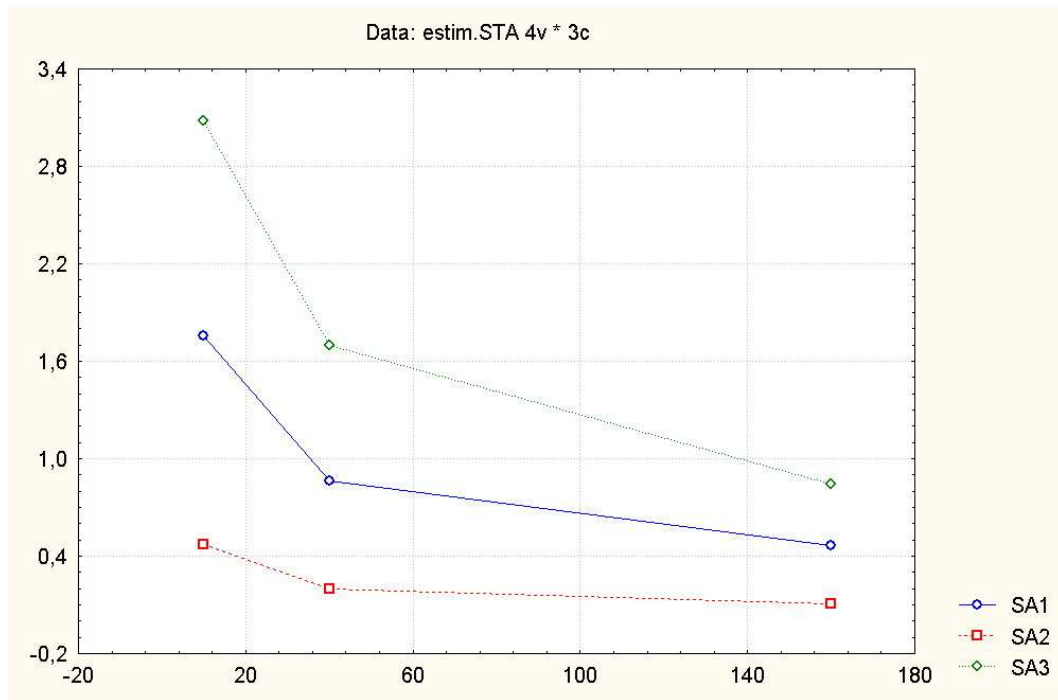


Рис. 2.4. Порівняння трьох оцінок за стандартним відхиленням

Побудуємо графіки $S_a(n)$:
 виділимо змінні $Sa1, Sa2, Sa3$ – *Graphs – Custom Graphs – 2D Graphs*
 введемо для *Plot1* $X : N, Y : Sa1$,
 для *Plot2* $X : N, Y : Sa2$,
 для *Plot3* $X : N, Y : Sa3$ – *OK*.

Спостерігаємо три криві $S_a(n)$ як функції n (див. рис. 2.4).

Очевидно, що оцінка $\hat{a}_2(X_1, \dots, X_n)$ найточніша, а найменш точна – оцінка $\hat{a}_3(X_1, \dots, X_n)$.

Графік виводимо на друк: *File – Print Graphs*.

Розділ 3

Оцінювання частки та кількості елементів із певною ознакою

3.1 Точкові оцінки для P і N_1

Нехай є генеральна сукупність обсягом N одиниць, серед яких N_1 одиниць мають певну ознаку C . Наприклад, N – кількість усіх магазинів у місті, N_1 – кількість магазинів самообслуговування (супермаркетів).

Частку елементів із певною ознакою визначають як

$$P = N_1/N, \quad 0 \leq P \leq 1.$$

У деяких випадках частку визначають також у відсотках, тоді вона дорівнює $P^* = P \cdot 100\%$.

Приклад. Серед $N = 400$ вступників на екзамені $N_1 = 60$ отримали відмінну оцінку. Тоді $P = 60/400 = 0.15$ або $P^* = P \cdot 100\% = 15\%$.

Обидві характеристики N_1 та P можуть бути визначені при повному обстеженні всіх одиниць сукупності. Якщо ж висновки роблять за вибіркою обсягу n , в якій виявилось n_1 одиниць з певною ознакою, то оцінки параметрів N_1 та P визначають за формулами

$$\hat{P} = n_1/n,$$

$$\hat{N}_1 = \hat{P}N = n_1N/n.$$

Ці оцінки є незміщеними та конзистентними оцінками відповідних параметрів. Нагадаємо, що вони є реалізаціями випадкових величин. Для вибору без повернення випадкова величина n_1 має гіпергеометричний розподіл, тобто

$$\mathbf{P}\{n_1 = k\} = \frac{C_k^{N_1} C_{n-k}^{N-N_1}}{C_n^N}, \quad k = 0, 1, \dots, n,$$

з математичним сподіванням $\mathbf{E}(n_1) = nP$ та дисперсією

$$\mathbf{D}(n_1) = nP(1 - P)(N - n)/(N - 1).$$

Для кількісної характеристики точності оцінювання потрібно знати дисперсії та стандартні похибки оцінок \hat{P} та \hat{N}_1 . Враховуючи властивості гіпергеометричного розподілу, для дисперсій і середніх квадратичних похибок оцінок отримаємо такі вирази:

$$\sigma_{\hat{N}_1}^2 = \frac{N^2 P(1 - P) N - n}{n(N - 1)}, \quad \sigma_{\hat{N}_1} = N \sqrt{\frac{P(1 - P)}{n}} \sqrt{\frac{N - n}{N - 1}}, \quad (3.1)$$

$$\sigma_{\hat{P}}^2 = \frac{1}{N^2} \sigma_{\hat{N}_1}^2 = \frac{P(1 - P) N - n}{n(N - 1)}, \quad \sigma_{\hat{P}} = \sqrt{\frac{P(1 - P)}{n}} \sqrt{\frac{N - n}{N - 1}}, \quad (3.2)$$

Коли обсяг вибірки досить великий, то можна вважати, що $N - 1 \approx N$. Тоді формули (3.1), (3.2) трансформуються як:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1 - P)}{n}} \sqrt{1 - \frac{n}{N}} = \sqrt{\frac{P(1 - P)}{n}} \sqrt{1 - f}, \quad (3.3)$$

$$\sigma_{\hat{N}_1} = N \sqrt{\frac{P(1 - P)}{n}} \sqrt{1 - f}. \quad (3.4)$$

Проаналізуємо ці вирази. Бачимо, що стандартні похибки залежать від:

- 1) частини відбору $f = n/N$;
- 2) обсягу вибірки n ;
- 3) частки P у генеральній сукупності.

Множник $k = \sqrt{\frac{N - n}{N - 1}} \approx \sqrt{1 - f}$ називають *коректуючим множником*, або *поправкою на скінченність генеральної сукупності* (ПСС).

Зрозуміло, що $0 \leq k \leq 1$. У таблиці наведено значення множника k залежно від f :

f	0.001	0.01	0.05	0.10	0.20
$k = \sqrt{1 - f}$	1	0.995	0.975	0.949	0.894
f	0.40	0.60	0.80	0.90	0.95
$k = \sqrt{1 - f}$	0.775	0.632	0.447	0.316	0.224

Очевидно, що при $f \leq 0.05$, тобто коли вибірка не перевищує 5% генеральної сукупності, коректуючий множник k практично не відрізняється від одиниці.

Зауважимо, що при виборі з поверненням, n_1 розподілено за біноміальним законом, тобто

$$\mathbf{P}\{n_1 = k\} = C_n^k P^k (1 - P)^{n-k}, \quad k = 0, 1, \dots, n$$

з математичним сподіванням $\mathbf{E}n_1 = nP$ та дисперсією $\mathbf{D}(n_1) = nP(1-P)$.

Крім того, відомо, що коли обсяг генеральної сукупності необмежено зростає ($N \rightarrow \infty$) тоді, коли n і N_1/N залишаються фіксованими, то гіпергеометричний розподіл збігається до біноміального.

Вважають, що апроксимацію біноміальним розподілом варто вживати, якщо $f < 0.05$. Деякі автори рекомендують застосовувати її вже при $f < 0.1$.

Отже, коли вибірка мала (N -велике) або ж при виборі з поверненням, користуємося таким виразом для середньоквадратичної похибки

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}, \quad \sigma_{\hat{N}_1} = N \sqrt{\frac{P(1-P)}{n}}. \quad (3.5)$$

Таким чином, вибір без повернення більш точний, ніж із поверненням, оскільки формули (3.5) дають завжди більші значення, ніж формули (3.3), (3.4).

Обсяг вибірки n – єдина величина, що впливає на похибку оцінювання, яку може варіювати статистик. З ростом обсягу вибірки похибка оцінювання зменшується пропорційно $1/\sqrt{n}$.

Нарешті, стандартна похибка оцінки залежить від величини P , точніше від $P(1-P)$. Наведемо значення $P(1-P)$ і $\sqrt{P(1-P)}$ для різних P :

P	0.01	0.05	0.10	0.20	0.30
$\sqrt{P(1-P)}$	0.99	0.218	0.300	0.400	0.458
P	0.50	0.70	0.80	0.90	0.95
$\sqrt{P(1-P)}$	0.500	0.468	0.400	0.300	0.218

Зауважимо, що при значеннях P , не дуже близьких до 0 і 1 ($0.20 \leq P \leq 0.80$), значення $\sqrt{P(1-P)}$ мало змінюється залежно від P . Цей факт має велике практичне значення. У формулах (3.1) – (3.4), (3.5) для дисперсій і стандартних похибок фігурує значення параметра P , проте це значення, здебільшого, невідоме, а вибіркоче обстеження проводиться саме з метою його оцінки. Отже, (3.1) – (3.4), (3.5) мають лише теоретичний інтерес. Проте для $0.2 \leq P \leq 0.8$ значення P можна замінити

на його оцінку \hat{P} (близьку до справжнього значення P), тобто замінити $\sqrt{P(1-P)}$ на $\sqrt{\hat{P}(1-\hat{P})}$ і при цьому не зважати на приблизний характер такої заміни.

Тому необхідні для розрахунків формули мають вигляд:

параметр – P , оцінка – $\hat{P} = n_1/n$, оцінка стандартної похибки –

$$S_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{1 - \frac{n}{N}};$$

параметр – N_1 , оцінка – $\hat{N}_1 = \hat{P}N = n_1N/n$, оцінка стандартної похибки

$$S_{\hat{N}_1} = N \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{1 - \frac{n}{N}}.$$

Ці формули містять уже тільки відомі значення: обсяг генеральної сукупності N , обсяг вибірки n та результат вибіркового обстеження n_1 – кількість елементів із фіксованою ознакою. Для того, щоб відрізнити значення дисперсій і стандартних похибок, які розраховують за вибіркою, від теоретичних, вживаємо для них позначення S^2 і S на відміну від σ^2 і σ .

3.2 Довірчі інтервали для P і N_1

3.2.1 Точні методи

Нехай n_1 елементів вибірки обсягу n належать до класу C . Треба зробити висновки стосовно кількості N_1 одиниць генеральної сукупності, які належать до C .

Точні методи побудови *довірчих інтервалів (надійних проміжків)* для N_1 ґрунтуються на властивостях гіпергеометричного розподілу.

Означення. Верхньою довірчою межею для N_1 вважаємо таке значення N_1^B , для якого ймовірність отримати у вибірці не більше n_1 елементів із класу C дорівнювала заданій імовірності α_B . Формально N_1^B знаходять з рівняння

$$\sum_{j=0}^{n_1} P(j, n-j; N_1^B, N - N_1^B) = \alpha_B, \quad (3.6)$$

де $P(\cdot)$ – значення ймовірності для гіпергеометричного розподілу.

Зауважимо, що розв'язок (3.6) не завжди ціле число, проте за змістом шукане N_1^B має бути цілим. Тому N_1^B знаходять як найменше ціле, для якого справедлива нерівність

$$\sum_{j=0}^{n_1} P(j, n-j; N_1^B, N - N_1^B) \leq \alpha_B. \quad (3.7)$$

Аналогічно нижню довірчу межу для N_1 знаходять як найбільше ціле, для якого справедлива нерівність

$$\sum_{j=n_1}^n P(j, n-j; N_1^H, N - N_1^H) \leq \alpha_H. \quad (3.8)$$

Гіпергеометричний розподіл табульовано. Довірчі межі для P легко знайти із співвідношень

$$P_H = N_1^H / N, \quad P_B = N_1^B / N,$$

де N_1^B , N_1^H знаходять за (3.7), (3.8).

3.2.2 Повторна вибірка. Біноміальний розподіл

У разі повторної вибірки для знаходження довірчого інтервалу для P використовуємо процедуру побудови довірчого інтервалу для параметра біноміального розподілу. Це означає, що верхню P_B і нижню P_H довірчі межі для P знаходимо, як розв'язки нерівностей

$$\sum_{j=0}^{n_1} P_{P,N}(j) \leq \alpha, \quad \alpha < 1/2,$$

$$\sum_{j=n_1}^n P_{P,N}(j) \leq 1 - \alpha, \quad \alpha > 1/2,$$

де $P_{P,N}(j) = C_N^j P^j (1-P)^{N-j}$ – біноміальні ймовірності.

Якщо $n_1 = 0$, то вважають, що $P_N = 0$, якщо ж $n_1 = n$, то $P_b = 1$.

Довірчі межі для параметра біноміального розподілу наведені в спеціальних статистичних таблицях.

3.2.3 Нормальна апроксимація. Симетричні інтервали

Вважають, що для розподілу оцінок \hat{P} і $\hat{N}_1 = \hat{P}N$ при $nP(1-P) > 1$ має місце задовільна, а при $nP(1-P) > 9$ – гарна апроксимація нормальним

розподілом з математичними сподіваннями P і PN та стандартними похибками σ_P і σ_{N_1} , відповідно, які визначають за (3.3), (3.4). Таким чином, за цих умов довірчі інтервали для P і N_1 з надійністю $\mathcal{P} = 1 - \alpha$ визначають за формулами

$$\hat{p} \pm u_\alpha \sigma_P, \quad \hat{N}_1 \pm u_\alpha \sigma_{N_1},$$

де u_α – квантиль рівня $1 - \alpha/2$ для стандартного нормального розподілу.

Оскільки теоретичні дисперсії σ_P і σ_{N_1} здебільшого невідомі, а відомі лише їх оцінки S_P та S_{N_1} , то безпосередні розрахунки проводять за формулами

$$\begin{aligned} \hat{P} \pm u_\alpha \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{1-f}, \\ \hat{N}_1 \pm u_\alpha N \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{1-f}. \end{aligned} \quad (3.9)$$

У.Кокрен (1976) замість (3.9) рекомендує застосовувати дещо ширший інтервал

$$\hat{P} \pm u_\alpha \left[\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{1-f} + \frac{1}{2n} \right].$$

3.2.4 Несиметричні інтервали

Наведений вище метод розрахунку симетричних довірчих інтервалів дає помилкові результати, коли частка P помітно відрізняється від 0,5. Існує інший метод побудови довірчих інтервалів, який теж використовує квантиль нормального розподілу, проте дає задовільні результати і при P близьких до 0 або 1. При цьому вимагають, щоб $f = n/N < 0,05$.

Довірчий інтервал (P_H, P_B) визначають співвідношеннями

$$P_H, P_B = \frac{1}{n + u_\alpha^2} \left\{ n_1 + \frac{u_\alpha^2}{2} \pm u_\alpha \sqrt{\frac{n_1(n - n_1)}{n} + \frac{u_\alpha^2}{n}} \right\}. \quad (3.10)$$

Якщо у вираз (3.10) підставити $n_1 = 0$, то навіть при нульовому результаті отримаємо верхню межу можливих значень частини P

$$P_B \leq u_\alpha^2 / (n + u_\alpha^2). \quad (3.11)$$

Зауважимо, що формула (3.11) дає дещо завищену оцінку, ніж при використанні точного розподілу, проте вона простіша для розрахунків.

Приклад. При перевірці в установі серед $n = 500$ випадково відібраних довідок не було виявлено жодної помилкової, хоч можливість появи помилок слід брати до уваги. Треба вказати інтервал, в якому з ймовірністю 0,955 знаходиться кількість помилкових довідок. За формулою (3.11) ($u_\alpha = 2$) ця кількість лежить у інтервалі $(0; 4/504)$, тобто з ймовірністю 0,955 кількість помилкових довідок не перевищує 0,79% усіх довідок.

3.3 Виконання в пакеті STATISTICA

Оцінювання частки та кількості елементів із певною ознакою та числових характеристик точності їх оцінювання проводиться за відповідними точними або наближеними теоретичними формулами. Для того, щоб їх застосовувати, нам потрібно знати лише значення функції розподілу для біноміальних та гіпергеометричних випадкових величин.

Покажемо, як знаходити в пакеті STATISTICA ці показники.

Значення функції розподілу для біноміальних випадкових величин

Скористаємося модулем *Nonparametrics & Distributions*.

Створюємо нову таблицю з кількістю випадків не меншою ніж параметр N (загальна кількість випробувань) у біноміальному розподілі, що нас цікавить. Зручно взяти їх цілими числами від 0 до N .

Виберемо на панелі кнопку *Distribution Fitting* і у віконечку тип дискретного розподілу (*Distributions*) вибираємо *Binomial* (біноміальний) (див. рис. 3.1)

Заповнюємо вікна для кількості категорій, нижньої та верхньої меж категорій (згідно з тим біноміальним розподілом, що нас цікавить) – див., наприклад, рис. 3.2.

Натискаємо кнопку ОК.

Отримуємо таблицю частот – див. рис. 3.3.

Значення з стовпчика "*cumul. % expected*" і дають, при переході від відсотків до часток (просто ділимо на 100), потрібні ймовірності для біноміального розподілу.

Підставивши ці числа в формули, за допомогою калькулятора легко знаходити відповідні характеристики. Якщо ці показники потрібно рахувати часто, то процедуру легко запрограмувати за допомогою мови *Statistica Basic*.

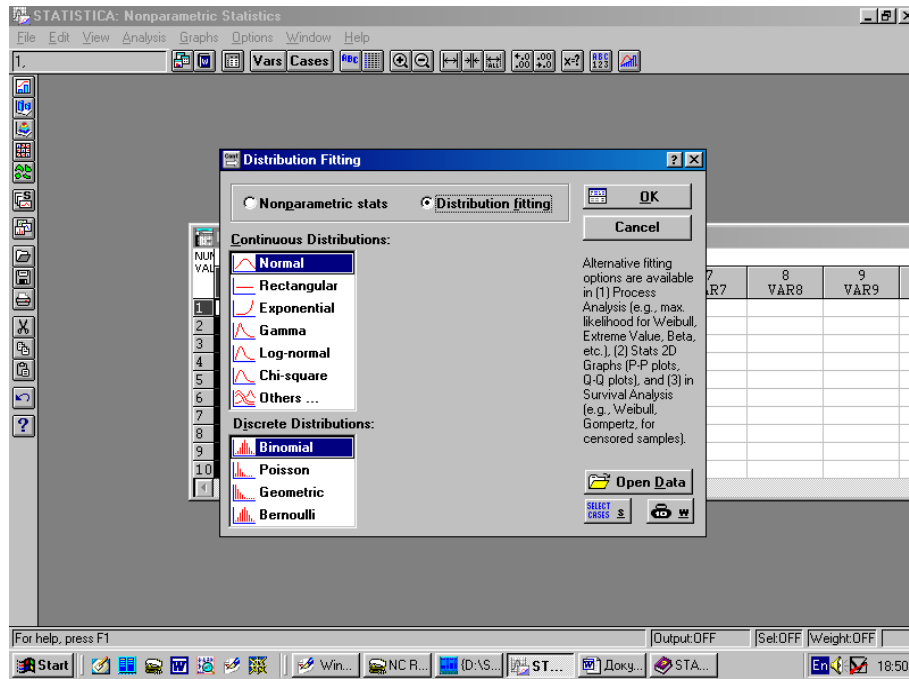


Рис. 3.1. Вибір типу розподілу

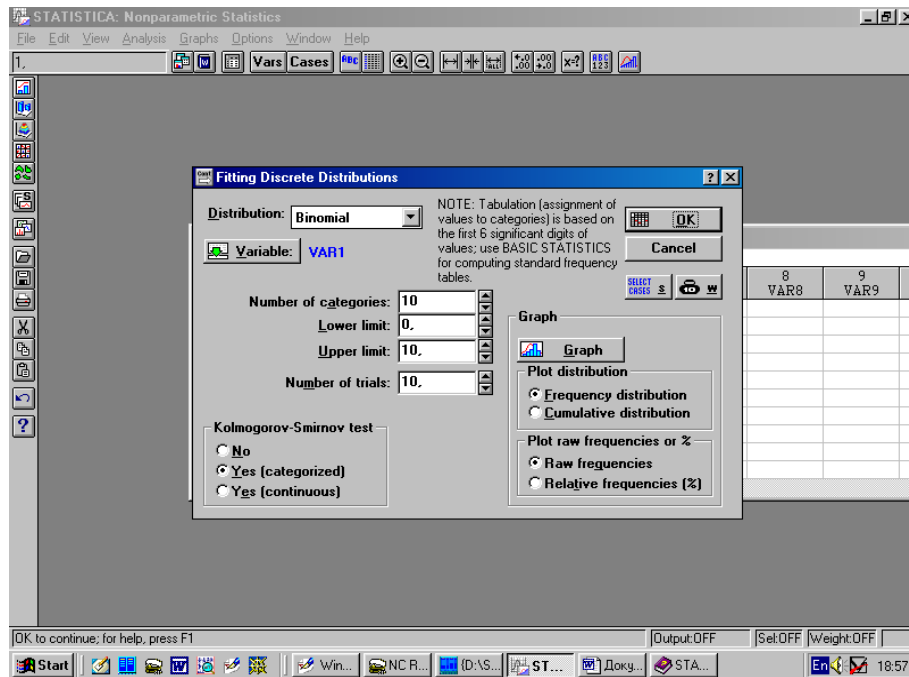


Рис. 3.2. Вибір параметрів розподілу

STATISTICA: Nonparametric Statistics

File Edit View Analysis Graphs Options Window Help

0. Columns Rows

Variable VAR1 ; distribution: Binomial p = .55000

Continue... Kolmogorov-Smirnov d = .2004403, p = n.s.
Chi-Square: -----, df = 0, p = ---

Upper Boundary	observed freq-cy	cumulativ observed	percent observed	cumul. % observed	expected freq-cy	cumulativ expected	percent expected	cumul. % expected
0.	0	0	0.00000	0.0000	.003405	.00341	.03405	.0341
1.	1	1	10.00000	10.0000	.041617	.04502	.41617	.4502
2.	1	2	10.00000	20.0000	.228896	.27392	2.28896	2.7392
3.	1	3	10.00000	30.0000	.746031	1.01995	7.46031	10.1995
4.	1	4	10.00000	40.0000	1.595677	2.61563	15.95677	26.1563
5.	1	5	10.00000	50.0000	2.340327	4.95595	23.40327	49.5595
6.	1	6	10.00000	60.0000	2.383667	7.33962	23.83667	73.3962
7.	1	7	10.00000	70.0000	1.664783	9.00440	16.64783	90.0440
8.	1	8	10.00000	80.0000	.763026	9.76743	7.63026	97.6743
9.	1	9	10.00000	90.0000	.207241	9.97467	2.07241	99.7467
Infinity	1	10	10.00000	100.0000	.025330	10.00000	.25330	100.0000

10 10,000

Fitting Disc...

Ready Output:OFF Sel:OFF Weight:OFF

Start Win... NCR... ID:\S... ST... Доку... STA... 19:01

Рис. 3.3. Таблиця частот

Розділ 4

Оцінювання середніх і сумарних значень

4.1 Оцінювання середнього та мінливості

Нехай кожному елементові генеральної сукупності відповідає певне кількісне значення ознаки x_i , $i = 1, 2, \dots, N$.

Отже, генеральна сукупність може бути описана такими статистичними характеристиками (параметрами):

- сумарним значенням ознаки для генеральної сукупності

$$X' = \sum_{i=1}^N x_i,$$

- середнім значенням ознаки (математичним сподіванням)

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} X'.$$

Мінливість ознаки характеризують дисперсією σ^2 та середньою квадратичною похибкою σ :

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2, \quad \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Усі ці характеристики можуть бути обчислені лише за суцільного обстеження. Якщо ж дослідник має лише n вибірових даних x_1, \dots, x_n , то він може використати *вибіркове сумарне значення*

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

та вибіркоче середнє

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

як оцінки для відповідних параметрів генеральної сукупності.

Приклад. Середнє значення витрат на харчування, отримане при опитуванні 120 випадково відібраних студентів гуртожитку, можна трактувати як оцінку середніх витрат на харчування всіх студентів гуртожитку.

Вибіркові оцінки \bar{x} та x' є конзистентними оцінками відповідних параметрів генеральної сукупності; їх теоретичні дисперсії та середні квадратичні похибки мають вигляд:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \quad (4.1)$$

$$\sigma_{x'}^2 = \frac{N^2 \sigma^2}{n} \frac{N-n}{N-1}, \quad \sigma_{x'} = \frac{N\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \quad (4.2)$$

а коефіцієнти варіації вибіркового середнього $V_{\bar{x}}$ та сумарного значення $V_{x'}$ дорівнюють один одному:

$$V_{\bar{x}} = \frac{\sigma_{\bar{x}}}{\bar{x}} = \frac{N\sigma_{\bar{x}}}{N\bar{x}} = V_{x'}, \quad (4.3)$$

Деякі спрощення можна досягти за умов:

$$\begin{aligned} 1) \quad N-1 \approx N, \quad \text{тоді} \quad \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \sqrt{1-f}, \\ 2) \quad N \rightarrow \infty, \quad \text{тоді} \quad \sigma_{\bar{x}} &\rightarrow \frac{\sigma}{\sqrt{n}}. \end{aligned} \quad (4.4)$$

Зауважимо, що формули (4.1) – (4.4) мають лише теоретичне значення, оскільки в них входить σ , яке можна знайти, обстеживши всю генеральну сукупність. Проте аналіз цих формул дає змогу зробити низку корисних висновків стосовно величин, що впливають на точність оцінки (середньоквадратичну похибку). Так середня квадратична похибка середнього $\sigma_{\bar{x}}$, за всіх інших рівних умов, зменшується зі збільшенням обсягу вибірки n і зростає при зростанні обсягу генеральної сукупності N та прямує до сталої величини σ/\sqrt{n} при $N \rightarrow \infty$.

Для практичних обчислень теоретичну дисперсію σ^2 або теоретичну стандартну похибку σ у (4.1), (4.2), (4.4) замінюють на їх вибіркові аналоги:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

і остаточно отримують такі формули:

$$S_{\bar{x}}^2 = \frac{S^2}{n} \frac{N-n}{N-1}, \quad S_{\bar{x}} = \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

або

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \sqrt{1-f} \text{ при } N-1 \approx N,$$

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \text{ при } N \rightarrow \infty;$$

та

$$S_{x'}^2 = N^2 S_{\bar{x}}^2 = \frac{N^2 S^2}{n} \frac{N-n}{N-1}, \quad S_{x'} = \frac{NS}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

або

$$S_{x'} = \frac{NS}{\sqrt{n}} \sqrt{1-f} \text{ при } N-1 \approx N.$$

4.2 Надійні проміжки для параметрів нормальних випадкових величин

1. Надійний проміжок для математичного сподівання θ з рівнем надійності $1 - \alpha$, при відомому стандартному відхиленні σ :

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

де d_α – квантиль нормального $N(0, 1)$ розподілу рівня α .

2. Надійний проміжок для математичного сподівання θ з рівнем надійності $1 - \alpha$, якщо стандартне відхилення невідоме:

$$\left(\bar{x} - \frac{t_{1-\frac{\alpha}{2}, n-1}}{\sqrt{n}} \hat{S}, \bar{x} + \frac{d_{1-\frac{\alpha}{2}}}{\sqrt{n}} \hat{S} \right),$$

де $t_{1-\alpha, n-1}$ – квантиль рівня $1 - \alpha$ розподілу Стьюдента з $(n - 1)$ ступенем вільності, а $\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

3. Надійний проміжок для дисперсії σ^2 з рівнем надійності $1 - \alpha$, при відомому математичному сподіванні a :

$$\left(\frac{n\hat{\sigma}^2}{\chi_{1-\frac{\alpha}{2}, n}^2}, \frac{n\hat{\sigma}^2}{\chi_{\frac{\alpha}{2}, n}^2} \right),$$

де $\chi_{\alpha,n}^2$ – квантиль χ^2 розподілу з n ступенями вільності рівня α , а $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$.

4. Надійний проміжок для дисперсії σ^2 з рівнем надійності $1 - \alpha$, якщо математичне сподівання невідоме:

$$\left(\frac{(n-1)\hat{S}^2}{\chi_{1-\frac{\alpha}{2},n-1}^2}, \frac{(n-1)\hat{S}^2}{\chi_{\frac{\alpha}{2},n-1}^2} \right).$$

5. Надійний проміжок з рівнем надійності $1 - \alpha$ для різниці математичних сподівань $a_1 - a_2$ двох випадкових величин ξ та η , при відомих стандартних відхиленнях σ_1 та σ_2 :

$$\left(\bar{\xi} - \bar{\eta} - d_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}; \bar{\xi} - \bar{\eta} + d_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right).$$

6. Надійний проміжок з рівнем надійності $1 - \alpha$ для різниці математичних сподівань $a_1 - a_2$ двох випадкових величин ξ (обсяг вибірки n) та η (обсяг вибірки m), якщо стандартні відхилення однакові, але невідомі:

$$\left(\bar{\xi} - \bar{\eta} - t_{1-\frac{\alpha}{2},n+m-2} s \sqrt{\frac{n+m}{nm}}; \bar{\xi} - \bar{\eta} + t_{1-\frac{\alpha}{2},n+m-2} s \sqrt{\frac{n+m}{nm}} \right),$$

де

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

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

4.3 Розподіл вибіркового середнього

Лише в окремих випадках, знаючи розподіл генеральної сукупності, можна точно знайти розподіл вибіркового середнього \bar{x} . У загальному випадку допоможе центральна гранична теорема, згідно з якою за досить великого обсягу вибірки n розподіл випадкової величини \bar{x} є приблизно нормальним із математичним сподіванням m та дисперсією σ^2 .

Цей факт широко використовують для побудови надійних проміжків. За досить великих обсягів вибірки для знаходження надійних проміжків використовують наведені вище формули.

4.4 Виконання в пакеті STATISTICA

Згенеруємо випадкову величину з певним розподілом і з відомими параметрами.

Знаходження вибірових характеристик

Перший спосіб:

клацнемо правою клавішею миші на назві стовпчика з вибіркою;

- *Quick Basic Stats*
- *Descriptives of var*
- ОК.

отримаємо таблицю з характеристиками:

mean (середнє),

Confid 95% (межі довірчого інтервалу з рівнем довіри 0.95: нижня та верхня),

Sum (сума),

Minimum, Maximum,

Range (розмах),

Variance (дисперсія),

Std. Dev. (стандартне відхилення).

Порівняємо вибірове середнє, медіану та стандартне відхилення з відповідними теоретичними значеннями.

Другий спосіб: на назві стовпчика з вибіркою клацнемо правою клавішею миші – *Block Stats / Columns* (блок статистик за колонками) – виділимо необхідне або *All*.

Знайти оцінки середнього, дисперсії, проміжки надійності для дисперсії можна також в модулі *Basic Statistics – Descriptive Statistics*.

Виберемо спочатку змінні з таблиці для аналізу – див. рис. 4.1. Наприклад – *All*. Натиснемо на кнопку *More statistics* та виберемо *All* – див. рис. 4.2. Натиснемо ОК і отримаємо оцінки різноманітних параметрів розподілів.

Власні формули для обчислення різноманітних оцінок параметрів за вибірками (стовпцями таблиці) можна написати за допомогою мови *Statistica Basic*, яка дає змогу запрограмувати різноманітні статистичні процедури.

Процедури для оцінювання інших параметрів розподілів (наприклад, моди) можна знайти в модулі *Nonparametrics – Descriptive Statistics* – розділ *Ordinal Descriptive Statistics*.

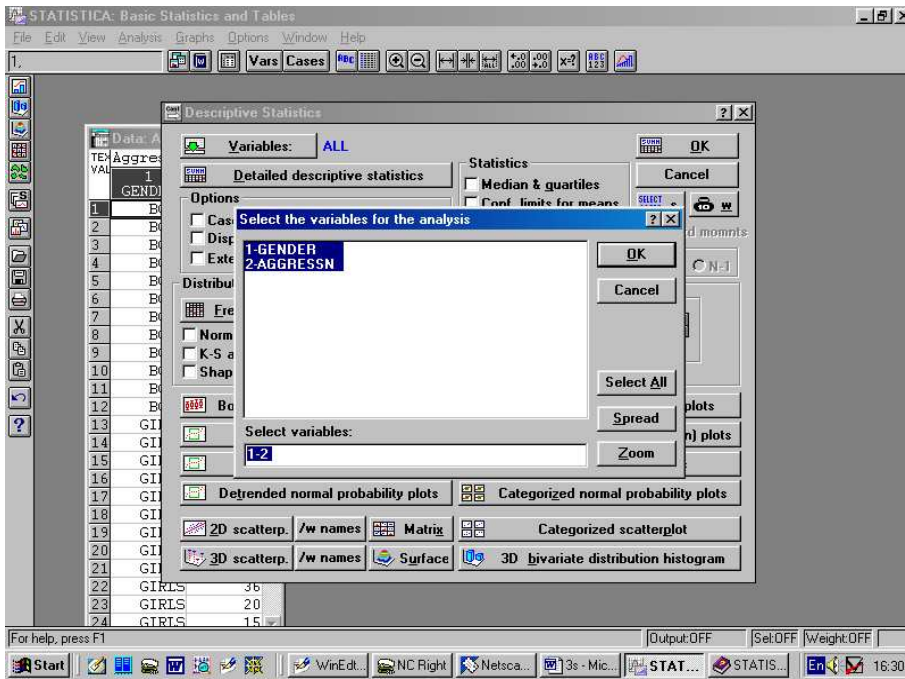


Рис. 4.1. Вибір змінних

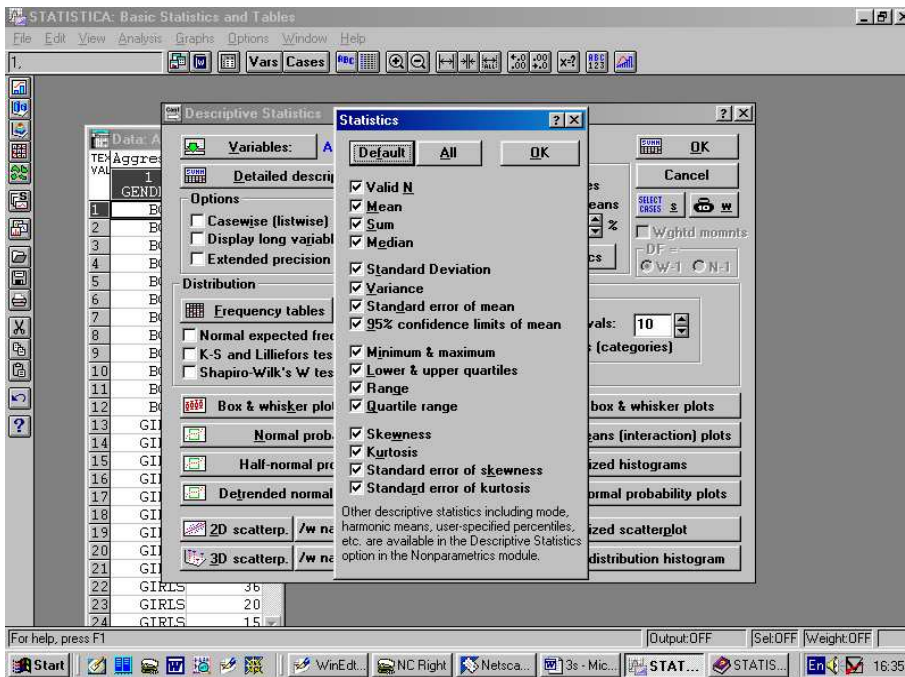


Рис. 4.2. Вибір оцінок різноманітних параметрів

Розділ 5

Визначення обсягу вибірки

У попередніх розділах нашу увагу було зосереджено на оцінюванні основних характеристик генеральної сукупності (частки, кількості елементів із певною ознакою) за вибірковими даними, а також на визначенні кількісної міри точності цих оцінок. При цьому обсяг вибірки вважався фіксованим. Проте вже аналіз середньоквадратичної похибки оцінок наочно показує, наскільки істотно обсяг вибірки впливає на точність оцінок.

При плануванні вибіркового обстеження постає питання, як забезпечити необхідну точність результатів і при цьому уникнути зайвих витрат.

У цьому розділі зосередимося на проблемі визначення кількості вибірових даних, необхідних для забезпечення заданої точності при оцінюванні частки та кількості елементів з фіксованою ознакою.

5.1 Показники точності оцінювання

Наведемо найбільш розповсюджені кількісні ймовірісно-статистичні міри точності оцінок, в основі яких лежить дисперсія або середньоквадратична похибка. Зауважимо, що остаточний вибір показника точності як основи для планування вибіркового обстеження належить не статистикам, а споживачам результатів обстеження. Суттєвим моментом повинна бути можливість змістовного тлумачення результатів.

Наведемо основні показники точності оцінок:

1. *Стандартна (середньоквадратична) похибка σ* : показує порядок величини можливого відхилення оцінки від справжнього значення параметра. За умови нормального розподілу оцінок у 2/3 випадків дійсне відхилення оцінки менше за стандартну похибку.

2. *Гранична похибка вибірки* e_α – практично максимально допустиме відхилення оцінки від параметра. Рідко, з імовірністю α , можливе відхилення більше за e_α . Ймовірність α вибирається дослідником малою.
3. *Величина довірчого інтервалу* $l = 2e_\alpha$ – діапазон, у якому знаходиться значення оцінюваного параметра.
4. *Показники відносної точності* – коефіцієнт варіації або показники, перелічені в пунктах 2 та 3, поділені на математичне сподівання оцінки.

Інколи буває потрібно задавати точність результатів обстежень у відносних показниках. Зрозуміло, що треба по-різному ставитись до середньої квадратичної похибки $\sigma_p = 0.04$, коли значення параметра $P = 1$ і коли $P = 0.1$. У першому випадку середня квадратична похибка дорівнює лише 4% від значення параметра P , а в другому – уже 40%.

Означення. Коефіцієнт варіації оцінки \hat{a} з математичним сподіванням $E\hat{a} = a$ і дисперсією $D\hat{a} = \sigma^2$ дорівнює $V = \sigma/a$.

Часто коефіцієнт варіації подають у відсотках.

5.2 Визначення обсягу вибірки n при оцінюванні часток P

Точність визначається середніми квадратичними похибками σ_p .

Якщо вважати величини σ_p , P та N відомими, то отримаємо формулу для n :

$$n = \frac{P(1 - P)N}{\sigma_p^2(N - 1) + P(1 - P)}, \quad (5.1)$$

коли вважати N досить великим, то $N \approx N - 1$ і формула (5.1) спрощується:

$$n = \frac{P(1 - P)N}{\sigma_p^2(N - 1) + P(1 - P)} = \frac{N}{1 + N\sigma_p^2/P(1 - P)} = \frac{n_0}{1 + n_0/N}, \quad (5.2)$$

де $n_0 = P(1 - P)/\sigma_p^2$.

Ця формула буде основою для подальших висновків. Зауважимо, що при $N \rightarrow \infty$ граничний обсяг вибірки дорівнює

$$n_0 = \lim_{N \rightarrow \infty} n = \frac{P(1 - P)}{\sigma_p^2}. \quad (5.3)$$

Формулу (5.3) варто застосовувати, коли немає жодної інформації стосовно N , або коли N дуже велике, а $f < 0,05$. Коли ж N не дуже велике, то формула (5.3) дає завищені результати щодо необхідної кількості спостережень n . Значення n_0 також використовують як перше наближення, що потребує подальшого уточнення.

При застосуванні формул (5.1) – (5.3) виникає ще кілька проблем.

У ці формули, крім N , входить величина P , яку якраз і треба оцінити за вибіркою. Інколи буває відома апіорна інформація відносно приблизного значення P , виходячи з теоретичних міркувань або з аналогічних досліджень. Тоді цю попередню оцінку можна використати в (5.1) – (5.3).

Якщо ж стосовно P нічого не відомо, то варто у формулу (5.1) підставити значення $P = 0,5$, яке дає максимальне значення добутку $P(1 - P) = 0,25$. При цьому дослідник отримує дещо завищене значення для n , яке гарантує необхідну точність.

У.Кокрен (1976), розвиваючи ідеї Кокса, рекомендує також проводити обстеження у два етапи. На першому етапі беруть просту випадкову вибірку обсягом m_1 , за допомогою якої знаходять оцінку частки P_1 . Цю оцінку використовують для знаходження потрібного обсягу n більшої вибірки за формулою

$$n = \frac{P_1(1 - P_1)}{\sigma_p^2} + \frac{3 - 8P_1(1 - P_1)}{P_1(1 - P_1)} + \frac{1 - 3P_1(1 - P_1)}{\sigma_p^2 m_1}.$$

Отже, потрібно додатково обстежити $n - m_1$ елементів, потім за всіма n даними знайти нову оцінку \hat{p}_1 , тобто $\hat{p}_1 = n_1/n$ і внести поправку на зсув. Остаточна оцінка для P знаходиться за формулою

$$\hat{P} = \hat{p}_1 + \frac{\sigma_p^2(1 - 2\hat{p}_1)}{\hat{p}_1(1 - \hat{p}_1)}.$$

Така процедура дає надійніші оцінки для P , проте подовжує терміни проведення обстежень.

5.2.1 Визначення n при заданій граничній похибці e_p

Вважатимемо, що всі оцінки мають нормальний (асимптотично нормальний) розподіл.

Необхідний обсяг вибірки знаходять за формулою

$$n = \frac{u_\alpha^2 P(1-P)N}{e_p^2(N-1) + u_\alpha^2 P(1-P)},$$

де u_α – квантиль рівня $1 - \alpha/2$ для стандартного нормального розподілу.

При $N \approx N - 1$ маємо

$$n = \frac{u_\alpha^2 P(1-P)}{e_p^2 + u_\alpha^2 P(1-P)/N} = \frac{N}{1 + Ne_p^2/u_\alpha^2 P(1-P)} = \frac{n_0}{1 + n_0/N}$$

та

$$n_0 = \lim_{N \rightarrow \infty} n = \frac{u_\alpha^2 P(1-P)}{e_p^2}$$

5.2.2 Визначення обсягу вибірки при заданій відносній точності

При заданому коефіцієнті варіації V необхідний обсяг вибірки дорівнює

$$n = \frac{(1-P)}{V^2 P + (1-P)/N} = \frac{N}{1 + NV^2 P/(1-P)} \quad (5.4)$$

і при $N \rightarrow \infty$ прямує до граничного значення

$$n_0 = \frac{(1-P)}{V^2 P} \quad (5.5)$$

Якщо немає апріорної інформації стосовно величини P , але з певних міркувань можна вказати нижню межу для P , то значення P_{\min} слід підставити в (5.4), (5.5) і отримане n гарантує необхідну точність. Дозволяють також застосування двохетапної схеми. Після першого етапу, на якому за вибіркою обсягом m_1 знайшли оцінку P_1 , треба додатково обстежити $n - m_1$ елементів, де

$$n = \frac{(1-P_1)}{V^2 p_1} + \frac{3}{P_1(1-P_1)} + \frac{1}{V^2 P_1 m_1},$$

а потім знайти

$$\hat{p} = \hat{p}_1 - V^2 \hat{p}_1(1 - \hat{p}_1).$$

5.3 Обсяг вибірки при дослідженні декількох ознак

У більшості обстежень дані збирають стосовно не однієї, а декількох ознак. Одним із методів визначення обсягу вибірки в цій ситуації є така послідовність дій: спочатку вибирають граничні значення похибок для кожної з цих ознак, визначають найважливіші для цього обстеження ознаки, знаходять необхідні значення для обсягу вибірки n_i окремо за кожною ознакою, а потім приймають компромісне рішення з урахуванням знайдених n_i для окремих ознак, вартості й термінів обстеження.

5.4 Визначення обсягу вибірки при оцінюванні середніх і сумарних значень

При прямому оцінюванні середніх і сумарних значень обсяг вибірки n – єдина величина, змінюючи яку дослідник може впливати на точність оцінки. Як і при оцінюванні частин, вимоги до точності можуть формулюватися в термінах дисперсії або середньої квадратичної похибки оцінок, граничної та відносної похибки. Розглянемо окремо ці випадки.

5.4.1 Визначення обсягу вибірки n при оцінюванні середнього при заданій середній квадратичній похибці

Нехай а рiогi задано бажане значення $s_{\bar{x}}$. Тоді при $N - 1 \approx N$ маємо формули для визначення n :

$$n = \frac{\sigma^2 N}{N\sigma_{\bar{x}}^2 + \sigma^2} = \frac{N}{1 + N(\sigma_{\bar{x}}/\sigma)^2} = \frac{\sigma^2}{\sigma_{\bar{x}}^2 + \sigma^2/N} \quad (5.6)$$

звідки

$$n_0 = \lim_{N \rightarrow \infty} n = \frac{\sigma^2}{\sigma_{\bar{x}}^2}, \quad (5.7)$$

де N – обсяг генеральної сукупності, σ^2 – дисперсія генеральної сукупності. Отже, коли нема точних даних про N , то у формулу (5.6) можна підставити число більше за N (таку верхню межу часто можна вказати) та отримати n , яке гарантує необхідну точність. Якщо ж про N нема ніякої інформації, слід використати формулу (5.7), яка дає дещо завищені значення для n , проте гарантує необхідну точність.

Певні складнощі на стадії планування вибірки можуть бути пов'язані і з відсутністю точних даних щодо σ^2 . У цьому разі можна використати:

1. Дані щодо дисперсії та середньої квадратичної похибки з аналогічних обстежень, які проводилися раніше;
2. Спеціально провести попереднє пробне обстеження з малим обсягом m_0 , наприклад, з $m_0 = 30$, за результатами якого оцінити вибіркочну дисперсію

$$S_0^2 = (1/(m_0 - 1)) \sum_{i=1}^{m_0} (x_i - \bar{x}_0),$$

і скористатися формулою (Кокрен (1976))

$$n \approx \frac{S_0^2}{\sigma_{\bar{x}}^2} (1 + 2/m_0);$$

3. Оцінити σ^2 на основі припущень щодо розподілу генеральної сукупності та співвідношень між середньоквадратичною похибкою та максимальним інтервалом вар'ювання ознаки $R = x_{\min} - x_{\max}$ для основних типів розподілів (Шварц (1978)).

5.4.2 Визначення обсягу вибірки при оцінюванні середнього при заданій граничній похибці $e_{\bar{x}}$

Нехай задане бажане значення граничної похибки $e_{\bar{x}} = u_{\alpha} \sigma_{\bar{x}}$ (розподіл \bar{x} вважаємо приблизно нормальним).

У цьому разі обсяг вибірки n визначають за формулами

$$n = \frac{N u_{\alpha}^2 \sigma^2}{N e_{\bar{x}}^2 + u_{\alpha}^2 \sigma^2} = \frac{N}{1 + N(e_{\bar{x}}/u_{\alpha} \sigma)^2} = \frac{u_{\alpha}^2 \sigma^2}{e_{\bar{x}}^2 + u_{\alpha}^2 \sigma^2 / N}$$

та

$$n \approx (u_{\alpha} \sigma / e_{\bar{x}})^2 \quad \text{при } N \rightarrow \infty,$$

де u_{α} – квантиль рівня $1 - \alpha/2$ для стандартного нормального розподілу.

5.4.3 Визначення обсягу вибірки при оцінюванні сумарного значення при заданій середньоквадратичній або граничній похибці

Довірчий інтервал для сумарного значення ознаки X' можна знайти лише для скінченної генеральної сукупності з відомим обсягом N . Отже,

необхідний обсяг вибірки, що гарантує задану середньоквадратичну похибку $\sigma_{X'} = N S_{\bar{x}}$ визначають за формулами

$$n = \frac{N\sigma^2}{N\sigma^2 + \sigma_{x'}^2} = \frac{N}{1 + N(\sigma_{x'}/N\sigma)^2},$$

а при заданій граничній похибці $e_{x'} = u_{\alpha}\sigma_{X'}$ – за формулами

$$n = \frac{N^2 u_{\alpha}^2 \sigma^2}{e_{x'}^2 + N^2 u_{\alpha}^2 \sigma^2} = \frac{N}{1 + N(e_{x'}/Nu_{\alpha}\sigma)^2}.$$

5.4.4 Визначення обсягу вибірки при заданій відносній точності

Перевага відносних оцінок у тому, що вони є величинами, які не мають розмірності, і за їх допомогою можна порівнювати точність оцінок різних ознак (і в різних обстеженнях). Припустимо, що $N \approx N - 1$. Тоді при заданому коефіцієнті варіації V оцінки середнього для необхідного обсягу вибірки n маємо співвідношення

$$n = \frac{V^2}{V_{\bar{x}}^2 + V^2/N} = \frac{N}{1 + N(V_{\bar{x}}/V)^2}, \quad (5.8)$$

і

$$n \approx (V/V_{\bar{x}})^2 \quad \text{при } N \rightarrow \infty \quad (f < 0.05). \quad (5.9)$$

Оскільки коефіцієнти варіації оцінок середнього та сумарного значення дорівнюють один одному, то формули (5.8), (5.9) можна використовувати і тоді, коли метою досліджень є сумарні значення ознак.

При практичному використанні формул (5.8), (5.9) виникають проблеми визначення $V_{\bar{x}}$, які ми вже обговорювали. У цьому разі слід скористатися результатами аналогічних попередніх обстежень або іншою апріорною інформацією, що дає верхню оцінку для $V_{\bar{x}}$.

5.5 Обсяг вибірки за необхідності отримати оцінки для підрозділів сукупності

Досить часто потрібно отримати оцінки не тільки для сукупності в цілому, але й для підрозділів. Якщо ці підрозділи можна виділити заздалегідь, як, наприклад, географічні (або адміністративні) райони, то n_i знаходиться окремо по кожному підрозділу. Припустимо, що середне

треба оцінити для кожного з підрозділів із заданою середньоквадратичною похибкою σ . Тоді для i -того підрозділу $n_i \approx s_i^2/\sigma^2$, а загальний обсяг вибірки $n = \sum_{i=1}^k n_i = \sum_{i=1}^k s_i^2/\sigma^2$. Якщо можна вважати, що $s_1^2 \approx \dots \approx s_k^2 \approx S^2$, S^2 – дисперсія всієї вибірки, то $n = kS^2/\sigma^2$. Це означає, що коли бажано отримати оцінки середнього з заданою точністю для кожного з k підрозділів сукупності, то треба провести в k разів більше обстежень, ніж коли б ця умова ставилася лише для всієї сукупності в цілому. На цей факт слід звертати увагу при плануванні обстежень.

У разі, коли класифікація елементів вибірки по підрозділах може бути проведена тільки після обстеження, виникають специфічні задачі, які розглядав, наприклад, Кокрен (1976), ним же досліджено задачу визначення обсягу вибірки з погляду теорії прийняття рішень.

5.6 Виконання в пакеті STATISTICA

Визначення обсягу вибірки n проводять за відповідними точними або наближеними теоретичними формулами. Для того, щоб їх застосовувати при відомій необхідній точності оцінювання, нам потрібно знати лише значення квантилей нормального розподілу й іноді оцінку частки P при першому етапі обстеження.

Покажемо, як знаходити в пакеті STATISTICA ці показники.

Квантилі нормального розподілу

Заходимо в модуль *Basic Statistics and Tables*. Вибираємо пункт *Probability Calculator* – див. рис. 5.1. Після його відкриття вибираємо нормальний розподіл – див. рис. 5.2.

Змінюючи значення у віконці p , у віконці X отримуватимемо значення квантилей рівня p . Відповідні характеристики будуть відображатися графічно як площі під графіком щільності і на кривій, що зображає функцію розподілу $N(0,1)$ випадкової величини.

Підставивши ці числа в формули, за допомогою калькулятора легко знаходиться відповідний обсяг вибірки n . Якщо цей показник n потрібно рахувати часто, то процедуру легко запрограмувати за допомогою мови *Statistica Basic*.

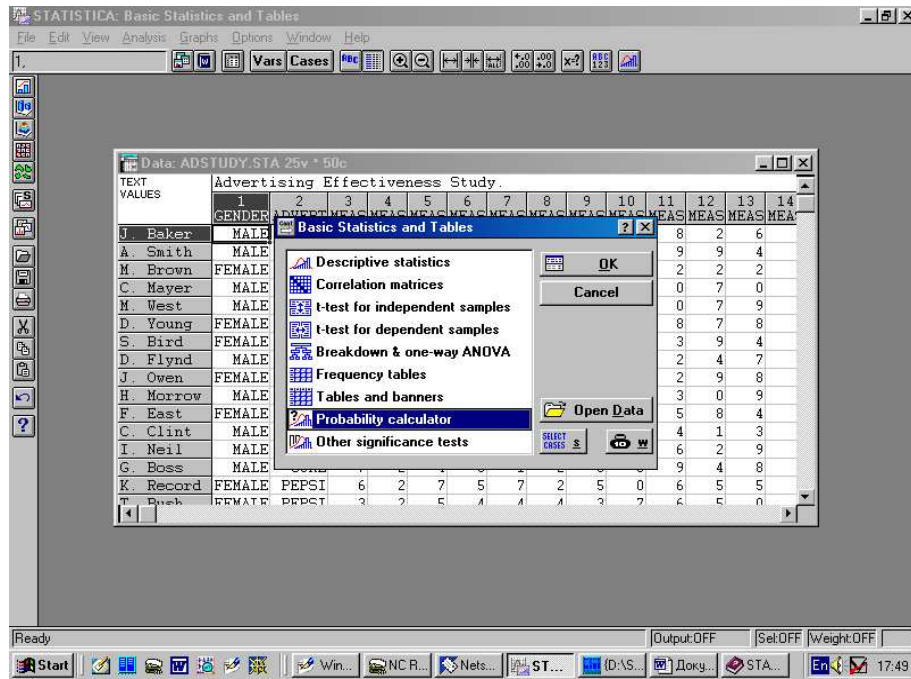


Рис. 5.1. Імовірнісний калькулятор

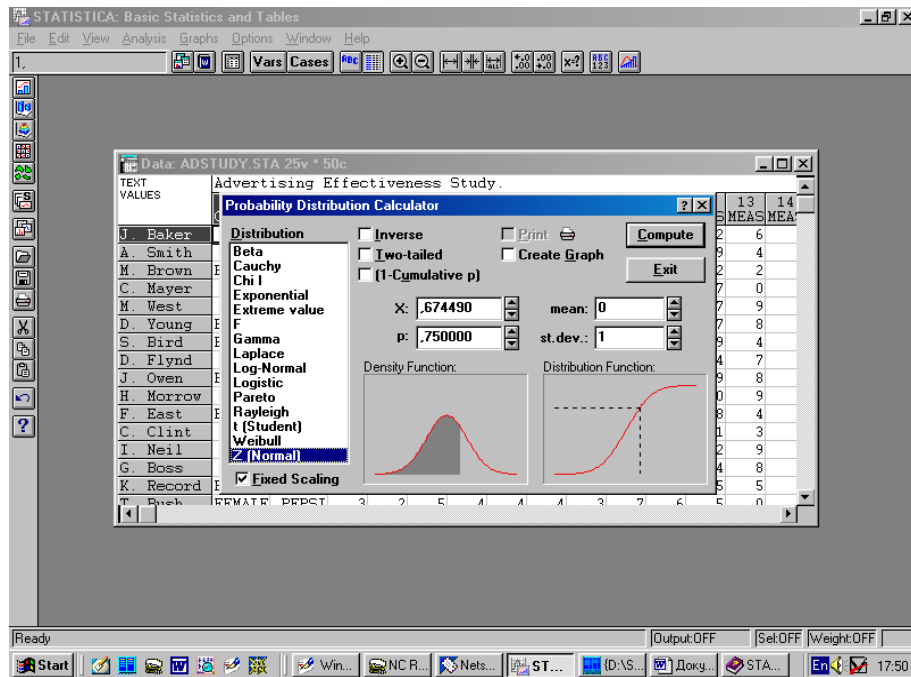


Рис. 5.2. Квантілі нормального розподілу

Розділ 6

Перевірка статистичних гіпотез

6.1 Постановка проблеми, основні поняття

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

Нехай щодо розподілу вибірки $\zeta = (\xi_1, \dots, \xi_n)$ 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 $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$) зробити висновок: розподіл ζ може бути із \mathcal{G} чи ні.

Статистичними гіпотезами будемо називати гіпотези про розподіли випадкових величин.

Нульова (основна) гіпотеза полягає у тому, що ζ має розподіл із підкласу \mathcal{G} . Позначення: H_0 . Решту гіпотез називають альтернативними чи конкурентними відносно H_0 . Позначення: H_1 .

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

Приклад. Нехай ξ має геометричний розподіл. Тоді

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

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

Якщо гіпотеза полягає у тому, що ζ 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 Θ . У протилежному випадку гіпотеза складна.

У попередньому прикладі H_0 – проста гіпотеза, а H_1 – складна.

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

Отже, потрібно 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” критерію.

Гіпотезу H_0 ми перевіряємо так: якщо реалізація вибірки $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ 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 $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ – 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:

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

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

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

Означення. Рівнем значущості α критерію \mathcal{S} називають число:

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

Рівень значущості α обмежує зверху ймовірність помилки першого роду.

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

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

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

Якщо при заданому рівні значущості α 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 значущості α має бути менший ймовірності помилки другого роду $\mathbf{P}_{H_1}(\bar{\mathcal{S}})$.

Означення. Нехай рівень значущості дорівнює α . 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 $\varphi : R^n \rightarrow R$ (статистику). Вона відображає критичну область \mathcal{S} у деяку множину $I \subset R$. Якщо при застосуванні φ до реалізації вибірки $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ $\varphi(\xi_1(\omega), \dots, \xi_n(\omega)) \in I$, то гіпотезу H_0 відхиляють, а якщо $\varphi(\xi_1(\omega), \dots, \xi_n(\omega)) \in R \setminus I$, то приймають. Звичайно розглядають такі I , що $R \setminus I = (a, b)$. (Проміжок не обов’язково скінченний. Одне з чисел a, b може бути $-\infty$ або $+\infty$ відповідно.)

6.2 Перевірка гіпотез для нормальних розподілів

6.2.1 Перевірка гіпотези про значення середнього

Випадок, коли дисперсія невідома

Нехай $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ – реалізація вибірки із генеральної сукупності з нормальним розподілом $N(a, \sigma^2)$ з невідомими параметрами a та σ^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

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

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

Нехай $t_{1-\frac{\alpha}{2}, n-1}$ – квантиль рівня $1 - \frac{\alpha}{2}$ розподілу Стюдента 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 – α . 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 значущості цього критерію – α .

Випадок, коли дисперсія відома

Якщо на відміну від попереднього випадку дисперсія σ^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 $d_{1-\frac{\alpha}{2}}$ – квантиль 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 значущості такого критерію α .

Аналогічно до попереднього при односторонній альтернативі $H_1: a > a_0$, критерій рівня значущості α :

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

6.2.2 Перевірка гіпотези про рівність середніх

Випадок, коли дисперсії невідомі

Нехай $\zeta_1 = (\xi_1, \dots, \xi_n)$ і $\zeta_2 = (\eta_1, \dots, \eta_m)$ – 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, σ^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: відхилити 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

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

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

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 Стюдента 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 значущості цього критерію – α .

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 значущості такого критерію також α .

Випадок, коли дисперсії відомі

Якщо, на відміну від попереднього випадку, дисперсії σ_1^2 та σ_2^2 для законів розподілу вибірок ζ_1 та ζ_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 значущості α і альтернативній 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 приймає в іншому випадку. При односторонній альтернативній гіпотезі $a_1 > a_2$, критерій:

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

6.2.3 Перевірка гіпотези про значення дисперсії

Випадок, коли математичне сподівання невідоме

Нехай $\zeta(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ – реалізація вибірки із генеральної сукупності з нормальним $N(a, \sigma^2)$ розподілом. Параметри a і σ^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 σ^2 . Отже, відхилення $\frac{s^2}{\sigma_0^2}$ від 1 в середньому менше, ніж відхилення $\frac{s^2}{\sigma_0^2}$ 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 χ^2 розподіл з $(n-1)$ ступенем вільності.

Нехай $\chi_{\gamma, n-1}^2$ – квантиль рівня γ розподілу χ^2 з $(n-1)$ ступенем вільності. Будемо приймати гіпотезу $H_0: \sigma^2 = \sigma_0^2$, якщо

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

і відхиляти в інших випадках. Рівень значущості такого критерію α . Цей критерій застосовуємо у випадку двосторонньої альтернативи. Для односторонньої альтернативи, наприклад $\sigma^2 > \sigma_0^2$, маємо критерій:

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

Рівень значущості такого критерію – α .

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 χ^2 розподіл з n ступенями вільності.

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

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

і відхиляємо в інших випадках. Рівень значущості такого критерію з двосторонньою альтернативою дорівнює α .

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

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

6.2.4 Перевірка гіпотези про рівність дисперсій

Випадок, коли математичні сподівання невідомі

Нехай $\zeta_1(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ і $\zeta_2(\omega) = (\eta_1(\omega), \dots, \eta_m(\omega))$ – 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

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

– незміщені слушні оцінки для параметрів σ_1^2 та σ_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: відхилити 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 $F_{\gamma, (n-1, m-1)}$ – квантиль рівня γ 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 значущості критерію α .

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 значущості α 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 значущості α 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 $\sigma_1^2 > \sigma_2^2$:

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

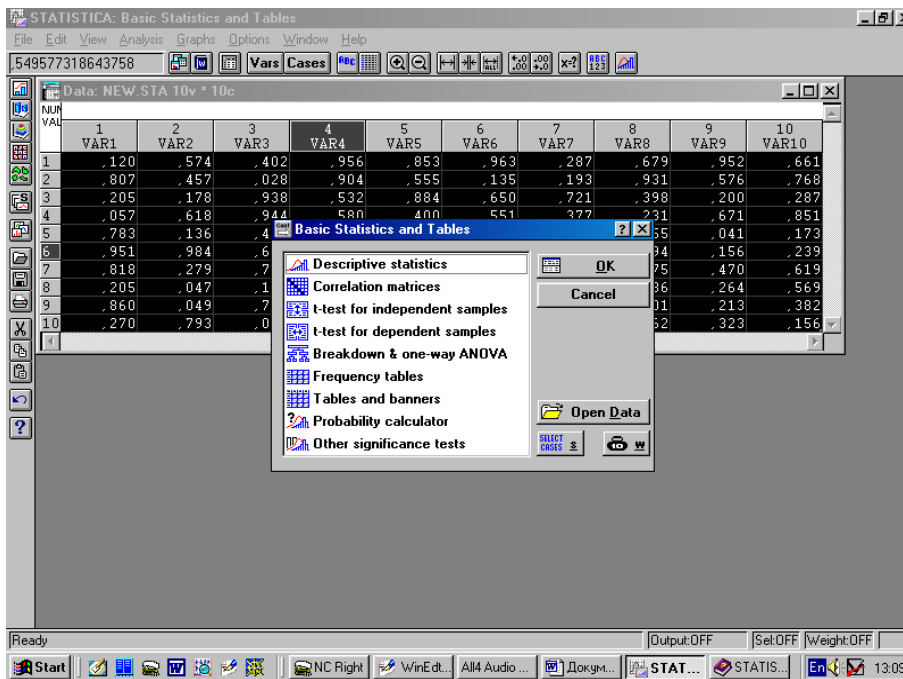


Рис. 6.1. t-критерій для незалежних вибірок

У багатьох випадках, при досить великому обсязі вибірки, запропоновані критерії використовують і для генеральних сукупностей, які не мають нормального розподілу. Це ґрунтується на тому факті, що, за центральною граничною теоремою, ξ та η розподілені асимптотично нормально.

6.3 Виконання в пакеті STATISTICA

Продемонструємо, як перевіряти різні гіпотези за допомогою пакету STATISTICA. Розглянемо два приклади – застосування t-тесту та застосування різноманітних тестів на виявлення різниці у важливих показниках розподілів.

Застосування t-тесту

t-тест знаходиться в модулі *The Basic Statistics and Tables*.

Вибираємо у вікні, що відкрилося, вихід *t-test for Independent Samples* або *t-test for Dependent Samples* – див. рис. 6.1.

t-критерій – найбільш часто використовуваний метод виявлення розбіжності між середніми двох вибірок. Наприклад, *t-критерій* можна ви-

користувати для порівняння середніх показників групи, в якій проводилась рекламна компанія, з контрольною групою, де реклама не проводилась.

Теоретично t -критерій може застосовуватися, навіть якщо обсяги вибірок дуже невеликі (наприклад, 10; деякі дослідники стверджують, що можна досліджувати вибірки меншого обсягу), і якщо змінні нормально розподілені (усередині груп), а дисперсії спостережень у групах не надто відрізняються. Припущення про нормальність можна перевірити, досліджуючи розподіл (наприклад, візуально за допомогою гістограми) або застосувавши якийсь критерій нормальності. Рівність дисперсій у двох групах можна перевірити за допомогою F -критерію чи використувати більш стійкий критерій Левена. Якщо умови застосовуваності t -критерію не виконані, варто використовувати непараметричні альтернативи t -критерію (див. розділ *Непараметричні критерії*).

p -рівень значущості t -критерію дорівнює імовірності помилково відкинути гіпотезу про рівність середніх двох вибірок, коли насправді ця гіпотеза істинна. Іншими словами, він дорівнює імовірності помилки прийняти гіпотезу про нерівність середніх, коли насправді середні рівні. Як це виглядає в пакеті STATISTICA – див. рис.6.2. Деякі дослідники пропонують, у випадку, коли розглядають відмінності тільки в одному напрямку (наприклад, розглядають альтернативу: середнє в першій групі більше (менше), ніж середнє в другій), використовувати *однобічний t -тест*.

Щоб застосувати t -критерій в пакеті STATISTICA для незалежних вибірок, потрібна, принаймні, одна незалежна (*що групує*) змінна (наприклад, стать: *чоловік/жінка*) і одна залежна змінна (наприклад, тестове значення деякого показника, оцінка, і т.д.). Вибір їх у пакеті STATISTICA видно на рис. 6.3.

За допомогою спеціальних значень незалежної змінної (ці значення називають кодами, наприклад, чоловік і жінка) дані розбиваються на дві групи. Можна зробити аналіз наступних даних за допомогою t -тесту, що порівнює середнє IQ для чоловіків і жінок.

	Стать	IQ
спостереження 1	чоловік	111
спостереження 2	чоловік	110
спостереження 3	чоловік	109
спостереження 4	жінка	102
спостереження 5	жінка	104
	середнє IQ для чоловіків = 110	
	середнє IQ для жінок = 103	

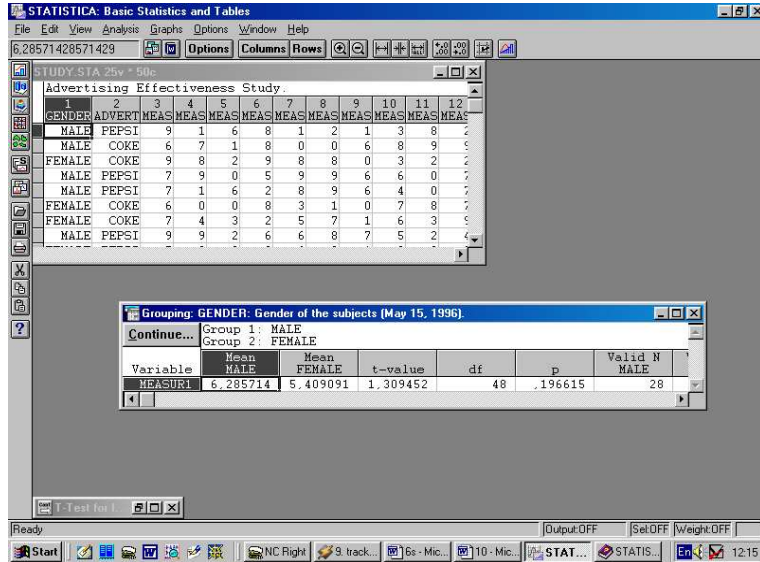


Рис. 6.2. p -рівень значущості t -критерію

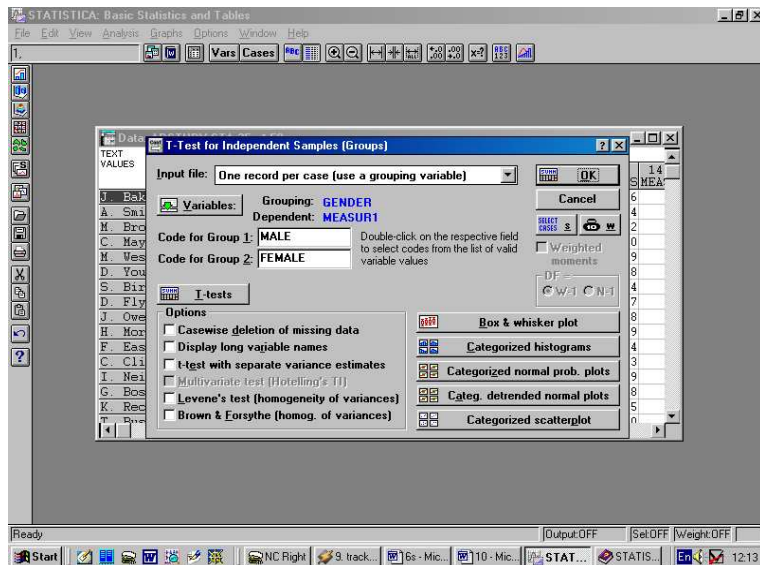


Рис. 6.3. t -критерій для незалежних вибірок

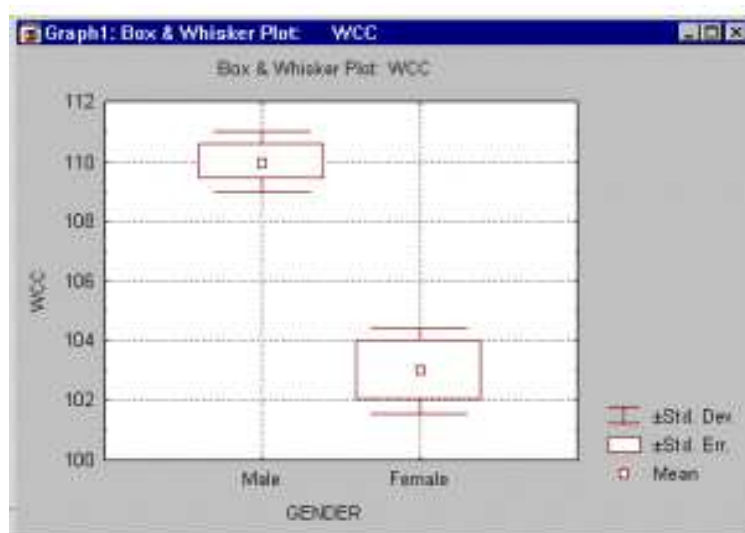


Рис. 6.4. Коробки з вусами

Візуально порівняти середні і міри відхилення від середнього в групах можна за допомогою *діаграм розмаху – коробок з вусами* (див. рис. 6.4).

На практиці часто потрібно порівнювати більше двох груп даних або порівнювати групи, створені більш ніж однією незалежною змінною. У таких більш складних дослідженнях варто використовувати *Дисперсійний аналіз*, який можна розглядати як узагальнення *t*-критерію. Фактично у випадку однофакторного порівняння двох груп дисперсійний аналіз дає результати, ідентичні *t*-критерію. Однак, якщо план істотно більш складний, слід віддати перевагу ANOVA перед *t*-критерієм (навіть якщо використовують послідовність *t*-критеріїв).

Продемонструємо використання *t*-критерію на конкретній задачі. Деякій групі людей було запропоновано оцінити в балах один із загальнодержавних телевізійних каналів за такими показниками:

Показник 1: якість інформаційних програм;

Показник 2: якість розважальних програм;

Показник 3: якість рекламних матеріалів;

Показник 4: врахування інтересів аудиторії.

При опитуванні фіксувалась стать респондентів. Було одержано такі результати:

№	Стать	1	2	3	4	№	Стать	1	2	3	4
1	Ч	8	5	3	5	19	Ч	8	5	3	5
2	Ж	5	8	6	5	20	Ж	5	8	6	5
3	Ж	4	9	5	3	21	Ж	4	9	5	3
4	Ж	6	7	4	4	22	Ч	7	7	4	4
5	Ч	7	2	4	6	23	Ж	7	3	4	6
6	Ч	9	1	1	7	24	Ч	9	1	4	7
7	Ж	3	6	2	9	25	Ж	3	6	2	9
8	Ч	5	4	4	2	26	Ч	4	4	1	2
9	Ж	2	8	4	1	27	Ж	2	6	4	1
10	Ж	4	9	2	3	28	Ч	4	9	2	3
11	Ч	9	4	2	4	29	Ж	9	5	2	4
12	Ч	5	5	2	5	30	Ч	6	4	2	3
13	Ж	6	9	6	7	31	Ч	6	8	2	7
14	Ж	2	7	9	2	32	Ж	3	8	5	2
15	Ч	9	3	5	2	33	Ж	8	3	5	2
16	Ж	3	8	8	7	34	Ж	2	8	2	7
17	Ч	8	2	4	8	35	Ч	7	3	2	8
18	Ж	5	6	8	6	36	Ж	8	5	2	6

Порівняємо середні значення оцінок, виставлених чоловіками та жінками, за всіма розглянутими показниками. Рівень значущості візьмемо $\alpha = 0,05$.

Застосуємо процедуру *t-test for independent samples* модуля Basic Statistics and Tables. Введемо таблицю даних із 5 змінних та 36 значень:

Перша змінна: стать (STAT) – М, F;

Друга змінна: бал за якість інформаційних програм (INF) – 0-9;

Третя змінна: бал за якість розважальних програм (ROZV) – 0-9;

Четверта змінна: бал за якість реклам (RECL) – 0-9;

П'ята змінна: бал за врахування інтересів аудиторії (INT) – 0-9.

У вікні, що відкриється, виберемо *Input file: One record per case (use a grouping variable)*, змінні *Variables (Grouping variable: 1-STAT, Dependent variables: 2-INF – 5-INT)*. Виберемо коди (*Code for group 1:*

M, Code for Group 2: F) та опції – Options (*t-text with separate variance estimates, Multivariate test (Hotelling's TI)*). Результатом роботи процедури є така таблиця:

<i>Grouping: STAT Group 1: M; Group 2: F</i>							
<i>TI(casewise MD)=26,0749 F(4,31)=5,9436 p<0,00114</i>							
Variable	Mean M	Mean F	t-value	df	p	t separ. var.est.	df
INF	6,94	4,55	3,53	34	0,0012	3,62	33,99
ROZV	4,19	6,90	-3,90	34	0,0004	-3,80	28,50
RECL	2,81	4,55	-2,84	34	0,0076	-3,01	30,79
INT	4,88	4,60	0,35	34	0,7271	0,36	33,81

p 2-sided	Val. N M	Val. N F	Std.D. M	Std.D. F	F-ratio variance	p variance
0,0009	16	20	1,77	2,19	1,53	0,4076
0,0007	16	20	2,32	1,86	1,55	0,3652
0,0052	16	20	1,22	2,19	3,20	0,0264
0,7225	16	20	2,13	2,48	1,36	0,5501

Результати обчислень дають змогу стверджувати, що для двох вибірок (чоловіча і жіноча) дисперсії змінної RECL можна вважати різними (p variances = 0,0264), різниці між оцінками дисперсій (Std.D.) інших змінних більше випадкові, ніж пов'язані з відмінностями дисперсій (p variances > 0,05). Звідси потрібно вважати статистично значущими різниці між середніми змінних INF (p = 0,0012), ROZV (p = 0,004), RECL (p 2-sided = 0,0052). Крім того, маємо результати багатовимірного тесту на відмінність векторів середніх обох вибірок. При рівні значущості $p < 0,00114$ вектори середніх є різними.

Для справедливості зауважимо, що обсяги обох вибірок є недостатньо великими, щоб одержані результати можна було вважати якісними.

t-критерій для залежних вибірок

Ступінь розбіжності між середніми в двох групах залежить від *внутрішньогрупової варіації* (дисперсії) змінних. Залежно від того, наскільки різні ці значення для кожної групи, різниця між груповими середніми показує більш сильний чи більш слабкий ступінь залежності між незалежною (що *групує*) і залежною змінними. Наприклад, якщо середне

IQ дорівнювало 102 для чоловіків і 104 для жінок, то різниця внутрішньогрупових середніх тільки на величину 2 буде надзвичайно важлива, коли всі значення IQ чоловіків лежать в інтервалі від 101 до 103, а всі значення IQ жінок – в інтервалі 103 – 105. У цьому випадку можна досить добре передбачити IQ (значення залежної змінної), виходячи зі статі суб'єкта (незалежної змінної). Однак якщо та ж різниця 2 отримана із сильно розкиданих даних (наприклад, що змінюються в межах від 0 до 200), то цією різницею цілком можна знехтувати. Таким чином, можна сказати, що зменшення внутрішньогрупової варіації збільшує чутливість критерію.

t-критерій для залежних вибірок корисний у тих ситуаціях, які часто виникають на практиці, коли важливе джерело внутрішньогрупової варіації (чи помилки) може бути легко визначене і вилучене з аналізу. Наприклад, це стосується експериментів, у яких дві порівнювані групи ґрунтуються на одній і тій же сукупності спостережень (суб'єктів), які тестували двічі (наприклад, до і після агітації, до і після виборів). У подібних експериментах значна частина внутрішньогрупової мінливості (варіації) в обох групах може бути пояснена індивідуальними розбіжностями суб'єктів. Зауважимо, що насправді така ситуація не надто відрізняється від тієї, коли порівнювані групи зовсім незалежні (див. *t*-критерій для незалежних вибірок), де індивідуальні відмінності також мають внесок у дисперсію помилки. Однак у випадку незалежних вибірок, неможливо з цим нічого зробити, тому що неможливо визначити (чи “видалити”) частину варіації, зв'язану з індивідуальними розбіжностями суб'єктів. Якщо ж ту саму вибірку тестують двічі, то можна вилучити цю частину варіації. Замість дослідження кожної групи окремо й аналізу початкових значень, можна розглядати просто різницю між двома вимірами (наприклад, “до прийому ліків” і “після прийому ліків”) для кожного суб'єкта. Віднімаючи перші значення від других (для кожного суб'єкта) і аналізуючи потім тільки ці “чисті (парні) різниці”, вилучають ту частину варіації, яка є результатом розбіжності початкових рівнів індивідуумів. Саме так і проводять обчислення в *t*-критерії для залежних вибірок. У порівнянні з *t*-критерієм для незалежних вибірок, такий підхід дає завжди “кращий” результат (критерій стає чутливішим).

Теоретичні припущення *t*-критерію для незалежних вибірок стосуються також до критерію для залежних вибірок. Це означає, що попарні різниці мають бути нормально розподілені. Якщо ж це не так, то можна скористатися одним із альтернативних непараметричних критеріїв.

Можемо застосовувати *t*-критерій для залежних вибірок до будь-якої пари змінних у наборі даних. Зауважимо, що застосування цього критерію невиправдане, якщо значення двох змінних непорівнянні. Напри-

клад, якщо ви порівнюєте середнє споживання у вибірці людей до і після реклами, але використовуєте різні методи обчислення кількісного показника або одиниці виміру, то високо значущі результати t -критерію можуть бути отримані штучно, саме за рахунок зміни одиниць виміру. Наступний набір даних може бути проаналізований за допомогою t -критерію для залежних вибірок.

	до	після
спостереження 1	111.9	113
спостереження 2	109	110
спостереження 3	143	144
спостереження 4	101	102
спостереження 5	80	80.9
...
	середня різниця між “до” і “після” = 1	

Середня різниця між показниками в двох стовпцях відносно мала ($d = 1$) у порівнянні з розкидом даних (від 80 до 143, у першій вибірці). Проте t -критерій для залежних вибірок використовує тільки парні різниці, “ігноруючи” вихідні чисельні значення і їхню варіацію. Таким чином, величина цієї різниці і буде порівнюватися не з розкидом вихідних значень, а з розкидом індивідуальних різниць, який відносно малий: 0.2 (від 0.9 у спостереженні 5 до 1.1 у спостереженні 1). У цій ситуації різниця 1 дуже велика і може привести до значущого t -значення.

t -критерій для залежних вибірок може бути обчислений для списків змінних і переглянутий далі як матриця. Пропущені дані при цьому обробляють або *пострічково*, або попарно, так само, як при обчисленні кореляційних матриць. Застереження при обчисленні матриць t -критеріїв за наявності пропусків:

- можлива поява артефактів (штучних результатів) через попарне видалення пропусків у t -критерії;
- можливе виникнення чисто “випадкових” значущих результатів.

Розглянемо конкретний приклад застосування t -критерію.

У групі водіїв проводиться навчання з удосконалення знань правил дорожнього руху. До навчання, після першого та другого тижнів навчання проводилося тестування. Результати тестування – кількість помилок

допущених кожним водієм у відповіді на тестові завдання, наведено в таблиці:

№	До	1 тиж	2 тиж	№	До	1 тиж	2 тиж
1	8	5	4	2	6	5	5
3	7	4	6	4	8	6	4
5	9	8	6	6	8	5	5
7	8	3	4	8	5	6	4
9	7	5	4	10	6	5	4
11	7	3	4	12	9	8	7
13	7	6	5	14	5	3	3
15	8	6	6	16	9	5	6
17	6	6	6	18	8	5	5
19	7	7	6	20	6	6	5

Визначити, чи відрізняються результати до і після кожного тижня навчання (рівень значущості прийняти рівним 0,05).

Введемо дані в пакеті STATISTICA. Для цього створимо три змінні з 20 значеннями. В першу змінну VAR1 занесемо кількості помилок, допущених до навчання, в другу VAR2 – після першого тижня і в третю VAR3 – після другого тижня навчання.

Виберемо в модулі *Basic Statistics and Tables* процедуру *t-test for dependent samples*. У вікні цієї процедури виберемо два списки змінних *Variables: (First variable list: 1-VAR1, Second variable list (optional): 2-VAR2 – 3-VAR3)*. Натиснувши T-test, одержимо таблицю результатів:

<i>T-test for Dependent Samples</i>								
<i>Marked differences are significant at p<0,05</i>								
	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
VAR1*	7,20	1,24						
VAR2*	5,35	1,42	20	1,85	1,57	5,29	19	0,00
VAR1*	7,20	1,24						
VAR3*	4,95	1,05	20	2,25	1,16	8,64	19	0,00

З таблиці видно, що середні пар змінних VAR1 і VAR2, VAR1 і VAR3 є суттєво різними (рівень значущості p значно менший 0,05). Порівняємо тепер середні змінних VAR2 і VAR3 між собою. Проробимо все те ж саме, тільки виберемо в першому списку змінну VAR2, а в другому списку змінну VAR3. Результат розміщено в таблиці:

<i>T-test for Dependent Samples</i>								
<i>Marked differences are significant at $p < 0,05$</i>								
	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
VAR2	5,35	1,42						
VAR3	4,95	1,05	20	0,40	1,10	1,63	19	0,12

Отже, немає підстав стверджувати, що середні змінних VAR2 і VAR3 різні ($p > 0,05$).

Якщо маємо більше двох “залежних вибірок”, то можна використувати дисперсійний аналіз із *повторними вимірами*. Повторні виміри в дисперсійному аналізі (ANOVA) можна розглядати як узагальнення t -критерію для залежних вибірок, що дозволяють збільшити чутливість аналізу. Наприклад, можна одночасно контролювати не тільки базовий рівень залежної змінної, але й інші фактори, а також включати в план експерименту більше однієї залежної змінної (багатовимірний дисперсійний аналіз MANOVA).

Тестування наявності різниці у показниках розподілів

Заходимо в модуль *The Basic Statistics and Tables*.

Вибираємо у вікні, що відкрилося, вихід *Other Significance Tests* – див. вище рис. 6.1.

Отримуємо вікно з трьома тестами на визначення наявності різниці у показниках розподілів – див. рис. 6.5.

Можемо провести тест на виявлення різниці двох коефіцієнтів кореляції, математичних сподівань або ймовірностей.

Для того, щоб провести обчислення, потрібно знати значення оцінок відповідних параметрів та обсяги відповідних вибірок. Кількості спостережень визначаються нашими масивами даних, а оцінки відповідних параметрів рахуємо в пакеті так, як про це розповідалося раніше (все оцінювання необхідних параметрів проводимо в цьому ж модулі *Basic Statistics and Tables*).

Наприклад, знаходження кореляційної матриці:

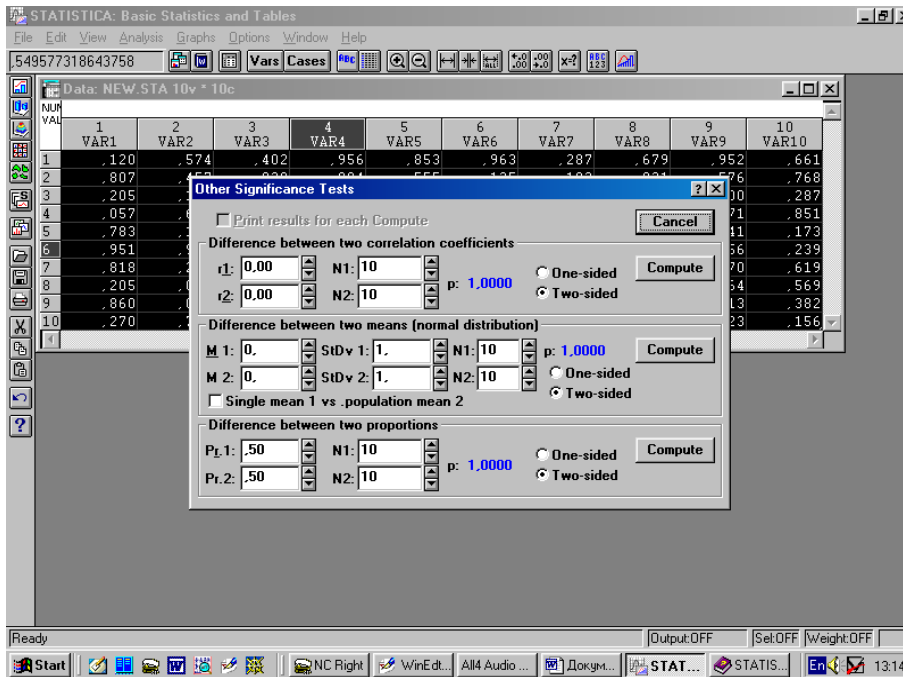


Рис. 6.5. Тестування наявності різниці у показниках розподілів

- Заходимо в *Analysis*;
- *Quick Basic Stats*;
- *Correlation matrices*;
- *Variables for the analyses – First list: All – Second list: All*;
- *OK*.

Друге значення знаходимо в результаті аналогічних процедур, але з іншими даними.

Заповнюємо відповідні віконця цими значеннями. Вибираємо кнопку для проведення одностороннього чи двостороннього тестування. Натискаємо *Compute*.

Біля “р” з’являється ймовірність, з якою ми можемо прийняти гіпотезу про рівність відповідних показників, див., наприклад, рис. 6.6.

Якщо ця ймовірність близька до 1, то приймаємо гіпотезу про рівність відповідних показників. Якщо ж вона близька до 0, то вважаємо що цю гіпотезу відхиляємо.

Збільшуючи кількості елементів у вибірках у відповідних віконцях, за тих же значень параметрів і перераховуючи ймовірності, помічаємо, що збільшення кількості елементів веде до більш точного розрізнення параметрів (ймовірності “р” прийняття гіпотез про рівність показників зменшуються), чого і слід було сподіватись із емпіричних міркувань.

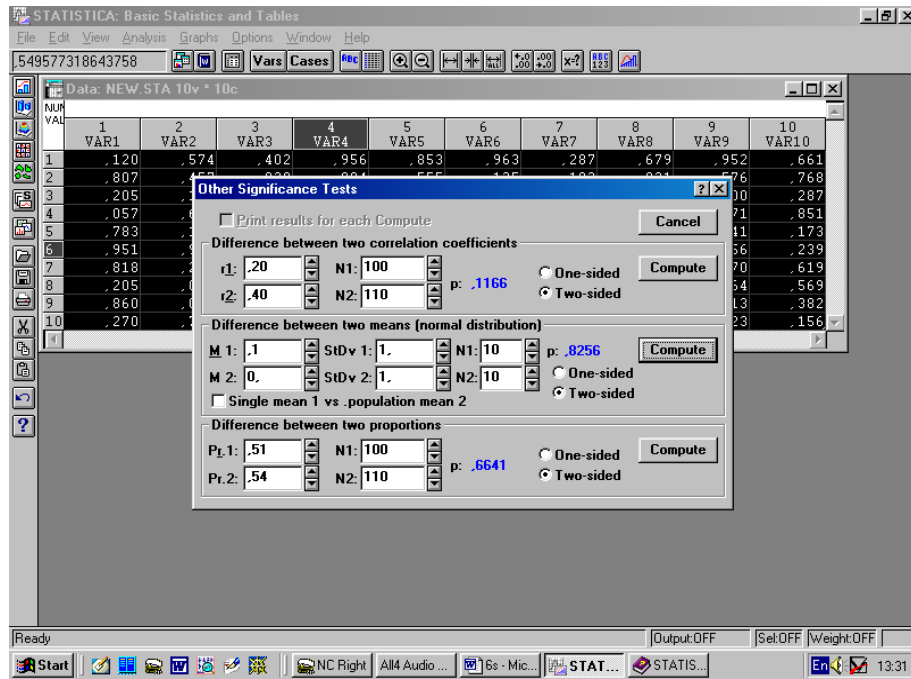


Рис. 6.6. Імовірності прийняття гіпотез про рівність показників

Розділ 7

Непараметричні критерії

Розглянуті раніше критерії перевірки статистичних гіпотез ґрунтувались 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 здебільшого 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 загальніших припущеннях та використовують простіші обчислення.

7.1 Критерій знаків

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

Будемо вважати, що розподіли F_ξ та F_η абсолютно неперервні, але нам не відомі. Якщо наші вибірки отримані із однорідних генеральних

сукупностей, то

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

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

Статистика критерію знаків – кількість знаків “+” чи “-” у послідовності знаків різниць $\xi_i - \eta_i$, $i = 1, 2, \dots, k$. 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 α – рівень значущості критерію.

Тоді гіпотезу H_0 відхиляють, якщо

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

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

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

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

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

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

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

ПРИКЛАД. Проводилось опитування двох соціальних груп про майбутній бюджет. Результати були такі (за 100-бальною шкалою):

v_1	70	85	63	54	65	80	75	95	52	55
v_2	72	86	62	55	63	80	78	90	53	57

Чи можна за такими результатами при рівні значущості $\alpha = 0,1$ стверджувати, що друга група дає завищену оцінку?

Знак різниці $v_1 - v_2$ набуває значень -, -, +, -, +, 0, -, +, -, -. Ненульових різниць $k = 9$, додатніх $r = 3$. Перевіримо гіпотезу H_0 про те, що різниця результатів опитування зумовлена випадковими помилками.

Альтернативна гіпотеза полягає у тому, що оцінки другої групи завищені.

Отже, $H_0: p = \frac{1}{2}$, $H_1: p < \frac{1}{2}$. Оскільки $\hat{F}_2 = \frac{9-3}{3+1} = 1,5$ і $F_{0,9; (8,12)} = 2,24$, то гіпотезу H_0 не відхиляють.

7.2 Критерій Вілкоксона

Нехай $\bar{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ та $\bar{\eta} = (\eta_1, \eta_2, \dots, \eta_m)$ – реалізації незалежних вибірок із 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 θ – невідомий параметр.

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

Розмістимо вибірки $\xi_1, \xi_2, \dots, \xi_n$ і $\eta_1, \eta_2, \dots, \eta_m$ у спільний варіаційний ряд. Для кожного ξ_i (η_j) визначимо його ранг як номер місця, на якому стоїть ξ_i (η_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, \dots, r_n – ранги ξ_1, \dots, ξ_n . Тоді:

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

Величина W описує міру “змішаності” значень вибірок. Якщо W велике (більшість вибіркових значень ξ_i розміщені праворуч значень η_j) чи мале (більшість значень ξ_i розміщені ліворуч значень η_j), то “змішаність” незначна, інакше “гарна” (значна).

Мінімально можливе значення статистики W : $1 + 2 + \dots + n = \frac{n(n+1)}{2}$, максимально можливе: $(m+1) + (m+2) + \dots + (m+n) = \frac{(2m+n+1)n}{2}$, “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 $H_0 : \theta = 0$ ($\mathbf{F} \equiv \mathbf{G}$) істинна, то $\bar{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ і $\bar{\eta} = (\eta_1, \eta_2, \dots, \eta_m)$ – незалежні вибірки з одного розподілу і змішаність “гарна”. W тоді близьке до середнього значення $\frac{n(n+m+1)}{2}$. В іншому випадку W істотно відхиляється від $\frac{n(n+m+1)}{2}$.

При заданому рівні значущості α 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 істотно, 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 $n, m \rightarrow +\infty$ 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_{α} – квантиль розподілу $N(0, 1)$.

7.3 Виконання в пакеті STATISTICA

Розглянемо приклад перевірки гіпотези за допомогою непараметричних критеріїв у пакеті. Більшість із них зібрані в модулі *Nonparametrics & Distributions*. Зокрема, відкривши цей модуль, у розділі *Nonparametric Statistics* ми зразу бачимо такі критерії перевірки гіпотез: χ^2 критерій, Вольда-Вольфовіца, Манна-Уїтні, критерій знаків, тест Вілкоксона та інші – див. рис. 7.1.

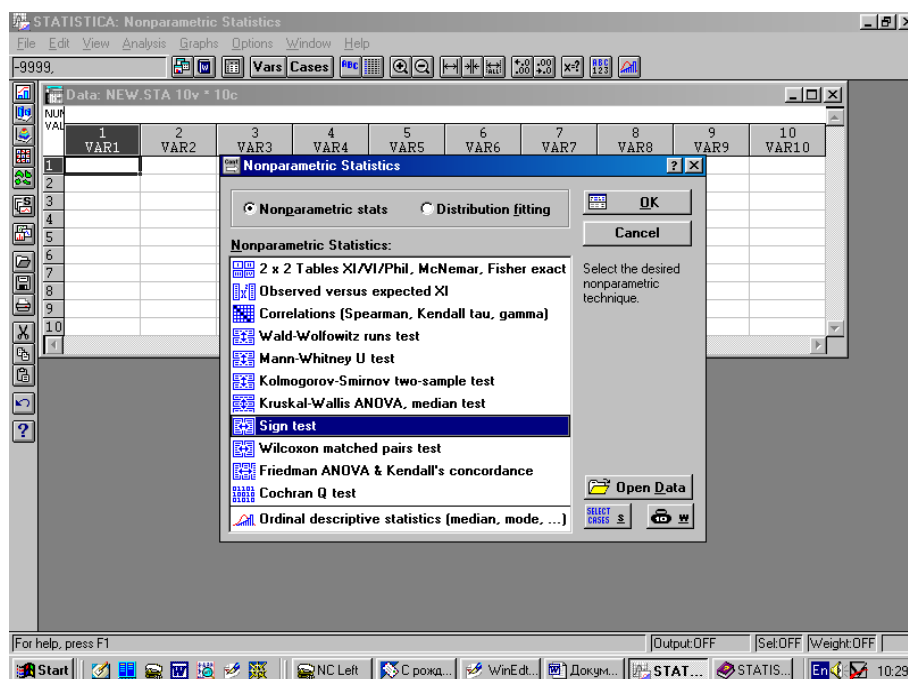


Рис. 7.1. Непараметричні критерії

Продемонструємо, наприклад, як використовувати критерій знаків. Заходимо в модуль *Nonparametrics & Distributions*.

Вибираємо опцію *Nonparametric Statistics* і в ній критерій знаків – див. рис. 7.1. У вікні, що відкривається, натискаємо кнопку *Variables* і вибираємо, наприклад, третю та четверту для аналізу за критерієм знаків – див. рис. 7.2.

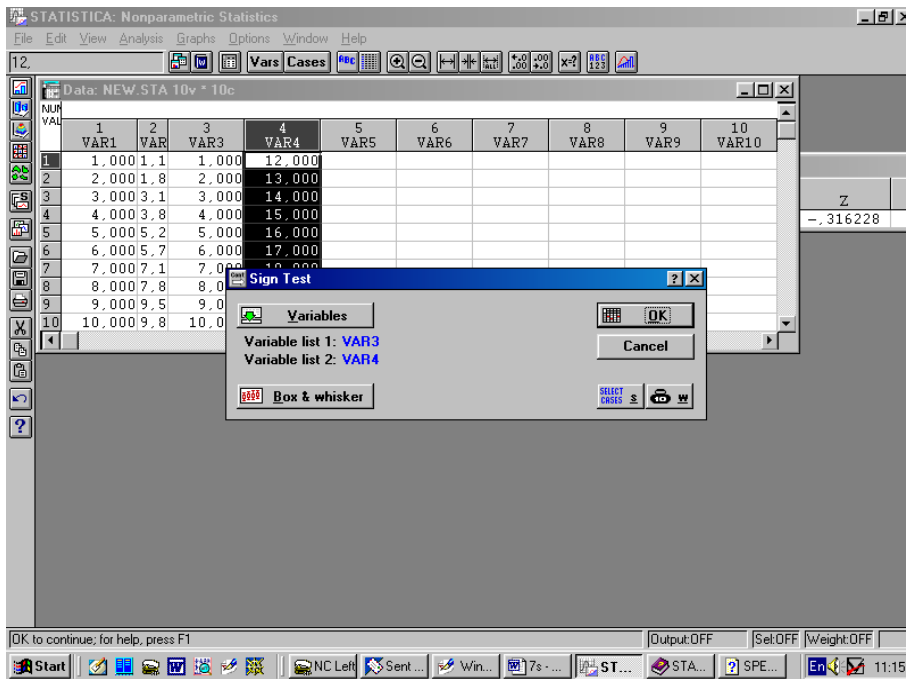


Рис. 7.2. Критерій знаків

Для початкового візуального аналізу можна використати опцію *Box and whisker plot*. Отримуємо коробки з вусами, які дають графічне уявлення про розподіл наших двох величин – див. рис. 7.3.

Для проведення самого критерію знаків повернемося до панелі критерію знаків. Натискаємо ОК і отримуємо таблицю з результатами аналізу (див. рис. 7.4).

Перша клітинка таблиці показує, скільки значень не збігаються у двох змінних і будуть використані для обчислення відповідної статистики у критерії знаків. Друга клітинка таблиці дає відсоток значень однієї змінної, які більші за значення другої змінної.

Якщо змінні не відрізняються одна від іншої (випадкові величини з однієї і тієї ж самої генеральної сукупності), то слід би було чекати 50% перевищень значень однієї над іншою.

Наступна клітинка дає значення статистики, а в четвертій клітинці отримуємо ймовірність, з якою приймаємо нашу гіпотезу. Якщо ця ймовірність близька до 1, то вважаємо, що змінні не відрізняються, а якщо

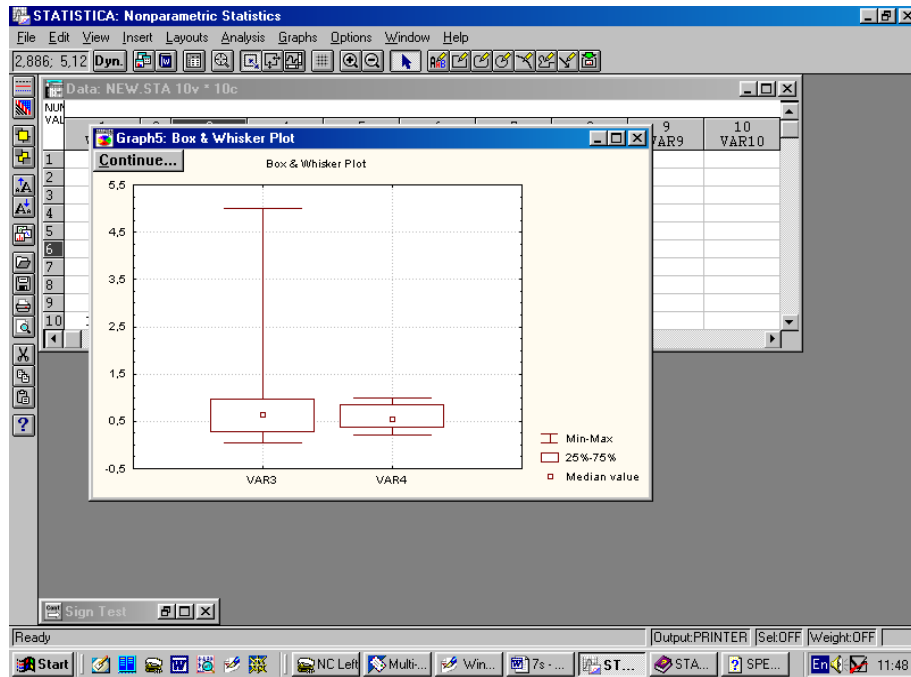


Рис. 7.3. Графічне зображення розподілу величин

The screenshot shows the 'Sign Test (new.sta)' dialog box with the following results:

	No. of Non-ties	Percent $v < V$	Z	p-level
VAR3 & VAR4	10	40,00000	,316228	,751830

The background data table shows the following values for variables VAR1 through VAR10 across 10 rows:

NUN	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
1	1.000	1.0	1.000	,434						
2	2.000									
3	3.000									
4	4.000									
5	5.000									
6	6.000									
7	7.000	7.1	,483	,362						
8	8.000	7.8	,965	,985						
9	9.000	9.5	,675	,213						
10	10.000	9.8	,619	,502						

Рис. 7.4. Результати аналізу критерію знаків

близька до 0, то приймаємо гіпотезу про значну різницю у розподілах двох показників.

ПРИКЛАД. Досліджувалась популярність естрадного співака до (VAR1) і після (VAR2) гастролей. Одержані результати (у %) в 25 регіонах:

12,0	26,0	15,0	24,0	25,0	29,0	14,0	16,0	12,0	13,0
14,0	20,0	23,0	21,0	25,0	30,0	24,0	28,0	16,0	15,0
17,0	20,0	13,0	21,0	8,00	10,0	6,00	9,00	13,0	12,0
14,0	13,0	24,0	25,0	31,0	34,0	26,0	29,0	25,0	16,0
18,0	12,0	13,0	16,0	7,00	9,00	14,0	11,0	6,00	4,00

Чи можна стверджувати, що гастролі вплинули на популярність співака? Дані наведено парами (VAR1 VAR2) для кожного регіону.

Розв'яжемо задачу з використанням процедури *Sign test* пакету *Nonparametric Statistics*. Зайдемо в пакет, створимо файл даних та виберемо відповідну процедуру (рис. 7.5).

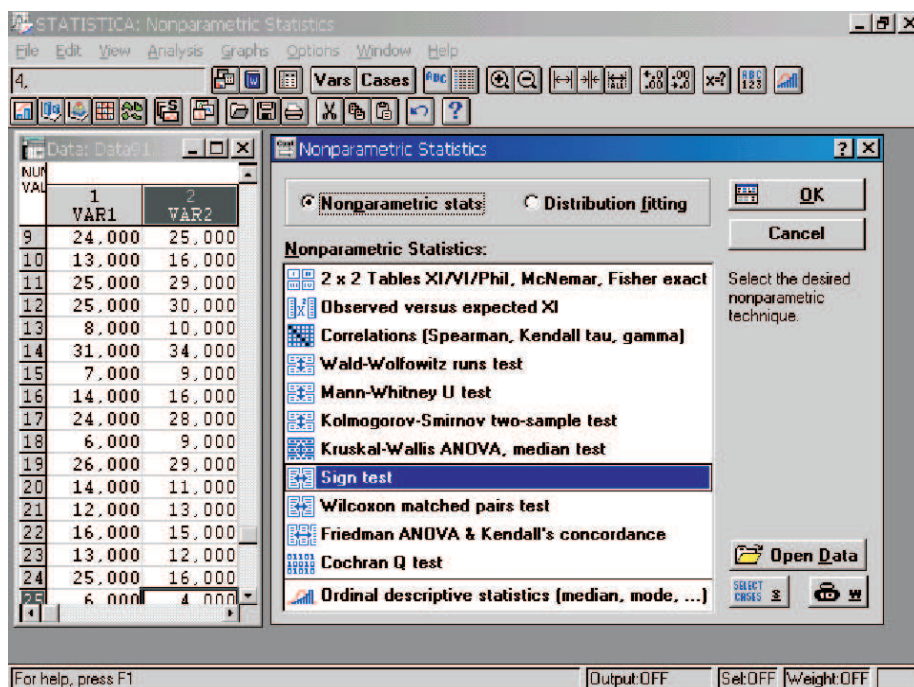


Рис. 7.5. Застосування критерію знаків

У вікні, що відкриється, виберемо *Variables (First variable list:1, Second variable list:2)* (рис. 7.6).

Натиснувши ОК, одержимо результат (рис. 7.7).

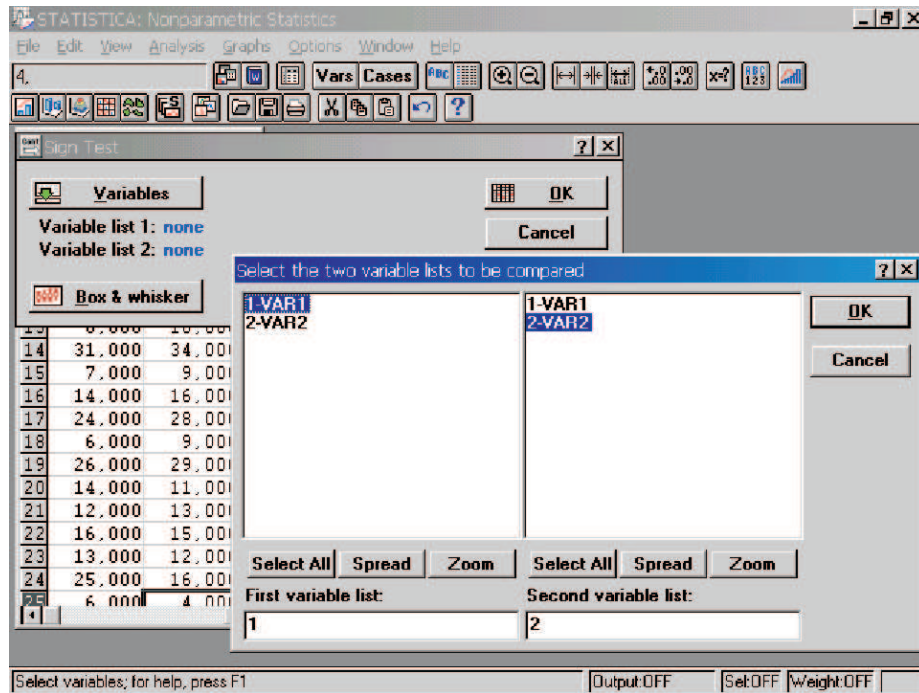


Рис. 7.6. Вибір змінних

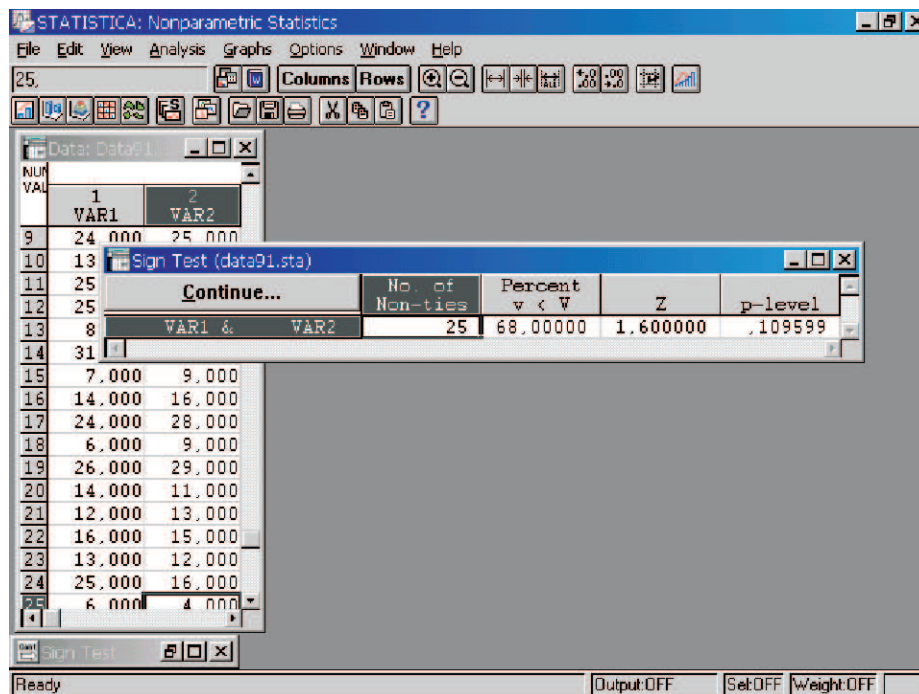


Рис. 7.7. Вікно результатів

На цьому рисунку *No. of Non-ties* – кількість варіантів з різними значеннями, *Percent $v < V$* – процент варіантів, у яких перше значення більше другого (процент знаків “+”), *Z* – значення статистики критерію, *p-level* – рівень значущості, при якому гіпотеза H_0 не суперечить вибірці.

Отже, вибірка не дає підстав вважати, що гастролі не вплинули на популярність співака. Ймовірність помилки при цьому становить 0,109599.

Зауважимо, що цю та подібні задачі можна було б розв’язати з допомогою критерію Вілкоксона (*Wilcoxon matched pairs test*). Результати, одержані з допомогою цього критерію, наведені на рис. 7.8.

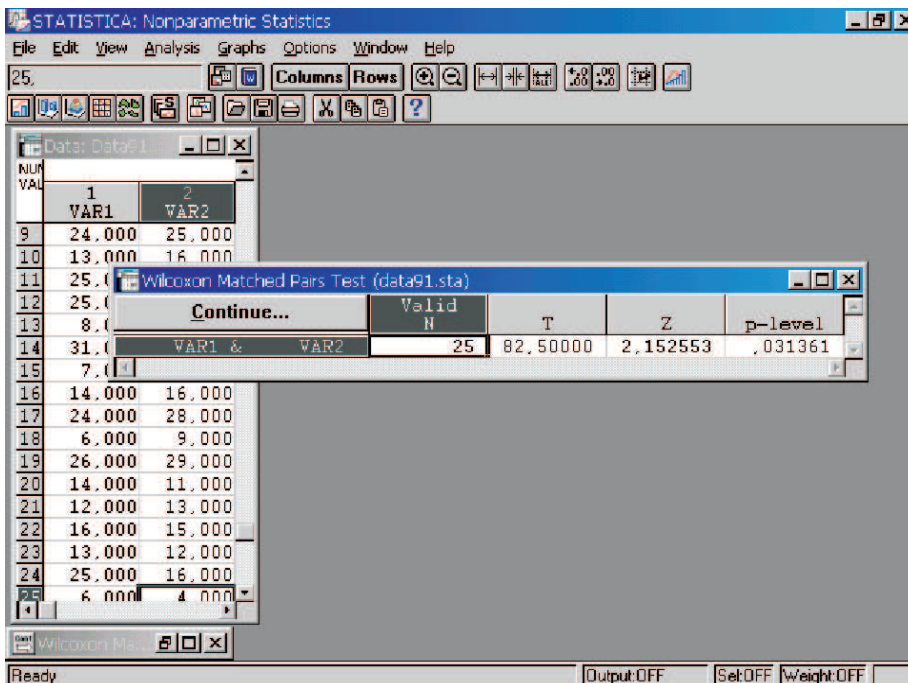


Рис. 7.8. Критерій Вілкоксона

Щоб одержати результат, потрібно виконати аналогічні операції, як і при використанні критерію знаків.

Як видно, критерій Вілкоксона дає меншу ймовірність помилки (*p-level*), ніж критерій знаків.

7.4 Критерій Манна і Уїтні

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 елементові 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_B статистики критерію є менше з чисел w_1 та w_2 ($W = \min(w_1, w_2)$). У статистичних таблицях наводяться ймовірності $p = P(W < x/H_0)$ за умови, 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 (за умови, що гіпотеза H_0 правильна) приблизно стандартний нормальний розподіл $N(0, 1)$. У цьому випадку гіпотезу H_0 відхиляють на рівні значущості α , якщо вибіркове значення Z_B статистики Z задовольняє нерівність (при двосторонній альтернативній гіпотезі)

$$|Z_B| > u_{1-\frac{\alpha}{2}}.$$

7.5 Виконання в пакеті STATISTICA

Нехай на екзамені з теорії ймовірностей в двох групах були отримані такі оцінки:

Перша група: 5, 3, 4, 2, 5, 4, 4, 3, 4, 3, 3, 5, 2, 4, 4, 3, 5, 2, 2, 4, 5, 3, 5, 4, 2, 3, 3, 3, 5, 2, 4, 4;

Друга група: 2, 5, 2, 4, 2, 4, 3, 3, 3, 2, 3, 2, 3, 4, 3, 4, 3, 2, 3, 5, 3, 4, 4, 3, 3, 2.

Чи можна стверджувати, що в цих групах абсолютна успішність є різною?

Сформуємо дві змінні VAR1 – об'єднана вибірка оцінок, VAR2 – номер групи, в якій відповідна оцінка виставлена. Виберемо в *Nonparametric Statistics* пункт *Mann-Whitney U test* (рис. 7.9). Натиснувши *Vari-*

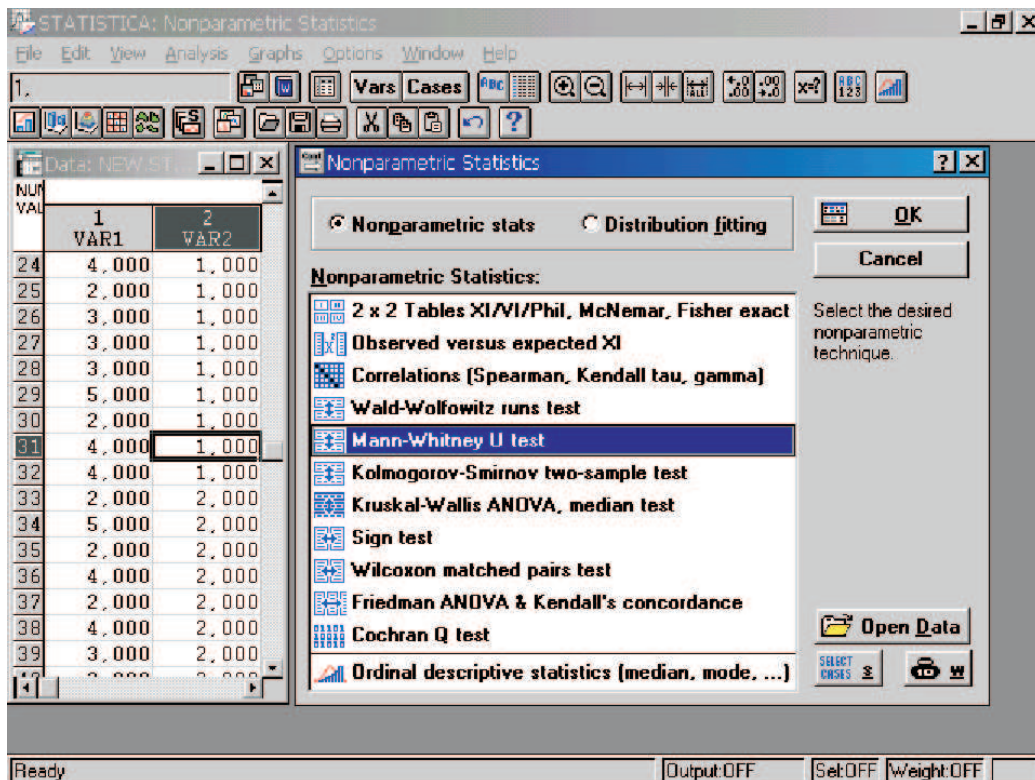


Рис. 7.9. Критерій Манна-Уїтні

ables, виберемо групуючу змінну (*Indep. (grouping) variable: 2-VAR2*) та залежну змінну (*Dependent variable list: 1-VAR1*) (рис. 7.10).

Результат роботи наведено в таблиці, зображеній на рисунку 7.11.

В одержаній таблиці *Rank Sum Group 1* – сума рангів першої вибірки, *Rank Sum Group 2* – сума рангів другої вибірки, *U* – вибіркоче значення статистики *W*, *Z* – вибіркоче значення статистики *Z*, *p-level* – рівень

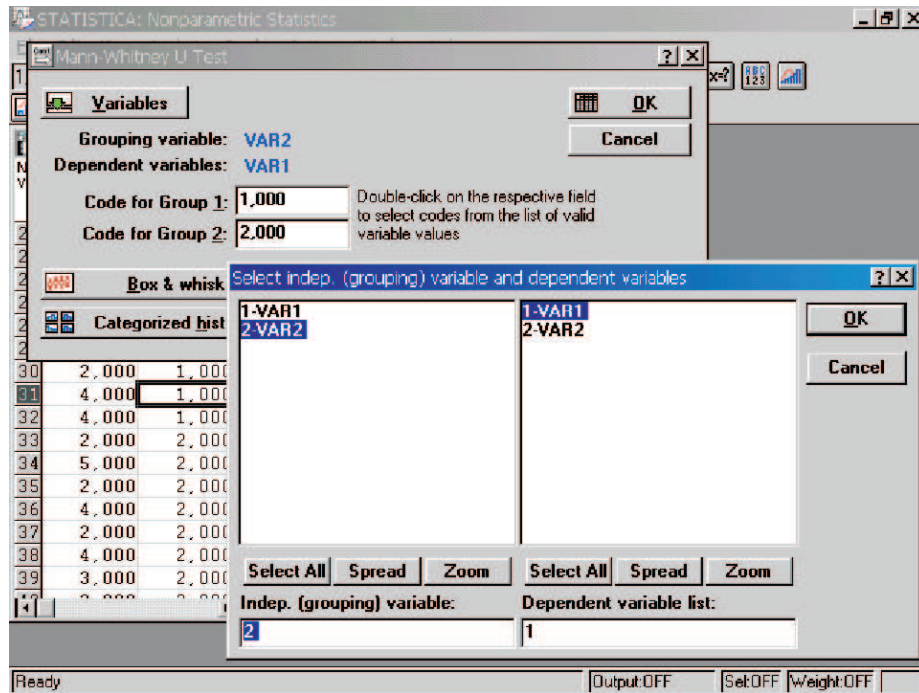


Рис. 7.10. Вибір змінних

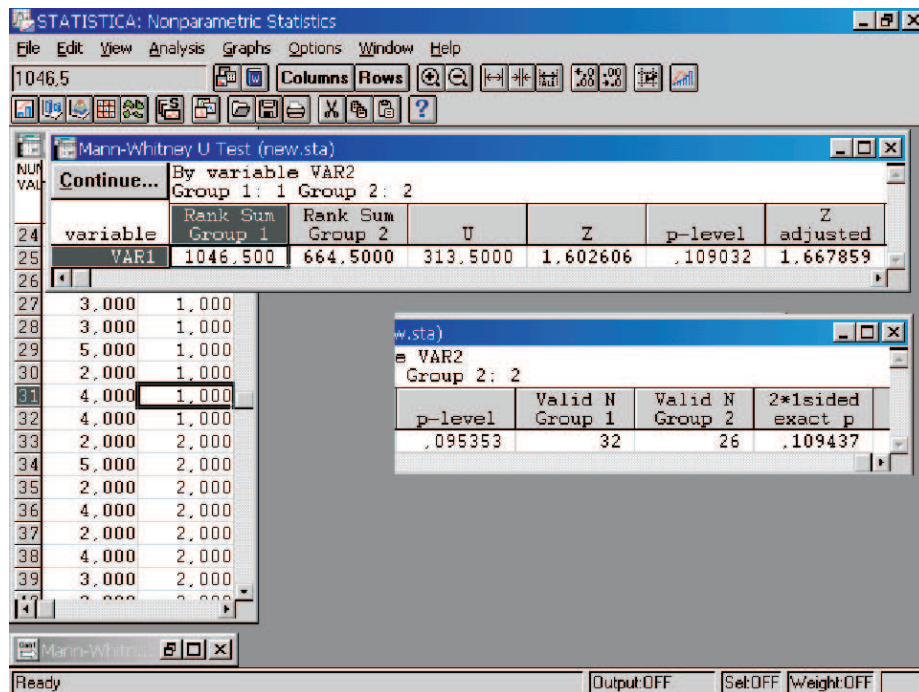


Рис. 7.11. Результати критерію Манна-Уїтні

значущості, при якому можна вважати, що гіпотеза H_0 не суперечить статистичним даним, Z adjusted – виправлене значення статистики Z та відповідний рівень значущості (p -level), *Valid N Group 1* – кількість елементів першої вибірки (n_1), *Valid N Group 2* – кількість елементів другої вибірки (n_2), *2*1sided exact p* – точне значення рівня значущості при двосторонній альтернативній гіпотезі.

Результати, одержані нами, дають змогу стверджувати, що немає підстав вважати абсолютні успішності в розглянутих групах різними. При цьому ймовірність помилитися не перевищує 0,1.

7.6 Рангова кореляція

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

Вибірковим значенням рангового коефіцієнта кореляції Спірмена ρ_s називають величину

$$r_s = 1 - \frac{6 \sum_{i=1}^n (\xi'_i - \eta'_i)^2}{n(n^2 - 1)}. \quad (7.1)$$

Ранговий коефіцієнт кореляції ρ_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 Стюдента з $n-2$ ступенями вільності. Отже, при заданому рівні значущості α , гіпотезу H_0 відхиляють, якщо

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

тобто між випадковими величинами є рангова кореляційна залежність. Приклад. При рівні значущості $\alpha = 0,1$ перевірити гіпотезу про існування рангової кореляційної залежності випадкових величин за такими вибірками

ξ_i	68,8	63,3	75,5	67,2	71,3	72,8	76,5	63,5	69,9	71,4
η_i	167	113,3	159,9	153,6	150,8	181,2	173,1	115,4	125,6	166,2

Обчислимо вибіркове значення рангового коефіцієнта кореляції. Якщо впорядкувати значення ξ_i , η_i з таблиці так, щоб ξ_i йшли у порядку зростання, то отримуємо таблицю

ξ_i	63,3	63,5	67,2	68,8	69,9	71,3	71,4	72,8	75,7	76,5
η_i	113,3	115,4	153,6	167	125,6	150,8	166,2	181,2	159,9	173,1

Таблиця рангів тоді буде така:

ξ'_i	1	2	3	4	5	6	7	8	9	10
η'_i	1	2	5	8	3	4	7	10	6	9
$\xi'_i - \eta'_i$	0	0	-2	-4	2	2	0	-2	3	1

Тому, за формулою 7.1:

$$r_s \approx 0,745.$$

Звідси

$$T_8 = 0,745 \sqrt{\frac{10 - 2}{1 - (0,745)^2}} \approx 3,159.$$

Оскільки $t_{0,95;8} \approx 1,86$, то випадкові величини рангово кореляційно залежні.

7.7 Виконання в пакеті STATISTICA

Групі з 20 студентів було запропоновано відповісти на запитання: як часто (завжди, зазвичай, іноді, ніколи) ви відвідуєте спортивні змагання з різних видів спорту (футбол, бейсбол, баскетбол, бокс, гімнастика). Результати опитування наведено в таблиці.

№	Футбол	Бейсбол	Баскетбол	Бокс	Гімнастика
1	завжди	зазвичай	зазвичай	ніколи	завжди
2	завжди	зазвичай	зазвичай	іноді	зазвичай
3	завжди	зазвичай	ніколи	іноді	зазвичай
4	завжди	зазвичай	іноді	ніколи	зазвичай
5	ніколи	зазвичай	іноді	іноді	іноді
6	ніколи	іноді	іноді	іноді	іноді
7	завжди	зазвичай	іноді	іноді	іноді
8	завжди	іноді	ніколи	іноді	іноді
9	завжди	завжди	іноді	іноді	іноді
10	зазвичай	завжди	іноді	іноді	іноді

11	зазвичай	зазвичай	іноді	зазвичай	ніколи
12	зазвичай	зазвичай	іноді	ніколи	зазвичай
13	зазвичай	зазвичай	іноді	іноді	ніколи
14	зазвичай	зазвичай	іноді	ніколи	іноді
15	завжди	зазвичай	іноді	іноді	іноді
16	зазвичай	іноді	іноді	ніколи	іноді
17	зазвичай	ніколи	іноді	іноді	іноді
18	зазвичай	іноді	іноді	іноді	іноді
19	зазвичай	іноді	ніколи	іноді	іноді
20	зазвичай	іноді	іноді	завжди	іноді

Задачу розв'яжемо з допомогою процедури Correlations (*Spearman*, *Kendall tau*, *gamma*) пакету *Nonparametric Statistics*. Зайдемо в пакет та виберемо відповідну процедуру. У вікні, що відкриється, виберемо змінні, натиснувши *Variables* (рис. 7.12). Якщо змінних кілька, як у на-

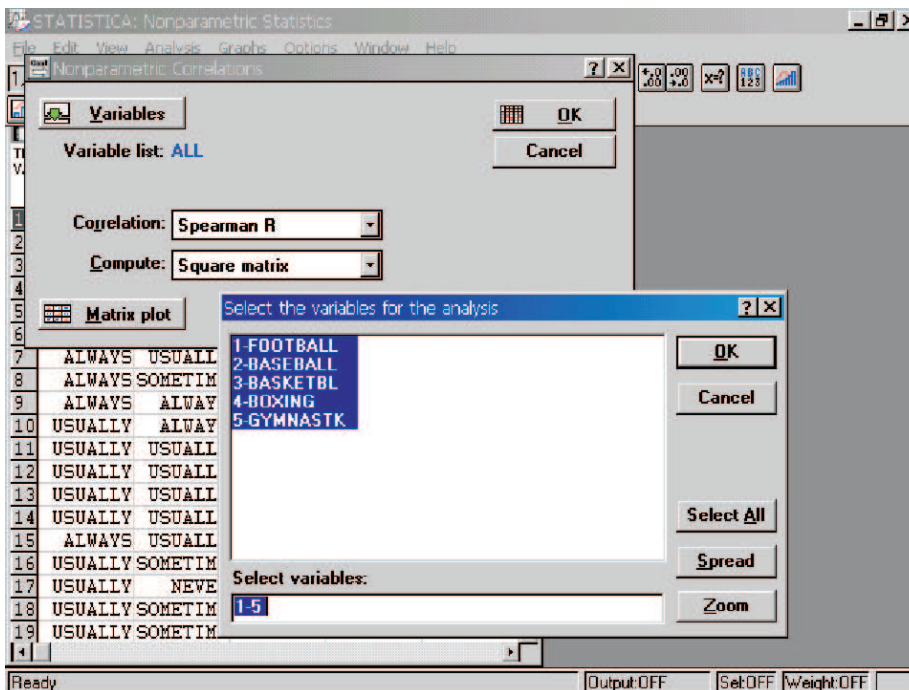


Рис. 7.12. Вибір змінних

шому випадку, і нас цікавлять кореляції між всіма парами, то зручно перед цим вибрати опцію *Compute: Square matrix*. Крім того, виберемо тип коефіцієнта кореляції, наприклад, *Correlation: Spearman R*. Натиснувши ОК, читаємо результат (рис. 7.13).

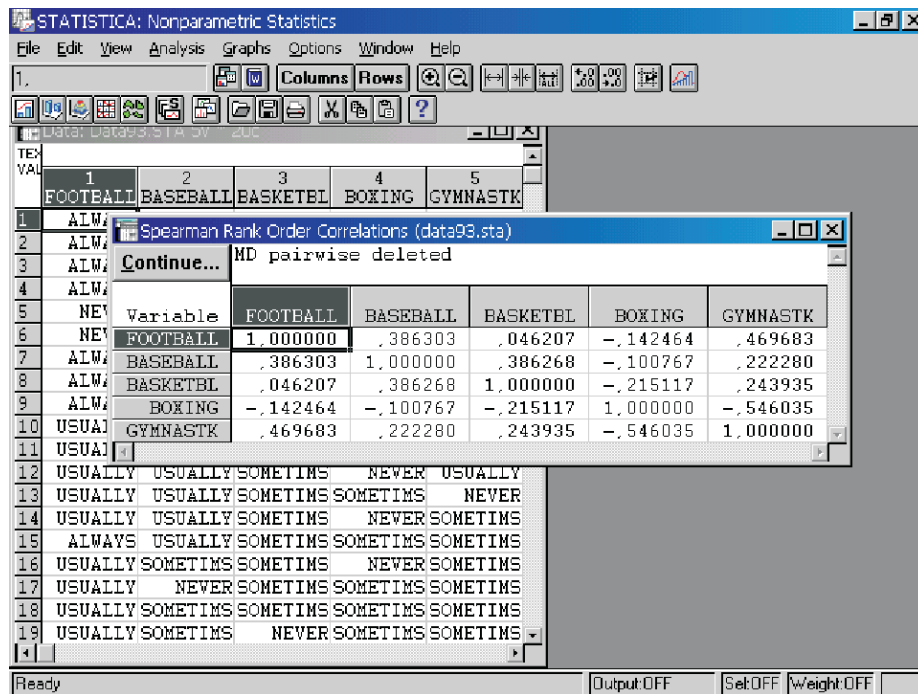


Рис. 7.13. Кореляція Спірмена

Розділ 8

Тести про вигляд розподілу

8.1 Двовибірковий критерій Колмогорова – Смірнова

Нехай x_1, x_2, \dots, x_n — вибірка з неперервно розподіленої генеральної сукупності з функцією розподілу $F(x)$, а y_1, y_2, \dots, 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_α — квантиль порядку α розподілу Колмогорова, то нульова гіпотеза не суперечить досліджуванім вибіркам (ймовірність помилки α).

8.2 Виконання в пакеті STATISTICA

Розв'яжемо задачу § 7.5 з допомогою критерію Колмогорова – Смірнова (*Kolmogorov – Smirnov two-sample test*). Виконуючи аналогічні дії, як для критерію Манна-Уїтні в § 7.5, одержимо такий результат (рис. 8.1):

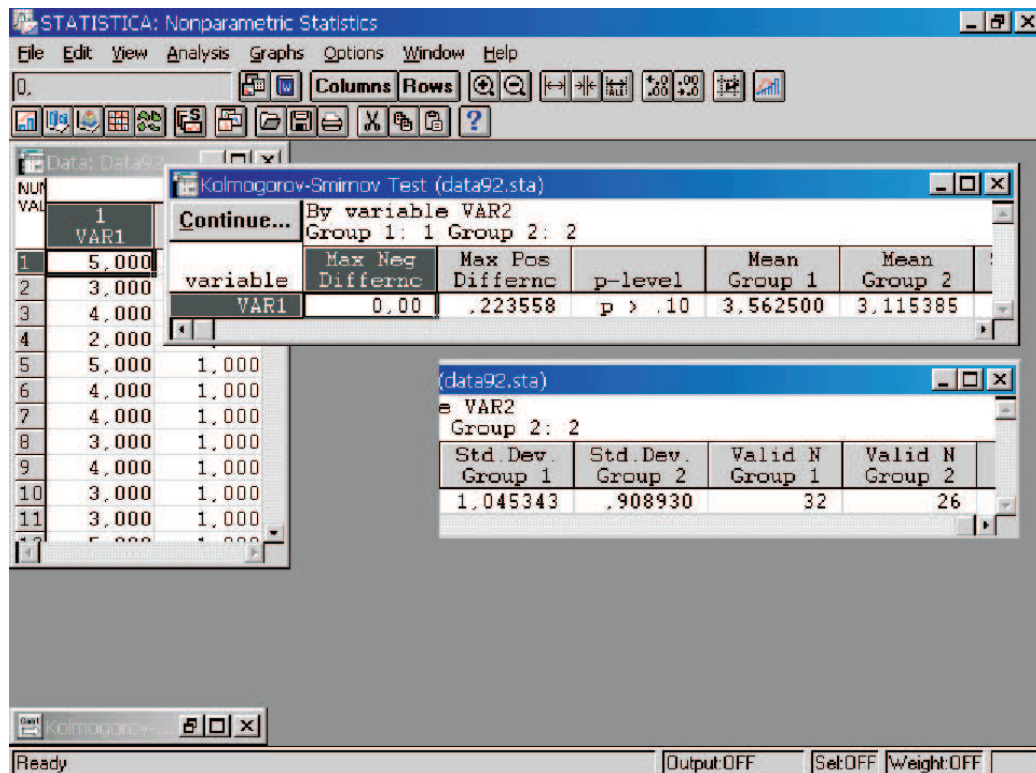


Рис. 8.1. Результати критерію Колмогорова – Смірнова

Підписи до кожної клітинки результуючої таблиці досить добре пояснюють її зміст. Як бачимо, критерій Колмогорова – Смірнова дає той же результат, що і тест Манна – Уїтні.

8.3 Критерій χ^2 та його застосування

8.3.1 Перевірка гіпотези про вид розподілу

Нехай у результаті експерименту отримали вибірку $\zeta = (\xi_1, \dots, \xi_n)$ 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 ґрунтується на тому, що емпіричний розподіл $\hat{\mathbf{F}}_n$, отриманий за вибіркою ζ , 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 Δ_i , $i = 1, 2, \dots, r$,

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 $p_i = \mathbf{P}\{\xi_k \in \Delta_i\}$ обчислюють за гіпотетичним розподілом \mathbf{G} , а ν_i – число елементів вибірки, які потрапили у множину Δ_i . Множини Δ_i вибирають так, щоб усі $p_i > 0$.

Якщо справедлива гіпотеза $H_0: \mathbf{F} = \mathbf{G}$, то частоти $\frac{\nu_i}{n}$ є слухними 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 $n \rightarrow \infty$ збігається до розподілу χ^2 з $r - 1$ ступенем вільності. Тому при досить великих n за розподіл $\hat{\chi}_n^2$ беруть розподіл χ^2 з $r - 1$ ступенем вільності.

Критерій χ^2 з рівнем значущості α полягає у тому, що гіпотезу H_0 відхиляють при

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

і приймають в іншому випадку.

8.3.2 Перевірка гіпотези про вид розподілу, який залежить від невідомих параметрів

Нехай $\zeta = (\xi_1, \dots, \xi_n)$ – вибірка із генеральної сукупності з невідомим розподілом \mathbf{F} . Гіпотеза H_0 полягає у тому, що $\mathbf{F} = \mathbf{G}(\theta)$, $\theta = (\theta_1, \dots, \theta_m)$, де розподіл \mathbf{G} визначається параметрами $\theta_1, \dots, \theta_m$, 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 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ параметрів і як гіпотетичний розглядають розподіл $\mathbf{G}(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$. Тоді розподіл відхилення

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

збігається до розподілу χ^2 з $(r - 1 - m)$ ступенем вільності.

У цьому випадку критерій χ^2 з рівнем значущості α полягає у відхиленні гіпотези H_0 при

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

ЗАУВАЖЕННЯ. Критерій χ^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

92 РОЗДІЛ 8

множин Δ_i має виконуватись умова $np_i > 10$. Якщо для деяких множин Δ_i ця умова не виконується, то їх потрібно об'єднати з сусідніми.

ПРИКЛАД. У таблиці вказані дані про кількість телефонних дзвінків до депутата з різних будинків району за 10000 годин. Перевірити гіпотезу про те, що кількість дзвінків підлягає розподілу Пуассона при рівні значущості $\alpha = 0,01$.

Кількість дзвінків, k	Кількість будинків, з яких було k дзвінків, n_k
0	427
1	235
2	72
3	21
4	1
5	1
≥ 6	0
Всього	757

Отже, при спостереженні за 757 будинками було $0 \cdot 427 + 1 \cdot 235 + 2 \cdot 72 + 3 \cdot 21 + 4 \cdot 1 + 5 \cdot 1 = 451$ дзвінків.

Оцінка $\hat{\lambda}$ параметра λ розподілу Пуассона дорівнює середній кількості дзвінків: $\hat{\lambda} = \frac{451}{757} \approx 0,6$. Тому маємо таблицю ймовірностей p_k та очікуваної кількості випадків з k дзвінками $n \cdot p_k$:

Кількість поломок, k	$p_k = \frac{0,6^k}{k!} e^{-0,6}$	np_k
0	0,54881	416
1	0,32929	249
2	0,09879	75
3	0,01976	15
4	0,00296	2
5	0,00036	0
≥ 6	0,00004	0

Оскільки для $k=4,5,6$ значення $np_k < 10$, об'єднуємо ці дані з даними для $k=3$:

k	n_k	np_k	$\frac{(n_k - np_k)^2}{np_k}$
0	427	416	0,291
1	235	249	0,787
2	72	75	0,120
≥ 3	23	17	2,118

Звідси, $\hat{\chi}_n^2 = 3,316$.

Оскільки ми оцінювали лише один параметр λ , то $m = 1$. Отже, кількість ступенів вільності $4 - 1 - 1 = 2$. За статистичними таблицями знаходимо: $\chi_{0,99;2}^2 = 9,21$. Отже, $\hat{\chi}_n^2 < \chi_{0,99;2}^2$, і гіпотезу про розподіл кількості дзвінків за законом Пуассона приймаємо.

8.3.3 Перевірка гіпотези про однорідність

Нехай у результаті k серій незалежних випробувань отримали результати $\xi_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{in_i})$, $i = 1, 2, \dots, k$. 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 $F_i(x)$ – функція розподілу спостережень i -ї серії, то ми повинні прийняти чи відхилити гіпотезу однорідності H_0 :

$$F_1(x) \equiv F_2(x) \equiv \dots \equiv F_k(x).$$

Для побудови критерію, як і в розділі 8.3.1, розіб'ємо область значень випадкових величин на скінченну кількість множин Δ_i , $i = 1, 2, \dots, r$, які не перетинаються. Нехай ν_{ij} – кількість результатів j -ої серії випробувань, які потрапили в множину Δ_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

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

де $\nu_i = \sum_{j=1}^k \nu_{ij}$.

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

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

8.4 Виконання в пакеті STATISTICA

Застосуємо *Observed versus expected XI* до прикладу з § 7.3. Будемо вважати, що рівні популярності співака після гастролей є спостережуваними частотами, а рівні до гастролей – прогнозованими. Такий вибір продиктований змістом нульової гіпотези. Вибравши змінні

у вікні *Observed vs. Expected Frequency* (*observed* — спостережувані, *expected* — прогнозовані) (рис. 8.2), одержимо результати (рис. 8.3), які свідчать, що відмінність між частотами є значущою ($p < 0,009204$).

8.5 Перевірка гіпотези про незалежність випадкових величин

Нехай ξ та η — дві дискретні випадкові величини, які можуть набувати значення x_1, x_2, \dots, x_k та y_1, y_2, \dots, y_l відповідно.

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

$$\mathbf{P}\{\xi = x_i, \eta = y_j\} = \mathbf{P}\{\xi = x_i\}\mathbf{P}\{\eta = y_j\} = p_i q_j, \quad 1 \leq i \leq k, \quad 1 \leq j \leq l.$$

Позначимо ν_{ij} — кількість спостережень ζ , 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	ν_{11}	ν_{12}	...	ν_{1l}	$\nu_{1\cdot}$
x_2	ν_{21}	ν_{22}	...	ν_{2l}	$\nu_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_k	ν_{k1}	ν_{k2}	...	ν_{kl}	$\nu_{k\cdot}$
Сума	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$...	$\nu_{\cdot l}$	n

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

Для перевірки гіпотези H_0 використовують критерій χ^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

$$\hat{\chi}_n^2 = n \sum_{i=1}^k \sum_{j=1}^l \frac{\nu_{ij}^2}{\nu_{\cdot j} \nu_{i\cdot}} - 1. \quad (8.1)$$

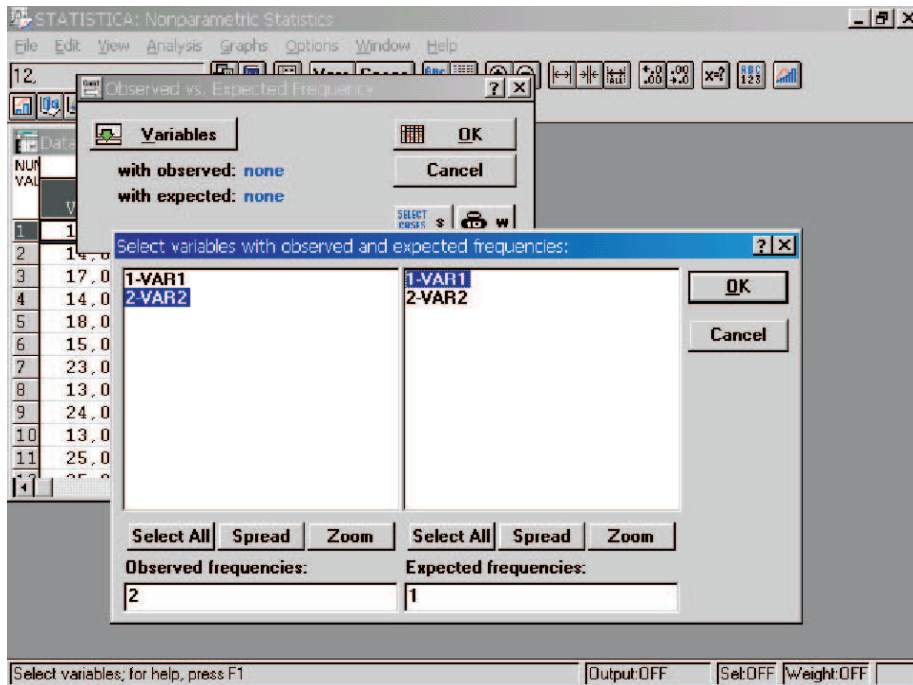


Рис. 8.2. Вибір змінних

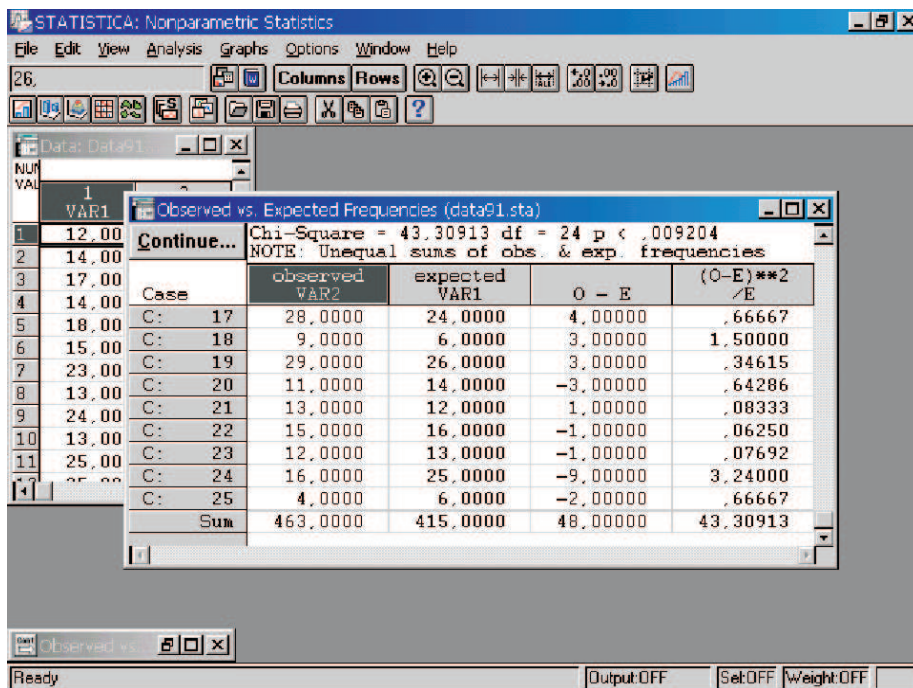


Рис. 8.3. Результати перевірки гіпотези про однорідність

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 kl частин. Тому $\hat{\chi}_n^2$ має граничний розподіл χ^2 з $lk-1-(k+l-2) = (k-1)(l-1)$ ступенями вільності.

Отже, гіпотезу H_0 відхиляють при рівні значущості α , якщо

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

ЗАУВАЖЕННЯ. Якщо ξ та η – неперервні випадкові величини, то область значення кожної з них розбивають на скінченну кількість проміжків, які відіграють роль x_i та y_j .

ПРИКЛАД. Вивчають три соціальні групи. Результати перевірки на наявність деякої ознаки такі:

Ознака	Група			Всього
	I	II	III	
Наявна	29	38	53	120
Відсутня	1	2	7	10
Всього	30	40	60	130

При рівні значущості $\alpha = 0,1$ перевірити гіпотезу про те, що наявність ознаки не залежить від групи.

Нам потрібно перевірити незалежність наявності ознаки та групи. За формулою (8.1):

$$\hat{\chi}_n^2 = 130 \left(\frac{29^2}{30 \cdot 120} + \frac{38^2}{40 \cdot 120} + \frac{53^2}{60 \cdot 120} + \frac{1^2}{30 \cdot 10} + \frac{2^2}{40 \cdot 10} + \frac{7^2}{60 \cdot 10} - 1 \right) \approx 2,546.$$

Число ступенів вільності: $(2-1)(3-1) = 2$. Оскільки $\hat{\chi}_{0,9;2}^2 \approx 4,605$, то робимо висновок, що наявність ознаки не залежить від групи.

Розділ 9

Лінійні регресійні моделі

Для того, щоб описувати та прогнозувати процеси, в економіці та соціології часто використовують математичні моделі цих процесів. У цьому випадку один або кілька параметрів процесу (ендогенних змінних) подають як функцію деяких зовнішніх факторів (екзогенних змінних). Деякі з цих факторів є суттєвими і чинять значний вплив на параметри процесу, а інші є несуттєвими, бо їх вплив є незначним. Як правило, суттєвих факторів є всього кілька, а несуттєвих — досить багато. Тому не можна повністю нехтувати впливом багатьох несуттєвих факторів на результуючі показники процесу. Позначимо результуючий показник процесу через y , набір суттєвих факторів (x_1, \dots, x_k) і набір несуттєвих факторів (ξ_1, \dots, ξ_k) . Загальний вигляд регресійної залежності:

$$y = F(x_1, \dots, x_k, \xi_1, \dots, \xi_k),$$

де F — функція регресії. Слід розрізняти крос-секційну регресію та регресію часових рядів. Крос-секційна регресія відображає зв'язок між ендогенними та екзогенними змінними в один і той же момент часу. Тому дані для всіх змінних вимірюють одночасно. А регресія часових рядів відображає зв'язок між змінними протягом певного проміжку часу, і тому дані для кожної змінної вимірюють періодично, у певні послідовні моменти часу, протягом деякого часового інтервалу.

9.1 Парна лінійна регресія

У тому разі, коли F є лінійною функцією, кажуть про лінійну регресію. Найпростішим випадком лінійної регресії є парна регресія: з однією ендогенною та однією суттєвою екзогенною змінною. Таку модель можна записати у наступному вигляді:

$$y = b_0 + b_1x + \xi,$$

де через ξ якраз і позначено вплив усіх несуттєвих чинників. Будемо вважати, що випадкова величина ξ підлягає $N(0, \sigma^2)$ розподілу. Нехай результати спостереження випадкових величин x та y подано у вигляді вибірки $(x_i, y_i), i = \overline{1, n}$. У такому разі можна вважати, що $y_i = b_0 + b_1 x_i + \varepsilon_i$, де ε_i — похибки спостережень, є некорельованими нормально розподіленими випадковими величинами з нульовим середнім та однаковою дисперсією (тобто підлягають розподілу $N(0, \sigma^2)$). Статистичні оцінки коефіцієнтів регресії \hat{b}_0 та \hat{b}_1 будують таким чином, щоб оцінки значень ендогенної змінної $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i, i = \overline{1, n}$ у точках $x_i, i = \overline{1, n}$ були якомога ближчими до вибірових значень $y_i, i = \overline{1, n}$.

За міру близькості вибирають, як правило, суму квадратів відхилень: $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, рідше — суму модулів відхилень: $\sum_{i=1}^n |y_i - \hat{y}_i|$. Якщо мірою близькості є сума квадратів відхилень, то кажуть, що відповідні оцінки коефіцієнтів регресії отримано методом найменших квадратів.

9.1.1 Метод найменших квадратів

Розглянемо міру близькості прогнозованих і спостережуваних значень $Q(\hat{b}_0, \hat{b}_1) = \sum_{j=1}^n (y_j - \hat{b}_0 - \hat{b}_1 x_j)^2$ як функцію двох змінних \hat{b}_0 і \hat{b}_1 та дослідимо її на екстремум. Запишемо необхідні умови існування екстремуму:

$$\begin{cases} \frac{\partial Q}{\partial b_0} = -2 \sum_{j=1}^n (y_j - \hat{b}_0 - \hat{b}_1 x_j) = 0, \\ \frac{\partial Q}{\partial b_1} = -2 \sum_{j=1}^n x_j (y_j - \hat{b}_0 - \hat{b}_1 x_j) = 0; \end{cases}$$

Розв'яжемо цю систему відносно оцінок коефіцієнтів регресії:

$$\begin{cases} n \cdot \hat{b}_0 + \hat{b}_1 \cdot \sum_{j=1}^n x_j = \sum_{j=1}^n y_j, \\ \hat{b}_0 \cdot \sum_{j=1}^n x_j + \hat{b}_1 \sum_{j=1}^n x_j^2 = \sum_{j=1}^n x_j y_j; \end{cases}$$

$$\begin{cases} \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}, \\ (\bar{y} - \hat{b}_1 \bar{x}) \cdot \sum_{j=1}^n x_j + \hat{b}_1 \sum_{j=1}^n x_j^2 = \sum_{j=1}^n x_j y_j; \end{cases}$$

$$\begin{cases} \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}, \\ \bar{y} \sum_{j=1}^n x_j - \hat{b}_1 \left(\bar{x} \cdot \sum_{j=1}^n x_j - \sum_{j=1}^n x_j^2 \right) = \sum_{j=1}^n x_j y_j; \end{cases}$$

$$\bar{y} \cdot \sum_{j=1}^n x_j - \sum_{j=1}^n x_j y_j = \hat{b}_1 \cdot \left(\bar{x} \cdot \sum_{j=1}^n x_j - \sum_{j=1}^n x_j^2 \right),$$

$$\sum_{j=1}^n x_j \cdot (\bar{y} - y_j) = \hat{b}_1 \cdot \left(\sum_{j=1}^n x_j (\bar{x} - x_j) \right).$$

Оскільки $\sum_{j=1}^n x_j \cdot (\bar{y} - y_j) = 0$, то

$$\sum_{j=1}^n x_j \cdot (\bar{y} - y_j) - \sum_{j=1}^n \bar{x} \cdot (\bar{y} - y_j) = \hat{b}_1 \cdot \left(\sum_{j=1}^n x_j \cdot (\bar{x} - x_j) - \bar{x} \cdot \sum_{j=1}^n (\bar{x} - x_j) \right),$$

звідки

$$\sum_{j=1}^n (x_j - \bar{x}) (\bar{y} - y_j) = \hat{b}_1 \cdot \sum_{j=1}^n (x_j - \bar{x}) (\bar{x} - x_j).$$

Остаточно отримуємо такі значення для оцінок коефіцієнтів парної лінійної регресії:

$$\begin{cases} \hat{b}_1 = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} = \hat{\rho} \cdot \frac{\hat{\sigma}_y}{\hat{\sigma}_x}, \\ \hat{b}_0 = \bar{y} - \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} \cdot \frac{1}{n} \sum_{j=1}^n x_j = \bar{y} - \hat{b}_1 \bar{x}. \end{cases}$$

Отже, емпірична лінія регресії має вигляд:

$$\hat{y} = \bar{y} - \hat{b}_1 \cdot \bar{x} + \hat{b}_1 \cdot x.$$

Або $\hat{y} = \bar{y} - \hat{b}_1 (x - \bar{x})$ і, отже, проходить через точку (\bar{x}, \bar{y}) .

Зауважимо, що отримані методом найменших квадратів оцінки коефіцієнтів парної лінійної регресії є найкращими у класі лінійних незміщених оцінок (*Best Linear Unbiased Estimator – BLUE*).

9.1.2 Перевірка моделі на адекватність

Якщо не розглядати питання правильності специфікації регресійної моделі з точки зору відповідної (економічної, соціологічної та іншої) теорії, то перевірка адекватності вибіркової регресії зводиться до перевірки її на відповідність раніше зробленим припущенням, а саме:

- похибки спостережень $\varepsilon_i = y_i - \hat{y}_i$ є нормально розподіленими випадковими величинами з нульовим середнім;
- дисперсія похибок є сталою (гомоскедастичність);
- похибки не є автокорельованими.

9.1.3 Перевірка моделі на значущість

Перевірку моделі на значущість здійснюють у два етапи. На першому етапі слід перевірити, чи включає довірчий інтервал для коефіцієнта b_1 нульове значення. Якщо при заданому рівні значущості довірчий інтервал накриває нульове значення, то знайдена регресійна залежність є незначущою. Якщо ж нульове значення не потрапляє у відповідний довірчий інтервал, то слід перейти до другого етапу, на якому встановлюється, яка саме частина варіації ендогенної змінної пояснюється варіацією екзогенної змінної.

Відношення суми квадратів відхилень, які пояснюються регресією, до загальної суми квадратів відхилень називають коефіцієнтом детермінації R^2 . Таким чином, значення коефіцієнту R^2 показує частку варіації ендогенної змінної, яку можна пояснити варіацією екзогенної змінної, а $1 - R^2$ частку варіації ендогенної змінної, викликану впливом випадкових факторів.

9.2 Множинна лінійна регресія

На жаль, парна лінійна регресія має досить обмежену сферу застосування на практиці, оскільки зазвичай доводиться досліджувати зв'язок між багатьма екзогенними змінними та однією ендогенною. Таку регресію називають множинною, або багатофакторною.

Загалом, множинна регресія дозволяє дослідникові визначити, "що є кращим поясненням для...". Наприклад, дослідникові в галузі освіти було б цікаво, які фактори є кращим поясненням успішного навчання в середній школі, психолога могло б зацікавити питання, які індивідуальні якості дозволяють краще прогнозувати ступінь соціальної адаптації індивіда, соціологам, імовірно, було б цікаво знайти ті соціальні індикатори, які найкраще пояснюють результат адаптації нової іммігрантської групи і ступінь її злиття із суспільством.

Модель множинної регресії можна записати у вигляді:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n + \xi,$$

де n – число факторів (екзогенних змінних) множинної регресії.

Застосувавши метод найменших квадратів, можна обчислити оцінки коефіцієнтів множинної регресії аналогічно до того, як це зроблено для парної регресії. Нагадаємо, що для парної регресії важливими були припущення про те, що:

- похибки спостережень є нормально розподіленими випадковими величинами з нульовим середнім;
- дисперсія похибок є сталою (гомоскедастичність);
- похибки не є автокорельованими.

Для множинної регресії до цих умов додається ще умова незалежності (некорельованості) екзогенних змінних. Якщо ж деякі з екзогенних змінних корелюють між собою, то говорять про явище мультиколінеарності.

Мультиколінеарність не дозволяє правильно визначити оцінки коефіцієнтів у регресійній моделі при корелюючих змінних, і тому варто переглянути питання специфікації моделі, виходячи з міркувань відповідної (економічної, соціологічної та іншої) теорії.

9.3 Виконання в пакеті STATISTICA

Відкриємо файл *Job_prof.sta*, який знаходиться у директорії *Examples* пакету (рис. 9.1). Перші чотири змінні (*Test1-Test4*) показують результати тестів на професійну придатність 25 претендентів на посаду службовців компанії. Незважаючи на результати тестування, всі 25 претендентів були взяті з випробувальним терміном, після закінчення якого робота кожного із службовців була оцінена за єдиною шкалою. Результати оцінок описує змінна *Job_prof*. Слід визначити, який (або які) з проведених тестів найбільш адекватно відображає рівень професійної придатності претендента.

Відкриємо вікно *Multiple Regression Startup Panel* (рис. 9.2), натиснемо на кнопку *Variables* та виберемо *Job_prof* як ендogenous змінну та *Test1-Test4* як екзогенні.

У клітинці *Perform default analysis* приберемо відповідне позначення (хрестик). Цим самим ми виберемо покроковий метод проведення регресійного аналізу.

Ми можемо вибрати наступні методи проведення регресійного аналізу: *Standard*, *Forward stepwise* та *Backward stepwise*.

STATISTICA: Multiple Regression

File Edit View Analysis Graphs Options Window Help

86. [Icons]

Data: JOB_PROF.STA 5v * 25c

NUMER Job proficiency data set from Neter, Wasserman, & Kutner, 1989

VALUE	1 TEST1	2 TEST2	3 TEST3	4 TEST4	5 JOB_PROF
1	86	110	100	87	88
2	62	97	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116
9	112	119	106	105	104
10	120	89	105	97	99
11	87	81	90	88	64
12	133	120	113	108	126
13	140	121	96	89	94
14	84	113	98	78	71
15	106	102	109	109	111
16	109	129	102	108	109
17	104	83	100	102	100
18	150	118	107	110	127
19	98	125	108	95	99
20	120	94	95	90	82
21	74	121	91	85	67
22	96	114	114	103	109
23	104	73	93	80	78
24	94	121	115	104	115
25	91	129	97	83	83

Рис. 9.1. Результати тестів на професійну придатність

Multiple Regression

Variables:

Independent: TEST1-TEST4
Dependent: JOB_PROF

Input file: Raw Data

MD deletion: Casewise

Mode: Standard

Perform default (non-stepwise) analysis
 Review descr. stats, corr. matrix
 Extended precision computations
 Batch processing/printing
 Print residual analysis

Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. deselect the default analysis check box.

OK
Cancel
Open Data
SELECT CRGES S W
 Weighted moments
DF = W-1 N-1

Рис. 9.2. Множинна регресія

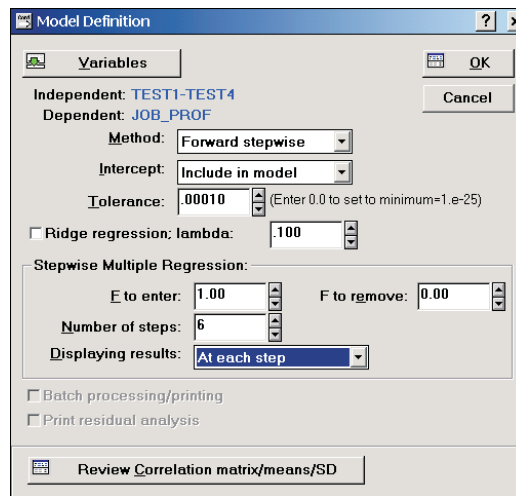


Рис. 9.3. Вибір методу

У вікні *Model definition* (рис. 9.3) оберемо метод *Forward stepwise* та виберемо опцію відображення процесу покроково у віконці *Displaying results*. Після вибору інших параметрів регресійного аналізу натиснемо *OK*.

Відкриється вікно *Multiple Regression Results* (рис. 9.4) для нульового кроку (Step 0). На цьому кроці можна переглянути описову статистику екзогенних змінних, натиснувши на кнопку *Correlations and desc. stats*.

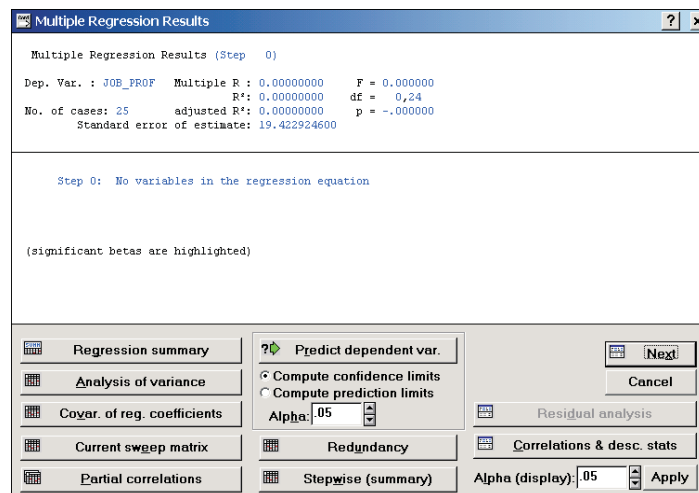


Рис. 9.4. Нульовий крок

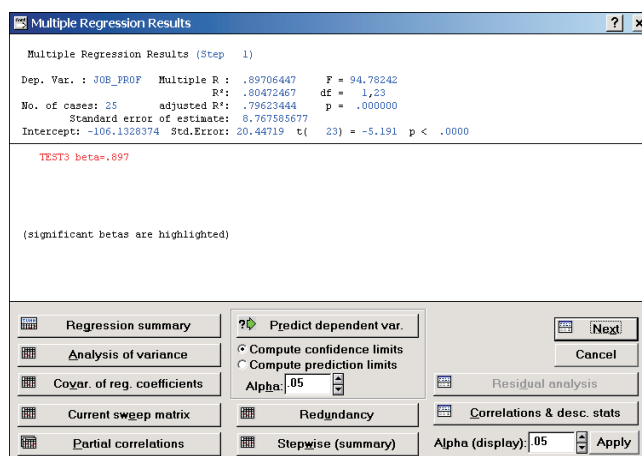


Рис. 9.5. Перший крок

На першому кроці (Step 1) (рис. 9.5) кожен з екзогенних змінних оцінюють окремо і ту з них, що має найбільший вплив (найбільше значення F-статистики), включають у регресійну модель.

На кожному наступному кроці (рис. 9.6 і 9.7) в модель включають наступну, найбільш впливову, згідно з F-критерієм, екзогенну змінну.

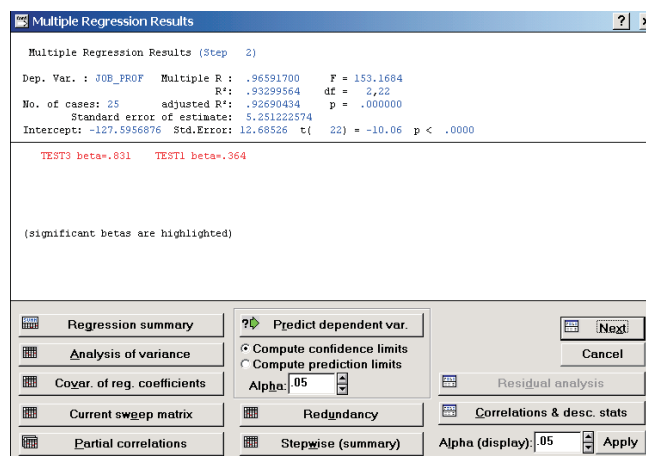


Рис. 9.6. Другий крок

Коли всі значущі екзогенні змінні увійдуть у модель, програма переходить до вікна *Residual Analysis* (рис. 9.8), за допомогою якого можна провести аналіз залишків регресійної моделі.

Наприклад, натиснувши кнопку *Plots of residuals*, можна частково перевірити модель на відповідність припущенню про нормальний розпо-

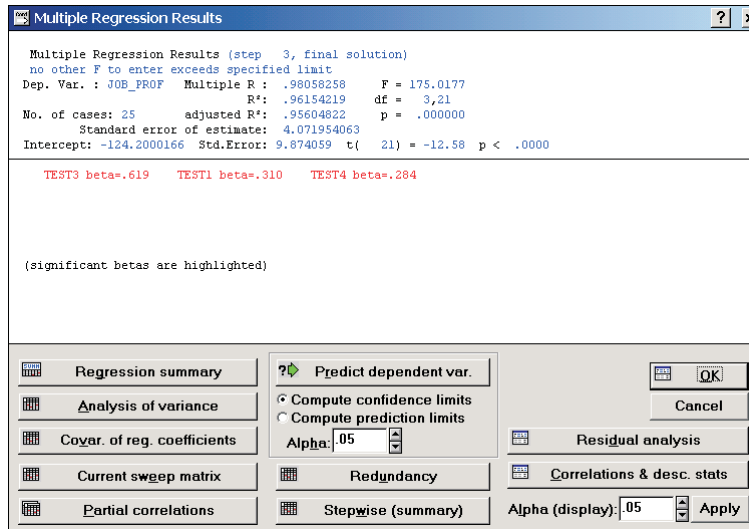


Рис. 9.7. Результати наступних кроків

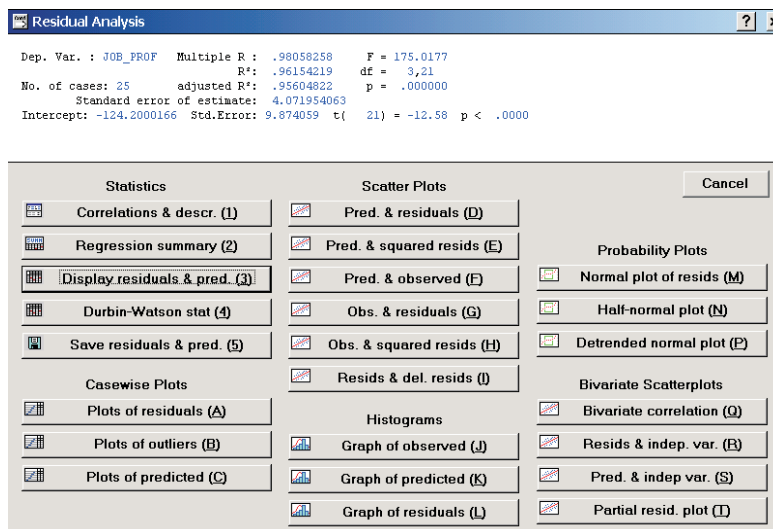


Рис. 9.8. Вікно аналізу залишків

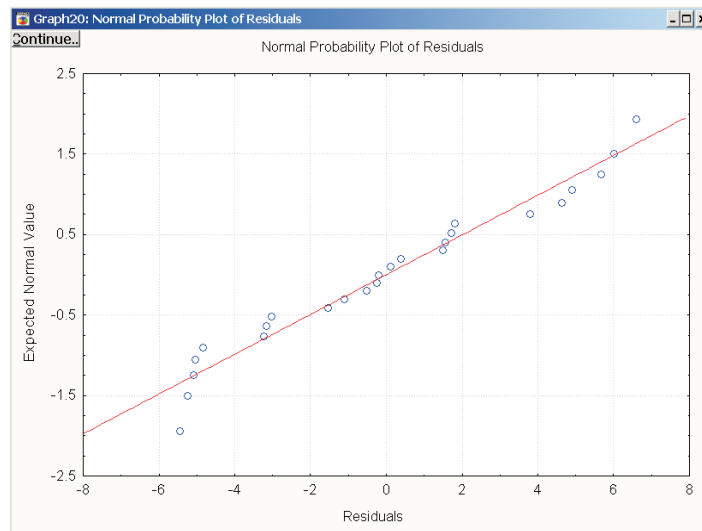


Рис. 9.9. Залишки на нормальному Q-Q графіку

діл залишків (рис. 9.9).

Також на панелі вікна *Residual Analysis* розміщена кнопка *Regression Summary*, натиснувши яку, можна отримати значення коефіцієнтів регресії, t -статистики та відповідної ймовірності помилки першого роду p , а також коефіцієнт детермінації R^2 .

Зауважимо, що у вікні *Regression Summary for Dependent Variable* (рис. 9.10) коефіцієнти екзогенних змінних, що увійшли в модель, зображені у двох варіантах.

Regression Summary for Dependent Variable: JOB_PROF						
Continue..						
R= .98058258 RI= .96154219 Adjusted RI= .95604822						
F(3,21)=175.02 p<.00000 Std.Error of estimate: 4.0720						
N=25	BETA	St. Err. of BETA	B	St. Err. of B	t (21)	p-level
Intercept			-124.200	9.874059	-12.5784	.000000
TEST3	.618670	.069224	1.357	.151832	8.9373	.000000
TEST1	.309670	.045646	.296	.043679	6.7841	.000001
TEST4	.284405	.072035	.517	.131054	3.9482	.000735

Рис. 9.10. Таблиця результатів аналізу

По-перше, це B коефіцієнти, які відповідають звичайним b регресійної моделі. По-друге, це $BETA$ коефіцієнти, які відповідають попередньо центрованим та стандартизованим значенням екзогенних змінних. Відповідно, величини коефіцієнтів $BETA$ дозволяють визначати відносний внесок відповідної екзогенної змінної у прогнозоване значення ендогенної змінної.

Розділ 10

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. Здебільшого таке розрізнення відбувається на інтуїтивному рівні. Проте коли ми маємо справу з новими явищами чи намагаємося згрупувати вже відомі об'єкти в нові класи за якимись ознаками, то в цьому разі не можна повністю покладатись на інтуїцію. 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 потрапляють.

У реальних ситуаціях постановка задачі не завжди така проста. Часто деякі спостереження втрачені, деякі характеристики ми не можемо виміряти. Іноді для різних об'єктів маємо характеристики різних типів і тощо.

Далі будемо розглядати різні математичні аспекти й моделі в кластерному аналізі. Які алгоритми кластеризації застосовувати, значною мірою залежить від реальної задачі. Питання це не є суто математичним і потребує глибокого розуміння не лише теорії, але й природи реального явища, яке розглядають.

Останніми роками кластерний аналіз дедалі частіше застосовують у соціологічних науках. Точне визначення соціальних, вікових груп дає змогу ефективніше працювати з ними, цілеспрямованіше застосовувати засоби впливу, робити точніші прогнози.

Перше питання, яке постає при кластеризації деякої сукупності об'єктів, – як вимірювати їх подібність чи відмінність між собою. Тобто, для об'єктів i та j та пари векторів їх характеристик $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 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.

10.1 Коефіцієнти подібності бінарних змінних

Нехай кожна зі змінних x_{ki} , яка використовується для опису характеристик індивіда, може набувати лише два значення (наприклад, високий чи низький, палить чи ні, соціально активний чи ні тощо). Тоді для двох індивідів ми можемо утворити 2×2 таблицю, у кожній клітинці якої стоїть число, що показує, скільки відповідних пар значень існує:

Значення характеристик	Індивід 1		Усього
	1	2	
Індивід 2 1	a	b	$a + b$
2	c	d	$c + d$
Усього	$a + c$	$b + d$	$a + b + c + d$

Ось список коефіцієнтів подібності, які часто використовують на практиці для такої ситуації:

$$1. \frac{a+d}{a+b+c+d} \quad 2. \frac{a}{a+b+c} \quad 3. \frac{2a}{2a+b+c} \quad 4. \frac{2(a+d)}{2(a+d)+b+c} \quad 5. \frac{a}{a+2(b+c)}$$

Як уже зазначалося, до яких даних застосовувати той чи інший коефіцієнт, визначається природою реальної ситуації. Для однакових даних різні коефіцієнти можуть мати дуже відмінні значення. Наприклад, нехай ми маємо двох індивідів, які описуються десятьма показниками:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \quad (10.1)$$

Відповідна 2×2 таблиця має вигляд:

Значення характеристик	Індивід 1		Усього
	1	0	
Індивід 2 1	2	1	3
0	2	5	7
Усього	4	6	10

Значення розглянутих раніше коефіцієнтів будуть для цих даних такі:

1. 0,7 2. 0,4 3. 0,57 4. 0,82 5. 0,25

Насправді різниця між коефіцієнтами ще глибша. Можна було б очікувати, що всі коефіцієнти подібності сумісно монотонні, тобто якщо значення якогось одного коефіцієнта для всіх пар індивідів мають певний порядок, то й значення інших коефіцієнтів для цих самих пар упорядковані таким самим чином.

Додамо до нашої матриці (10.1) новий рядок характеристик ще одного індивіда:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \quad (10.2)$$

У цьому випадку перший та другий із розглянутих коефіцієнтів набувають таких значень:

$$\begin{array}{ll} 1) & s_{12} = 0,70 \\ & s_{13} = 0,50 \\ & s_{23} = 0,80 \end{array} \quad \begin{array}{ll} 2) & \tilde{s}_{12} = 0,40 \\ & \tilde{s}_{12} = 0,00 \\ & \tilde{s}_{23} = 0,33 \end{array}$$

Отже, коефіцієнти не є сумісно монотонні.

Ситуація, коли деякий показник індивіда не є бінарним, але може набувати скінченної кількості значень, зводиться різними методами до попередньої. Скажімо, можна ввести нові бінарні характеристики, кожна з яких показує, чи набуває початкова характеристика якогось певного значення.

10.2 Подібність змінних із неперервними значеннями

Змінні, які набувають значень із деякої неперервної області, можна перетворити на бінарні, порівнюючи їх із деяким фіксованим значенням.

Наприклад, змінні, які показують термін перебування на одній посаді, проживання в певному регіоні, вживання наркотиків, можуть бути трансформовані в бінарну змінну з двома значеннями: більше двох років, менше чи точно два роки. Далі можемо застосувати розглянуті раніше коефіцієнти подібності. Зрозуміло, що за такого підходу значна частина інформації втрачається. На практиці застосовують різні інші коефіцієнти, зокрема пов'язані з кореляцією даних (див. Clifford, Stephenson [23]). Розглянемо ще один коефіцієнт (запропонований Gower [24]), який демонструє, як можна порівнювати індивідів за показниками різних типів:

$$s_{ij} = \frac{\sum_{k=1}^p w_{ijk} S_{ijk}}{\sum_{k=1}^p w_{ijk}}$$

Змінна w_{ijk} набуває значень 1 чи 0 залежно від того, враховуємо чи ні порівняння k -ї характеристики i -го та j -го індивідів. S_{ijk} – подібність між i -м та j -м індивідами за їх k -ю характеристикою. Скажімо, можемо розглянути такий випадок: для характеристик, які набувають дискретних значень, S_{ijk} дорівнює 1, якщо k -та характеристика однакова для i -го та j -го індивідів, і 0, якщо різна. Для характеристик з неперервною областю значень можемо розглянути

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

де x_{ik} та x_{jk} – значення k -ї характеристики i -го та j -го індивідів відповідно, а R_k – довжина проміжка, в якому змінюється k -та характеристика індивідів, що кластеризуються.

ПРИКЛАД. Нехай таблиця дає значення п'яти характеристик для п'яти осіб:

	1	2	3	4	5
1-й індивід	60	середній	0	ні	юний
2-й індивід	75	високий	1	ні	середнього віку
3-й індивід	55	низький	1	так	старий
4-й індивід	72,5	середній	0	так	старий
5-й індивід	60	середній	0	так	юний

Ми хочемо знайти коефіцієнти подібності між індивідами, не враховуючи пари нульових третіх характеристик і негативних четвертих.

Розглянемо першу характеристику як таку, що має неперервну область значень, решту – з дискретними. Оскільки $R_1 = 75 - 55 = 20$, то

$$s_{12} = \frac{1 \cdot (1 - \frac{15}{20}) + 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0 + 1 \cdot 0}{1 + 1 + 1 + 0 + 1} = 0,0625.$$

Обчисливши аналогічним чином решту коефіцієнтів, отримуємо симетричну матрицю подібності S для наших п'яти індивідів:

$$S = \begin{pmatrix} 1 & 0,062 & 0,15 & 0,344 & 0,75 \\ 0,062 & 1 & 0,2 & 0,175 & 0,005 \\ 0,15 & 0,2 & 1 & 0,425 & 0,350 \\ 0,344 & 0,175 & 0,425 & 1 & 0,475 \\ 0,75 & 0,005 & 0,35 & 0,475 & 1 \end{pmatrix}$$

10.3 Коефіцієнти відмінності

У багатьох випадках зручно вимірювати не подібність індивідів, а відмінність у їх характеристиках. Проте значної різниці тут немає. Часто для відповідних протилежних коефіцієнтів існує проста формула, яка дає їх зв'язок.

Ось декілька коефіцієнтів відмінності, які зустрічаються найчастіше:

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

Можливо, один із найчастіше вживаних коефіцієнтів – це перший. З математики відомо, що це квадрат евклідової відстані між двома точками. Проте для реальних даних цей коефіцієнт не завжди дає гарні результати. Наприклад, нехай у нас є дані про трьох дітей, вага яких визначена у фунтах, а зріст в футах:

	Вага	Зріст
Перша дитина	60	3,0
Друга дитина	65	3,5
Третя дитина	63	4,0

Відповідні евклідові відстані тоді дорівнюють:

$$d_{12} = 5,02; \quad d_{13} = 3,16; \quad d_{23} = 2,06.$$

Але якщо змінити шкалу вимірювання зросту, скажімо, на дюйми, то евклідові відстані будуть уже такі:

$$d_{12} = 7,81; \quad d_{13} = 12,37; \quad d_{23} = 6,32.$$

За першим методом шкалювання третя дитина ближча, ніж друга, до першої за своїми характеристиками, а при другому методі шкалювання – навпаки.

Евклідову відстань (перший із коефіцієнтів відмінності) звичайно застосовують у тому разі, коли вважають, що різні характеристики індивіда некорельовані між собою. За наявності кореляції, яка визначається коваріаційною матрицею S , віддалі між векторами x_i та x_j визначають формулою

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

Для багатьох бінарних показників із коефіцієнтами подібності $s_{ij} \in [-1, 1]$ часто застосовують таку формулу для визначення відмінностей:

$$d_{ij} = \sqrt{2(1 - s_{ij})}.$$

10.4 Міжгрупові відстані

До цього моменту ми порівнювали між собою два об'єкти: знаходили їх подібність, відмінність, вводили 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 у кластерному аналізі, й про них ми будемо говорити докладніше 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 $\bar{x}_A = (\bar{x}_{A,1}; \bar{x}_{A,2}; \dots; \bar{x}_{A,p})$, $\bar{x}_B = (\bar{x}_{B,1}; \bar{x}_{B,2}; \dots; \bar{x}_{B,p})$. 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 і S – коваріаційна матриця міжгрупових середніх, то тоді часто застосовують таку характеристику відстані:

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

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 .

Розділ 11

Ієрархічна кластерна техніка

Ієрархічна кластерна техніка полягає в тому, що будують деяку послідовність кластерних розбиттів множини, таку, що, з одного боку, розбиття складається з кластерів, кожен із яких містить тільки один елемент множини, а з іншого боку, маємо лише один кластер, що містить усю множину. Відповідно до напрямку, 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.

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. 11.1.

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

11.1 Агломеративні методи

Нехай ми маємо множину з n об'єктів. Метод полягає в утворенні послідовності P_n, P_{n-1}, \dots, 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, \dots, 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 кластери C_i та C_j знищуємо.

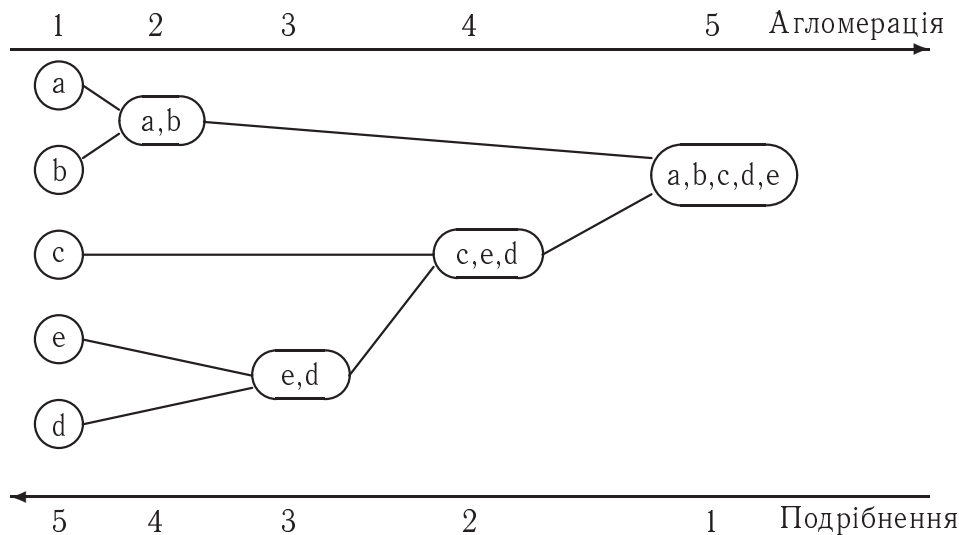


Рис. 11.1. Дендрограма

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 відбуватися по-різному.

11.1.1 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. 11.2).

ПРИКЛАД. Нехай маємо 5 об'єктів, відстані між якими задані такою матрицею:

$$\begin{matrix}
 & \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \\
 \{1\} & \left(\begin{array}{ccccc}
 0 & 2 & 6 & 10 & 9 \\
 2 & 0 & 5 & 9 & 8 \\
 6 & 5 & 0 & 4 & 9 \\
 10 & 9 & 4 & 0 & 3 \\
 9 & 8 & 9 & 3 & 0
 \end{array} \right) & & & & \\
 \{2\} & & & & & \\
 \{3\} & & & & & \\
 \{4\} & & & & & \\
 \{5\} & & & & &
 \end{matrix} \quad (11.1)$$

На першому кроці маємо розбиття

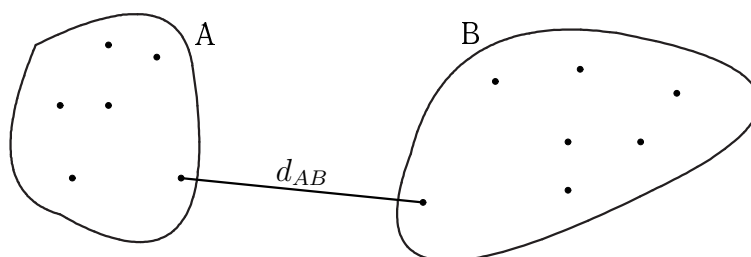


Рис. 11.2. Найближчі сусіди

$$P_5 : \{1\}, \{2\}, \{3\}, \{4\}, \{5\}.$$

Найменша відстань – між 1-м та 2-м об'єктами. Тому об'єднуємо їх у новий кластер і дістаємо розбиття

$$P_4 : \{1, 2\}, \{3\}, \{4\}, \{5\}.$$

Обчислимо відстань між кластером $\{1, 2\}$ та рештою:

$$d_{\{1,2\}\{3\}} = \min\{d_{13}, d_{23}\} = 5;$$

$$d_{\{1,2\}\{4\}} = \min\{d_{14}, d_{24}\} = 9;$$

$$d_{\{1,2\}\{5\}} = \min\{d_{15}, d_{25}\} = 8.$$

Нова матриця відстаней між кластерами має вигляд:

$$\begin{array}{c} \{1, 2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array} \begin{pmatrix} \{1, 2\} & \{3\} & \{4\} & \{5\} \\ 0 & 5 & 9 & 8 \\ 5 & 0 & 4 & 9 \\ 9 & 4 & 0 & 3 \\ 8 & 9 & 3 & 0 \end{pmatrix}$$

Найменша відстань між кластерами $\{4\}$ і $\{5\}$. Утворюємо новий кластер $\{4, 5\}$ і дістаємо розбиття

$$P_3 : \{1, 2\}, \{3\}, \{4, 5\}.$$

Оскільки

$$d_{\{1,2\}\{4,5\}} = \min\{d_{14}, d_{15}, d_{24}, d_{25}\} = 8;$$

$$d_{\{3\}\{4,5\}} = \min\{d_{34}, d_{35}\} = 4; \quad d_{\{1,2\}\{3\}} = 5,$$

то нова матриця відстаней

$$\begin{matrix} & \{1, 2\} & \{3\} & \{4, 5\} \\ \{1, 2\} & \left(\begin{matrix} 0 & 5 & 8 \\ 5 & 0 & 4 \\ 8 & 4 & 0 \end{matrix} \right) \\ \{3\} & & & \\ \{4, 5\} & & & \end{matrix}$$

Найменша відстань – між кластерами $\{4, 5\}$ і $\{3\}$. Тому наступне розбиття

$$P_2 : \{1, 2\}, \{3, 4, 5\}.$$

І на останньому кроці отримуємо розбиття

$$P_1 : \{1, 2, 3, 4, 5\}.$$

Відповідна дендрограма зображена на рис. 11.3.

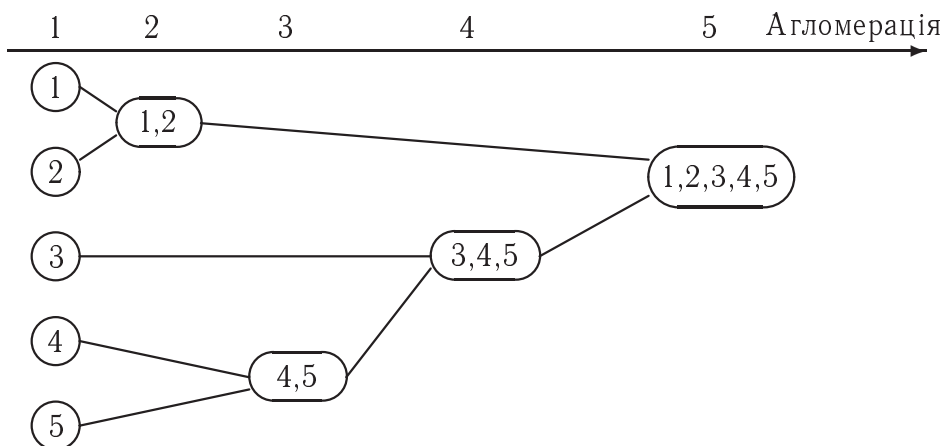


Рис. 11.3. Дендрограма 1

11.1.2 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. 11.4).

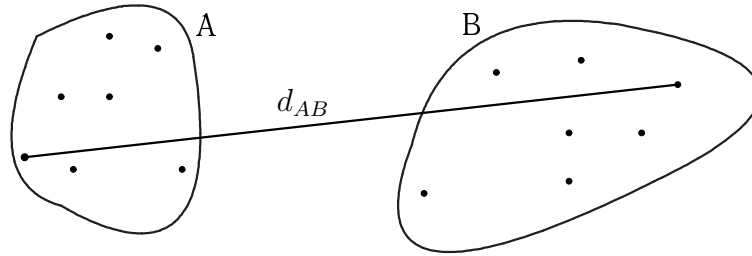


Рис. 11.4. Найвіддаленіші сусіди

Якщо застосовувати цей метод до матриці (11.1), то на першому кроці знову отримуємо новий кластер $\{1, 2\}$ і

$$P_2 : \{1, 2\}, \{3\}, \{4\}, \{5\}.$$

Оскільки

$$d_{\{1,2\}\{3\}} = \max\{d_{13}, d_{23}\} = 6;$$

$$d_{\{1,2\}\{4\}} = \max\{d_{14}, d_{24}\} = 10;$$

$$d_{\{1,2\}\{5\}} = \max\{d_{15}, d_{25}\} = 9,$$

то на наступному кроці маємо таке розбиття:

$$P_3 : \{1, 2\}, \{3\}, \{4, 5\}$$

і матрицю відстаней

$$\begin{array}{c} \{1, 2\} \\ \{3\} \\ \{4, 5\} \end{array} \begin{pmatrix} \{1, 2\} & \{3\} & \{4, 5\} \\ 0 & 6 & 10 \\ 6 & 0 & 9 \\ 10 & 9 & 0 \end{pmatrix}$$

Тому наступними розбиттями будуть:

$$P_4 : \{1, 2, 3\}, \{4, 5\};$$

$$P_5 : \{1, 2, 3, 4, 5\}.$$

Дендрограма буде вже така:

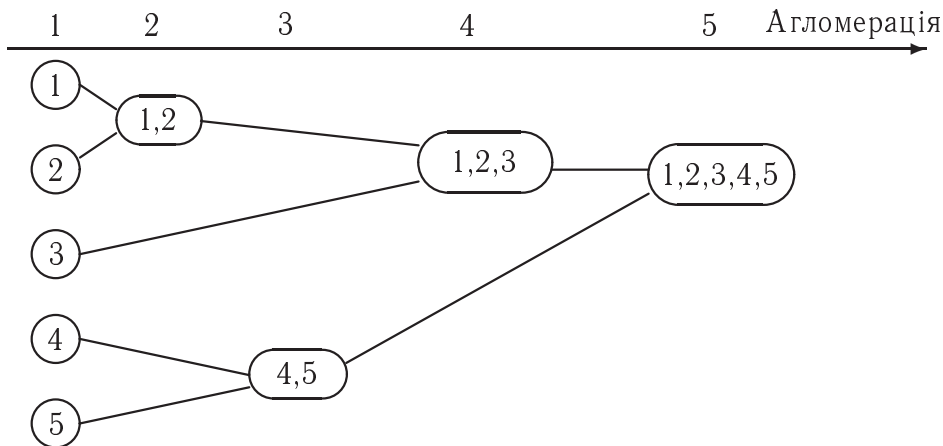


Рис. 11.5. Дендрограма 2

11.1.3 Метод середніх групових відстаней

У цьому методі відстань між двома кластерами визначають як середнє арифметичне відстаней усіх пар індивідів, по одному з кожного кластера. Розглянемо метод на прикладі матриці відстаней (11.1). На першому кроці без змін отримуємо розбиття:

$$P_2 : \{1, 2\}, \{3\}, \{4\}, \{5\}.$$

Оскільки

$$d_{\{1,2\}\{3\}} = \frac{1}{2}(d_{13} + d_{23}) = 5, 5;$$

$$d_{\{1,2\}\{4\}} = \frac{1}{2}(d_{14} + d_{24}) = 9, 5;$$

$$d_{\{1,2\}\{5\}} = \frac{1}{2}(d_{15} + d_{25}) = 8, 5,$$

то нова матриця відстаней буде

$$\begin{array}{c} \{1, 2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array} \begin{pmatrix} \{1, 2\} & \{3\} & \{4\} & \{5\} \\ 0 & 5, 5 & 9, 5 & 8, 5 \\ 5, 5 & 0 & 4 & 9 \\ 9, 5 & 4 & 0 & 3 \\ 8, 5 & 9 & 3 & 0 \end{pmatrix}$$

Найменша відстань – між $\{4\}$ та $\{5\}$. Отже, нове розбиття:

$$P_3 : \{1, 2\}, \{3\}, \{4, 5\}.$$

Оскільки

$$d_{\{1,2\}\{4,5\}} = \frac{1}{4}(d_{14} + d_{15} + d_{24} + d_{25}) = 9;$$

$$d_{\{1,2\}\{3\}} = 5, 5;$$

$$d_{\{4,5\}\{3\}} = \frac{1}{2}(d_{43} + d_{53}) = 6, 5,$$

то нова матриця відстаней:

$$\begin{array}{c} \{1, 2\} \\ \{3\} \\ \{4, 5\} \end{array} \begin{pmatrix} \{1, 2\} & \{3\} & \{4, 5\} \\ 0 & 5, 5 & 9 \\ 5, 5 & 0 & 6, 5 \\ 9 & 6, 5 & 0 \end{pmatrix}$$

і наступні розбиття такі:

$$P_4 : \{1, 2, 3\}, \{4, 5\};$$

$$P_5 : \{1, 2, 3, 4, 5\}.$$

Дендрограма залишається такою самою, як і на рис. 11.5.

11.1.4 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.

Продемонструємо застосування цього методу на прикладі. Нехай ми маємо 5 осіб, кожна з яких характеризується 2-ма показниками:

$$\begin{matrix} \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{matrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix}$$

Якщо взяти за відстань між індивідами евклідову відстань, то отримаємо таку матрицю відстаней:

$$\begin{matrix} & \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \\ \{1\} & \begin{pmatrix} 0 & 1 & 5,39 & 7,07 & 7,07 \\ 1 & 0 & 5,10 & 7,00 & 7,28 \\ 5,39 & 5,10 & 0 & 2,24 & 3,61 \\ 7,07 & 7,00 & 2,24 & 0 & 2 \\ 7,07 & 7,28 & 3,61 & 2 & 0 \end{pmatrix} \end{matrix}$$

Оскільки найменша відстань – між 1-ю та 2-ю особами, то утворюємо новий кластер $\{1, 2\}$. Отже,

$$P_4 : \{1, 2\}, \{3\}, \{4\}, \{5\}.$$

Для групи $\{1, 2\}$ середні характеристики це $(1; 1, 5)$. Нова матриця відстаней буде тоді:

$$\begin{matrix} & \{1, 2\} & \{3\} & \{4\} & \{5\} \\ \{1, 2\} & \begin{pmatrix} 0 & 5,22 & 7,02 & 7,16 \\ 5,22 & 0 & 2,24 & 3,61 \\ 7,02 & 2,24 & 0 & 2 \\ 7,16 & 3,61 & 2 & 0 \end{pmatrix} \end{matrix}$$

Оскільки найменша відстань – між $\{4\}$ і $\{5\}$, то маємо новий кластер $\{4, 5\}$ і нову матрицю відстаней:

$$\begin{matrix} & \{1, 2\} & \{3\} & \{4, 5\} \\ \{1, 2\} & \begin{pmatrix} 0 & 5 & 7,02 \\ 5 & 0 & 2,83 \\ 7,02 & 2,83 & 0 \end{pmatrix} \end{matrix}$$

Отже,

$$P_3 : \{1, 2\}, \{3\}, \{4, 5\}.$$

Наступні розбиття на кластери будуть такими:

$$P_2 : \{1, 2\}, \{3, 4, 5\};$$

$$P_4 : \{1, 2, 3, 4, 5\}.$$

Дендрограма така сама, як і на рис. 11.3.

11.1.5 Метод Уорда

Метод полягає в тому, що при переході від одного розбиття до наступного об'єднують ті два кластери, при об'єднанні яких відбувається мінімальне збільшення загальної втрати інформації. За втрату інформації для однієї групи беруть звичайно середньоквадратичне відхилення, а для кількох груп – суму всіх групових відхилень.

Наприклад, нехай маємо 10 осіб, для яких отримано значення деякої характеристики:

$$2, 6, 5, 6, 2, 2, 2, 0, 0, 0.$$

Якщо кожен кластер містить лише одного індивіда, то оскільки відхилення в кожному кластері – 0, то загальне відхилення дорівнює 0. Нехай маємо розбиття, яке складається лише з одного кластера, що містить усіх індивідів. Середнє значення характеристики тоді:

$$\frac{2 + 6 + 5 + 6 + 2 + 2 + 2 + 0 + 0 + 0}{10} = 2,5.$$

Відхилення в цьому разі:

$$(2 - 2,5)^2 + (6 - 2,5)^2 + (5 - 2,5)^2 + \dots + (0 - 2,5)^2 = 50,5.$$

Якщо ж кластеризувати індивідів таким чином:

$$\{0, 0, 0\}, \{2, 2, 2, 2\}, \{5\}, \{6, 6\},$$

то відповідні середні груп будуть:

$$\bar{x}_1 = 0, \quad \bar{x}_2 = 0, \quad \bar{x}_3 = 0, \quad \bar{x}_4 = 0.$$

Відхилення для кожної групи нульове, і тому загальне відхилення дорівнює нулеві також.

11.2 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 зумовлений 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, вказуючи відстані, на яких відбувається утворення нових кластерів, то часто великі зміни будуть свідчити про правильний вибір кластерів.

Наприклад, розглянемо дендрограму на рис. 11.6.

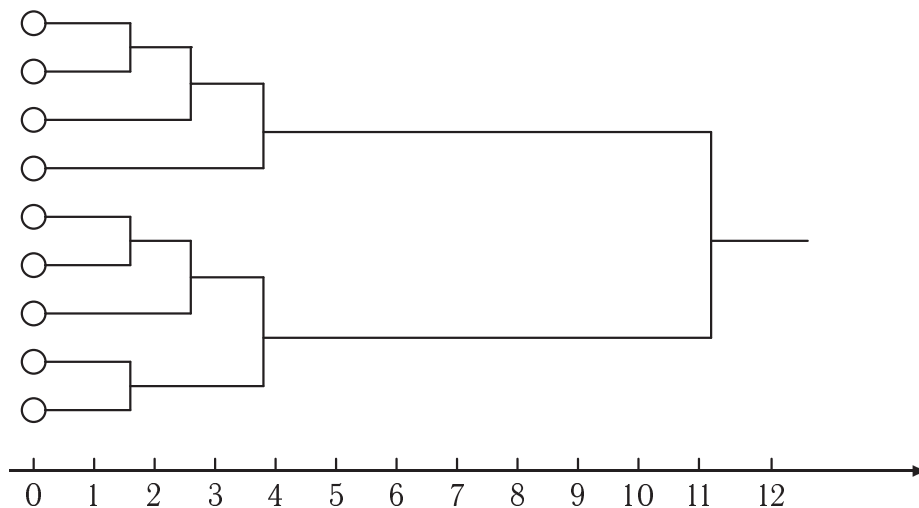


Рис. 11.6. Дендрограма з відстанями

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

11.3 Методи подрібнення

Методи подрібнення порівняно з агломеративними методами застосовують рідше. Тому опишемо лише два з них.

Перший метод полягає в тому, що спочатку з кластера, що містить усі індивіди, виділяють один, який найбільше відмінний від решти. Цей індивід і утворює новий кластер. Потім до новоутвореного приєднують елементи із залишку початкового кластера в порядку зменшення їх від-

мінності від елементів, що залишаються. Процес зупиняємо, якщо середня відмінність елемента від нового кластера стає більша, ніж середня відмінність від залишку початкового кластера. Далі процес повторюємо для кожного з двох новоутворених кластерів.

Наприклад, нехай ми маємо 7 осіб і матриця відстаней між їхніми векторами характеристик така:

$$\begin{matrix} & \{1\} & \{2\} & \{3\} & \{4\} & \{5\} & \{6\} & \{7\} \\ \begin{matrix} \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \\ \{6\} \\ \{7\} \end{matrix} & \begin{pmatrix} 0 & 10 & 7 & 30 & 29 & 38 & 42 \\ 10 & 0 & 7 & 23 & 25 & 34 & 36 \\ 7 & 7 & 0 & 21 & 22 & 31 & 36 \\ 30 & 23 & 21 & 0 & 7 & 10 & 13 \\ 29 & 25 & 22 & 7 & 0 & 11 & 17 \\ 38 & 34 & 31 & 10 & 11 & 0 & 9 \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix} \end{matrix}$$

У першого індивіда середня відстань до тих, що залишились, більша, ніж у решти індивідів. Тому маємо дві групи $\{1\}$ і $\{2, 3, 4, 5, 6, 7\}$. Легко отримати такі дані:

Індивід	Середня відстань до групи $\{2, 3, 4, 5, 6, 7\}$	Середня відстань до групи $\{1\}$	Різниця середніх відстаней
2	25	10	15
3	23,4	7	16,4
4	14,8	30	-15,2
5	16,4	29	-12,6
6	19	38	-19
7	22,2	42	-19,8

Отже, нові дві групи – це $\{1, 3\}$ та $\{2, 4, 5, 6, 7\}$. Для залишку основної групи отримуємо такі дані:

Індивід	Середня відстань до групи $\{2, 4, 5, 6, 7\}$	Середня відстань до групи $\{1, 3\}$	Різниця середніх відстаней
2	29,5	8,5	21
4	13,2	25,5	-12,3
5	15	25,5	-10,5
6	16	34,5	-18,5
7	18,7	39	-20,3

Отримуємо розбиття на групи: $\{1, 2, 3\}$ та $\{4, 5, 6, 7\}$. Тепер маємо:

Індивід	Середня відстань до групи $\{4, 5, 6, 7\}$	Середня відстань до групи $\{1, 2, 3\}$	Різниця середніх відстаней
4	10	24,3	-14,3
5	11,7	25,3	-13,6
6	10	34,3	-24,3
7	13	38	-25,0

Отже, маємо два кластери: $\{1, 2, 3\}$ та $\{4, 5, 6, 7\}$. Далі для кожного з них процес розбиття на групи повторюємо. Так продовжуємо доти, доки на дійдемо до елементарного розбиття – по індивіду в кластері.

Наступний метод застосовують до даних із бінарними характеристиками. Поділ на кластери відбувається таким чином: за деяким критерієм вибирають певну характеристику. Індивіди потрапляють у різні кластери, якщо вони мають різні значення цієї характеристики.

Наведемо декілька найбільш уживаних критеріїв для вибору відокремлювальної характеристики. Будемо використовувати позначення, введені в попередньому розділі.

За відокремлювальну характеристику беруть ту, для якої максимальний вираз

$$1. \sum_{j \neq k} x_{jk}^2; \quad 2. \sum_{j \neq k} \sqrt{x_{jk}^2}; \quad 3. \sum |ad - bc|; \quad 4. \sum (ad - bc)^2,$$

$$\text{де } x_{jk}^2 = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}.$$

Наприклад, нехай у нас є 5 індивідів, інформацію про три характеристики яких задано матрицею:

$$\begin{matrix} \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{matrix} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Будемо використовувати перший критерій.

Оскільки:

$$x_{12}^2 + x_{13}^2 = 1,87 + 2,22 = 4,09,$$

$$x_{12}^2 + x_{23}^2 = 1,87 + 0,83 = 2,7,$$

$$x_{13}^2 + x_{23}^2 = 2,22 + 0,83 = 3,05,$$

то розділення на кластери відбувається за значенням першої характеристики (сума 4,09 є максимальною). У результаті отримуємо два кластери {1, 5} та {2, 3, 4}.

11.4 Виконання в пакеті STATISTICA

В STATISTICA реалізовані такі методи кластеризації – агломеративні методи: *joining* (tree clustering), *two way joining*, а також метод *k*-середніх – *k-means clustering*.

Здебільшого перед початком класифікації дані стандартизують (обчислюють середнє, і ділять на квадратний корінь з дисперсії). Отримані в результаті стандартизації змінні мають нульове середнє й одиничну дисперсію. Дані, які ми розглядаємо далі, – вже стандартизовані.

У STATISTICA можна вибрати такі правила ієрархічного об'єднання кластерів:

Single linkage – метод одиночного зв'язку;

Complete linkage – метод повного зв'язку;

Unweighted pair group average – незважений метод “середнього зв'язку”;

Weighted pair group average – зважений метод “середнього зв'язку”;

Weighted centroid pair group – зважений центроїдний метод;

Ward method – метод Уорда.

Ці алгоритми відрізняються правилами об'єднання об'єктів у кластер.

У методі одиночного зв'язку на першому кроці об'єднують два об'єкти, які мають між собою максимальну міру подібності. На наступному кроці до них приєднують об'єкт з максимальною мірою подібності з одним із об'єктів кластера. У такий спосіб, процес продовжують далі. Отже, для включення об'єкта в кластер потрібна максимальна подібність лише з одним членом кластера. Звідси і назва методу одиночного зв'язку: потрібен тільки один зв'язок, для того, щоб приєднати об'єкт до кластера – зв'язок нового елемента з кластером визначається тільки за одним з елементів кластера. Вадю цього методу є утворення дуже великих “продовгуватих” кластерів.

Метод повних зв'язків дозволяє усунути цей недолік. Тут міра подібності між об'єктом – кандидатом на включення в кластер і всіма членами кластера не може бути меншою від деякого порогового значення.

У методі середнього зв'язку міра подібності між кандидатом і членами кластера середня, наприклад, беруть просто середнє арифметичне мір подібності.

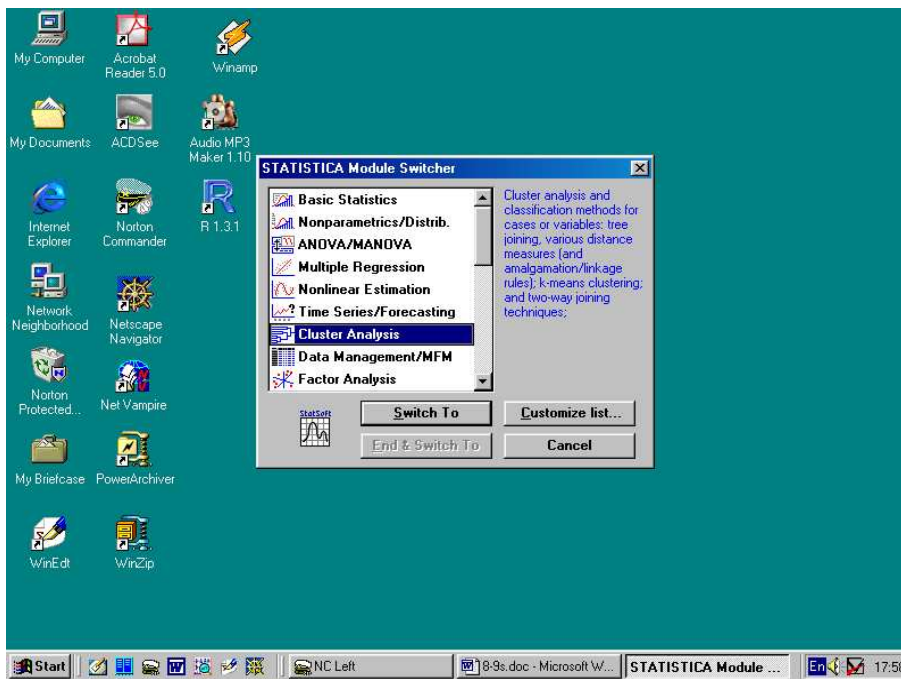


Рис. 11.7. Кластерний аналіз

Ідея ще одного агломеративного методу – методу Уорда полягає в тому, щоб проводити об'єднання, яке дає мінімальний приріст внутрішньогрупової суми квадратів відхилень. Зауважено, що метод Уорда приводить до утворення кластерів приблизно рівних розмірів, які мають форму гіперсфер.

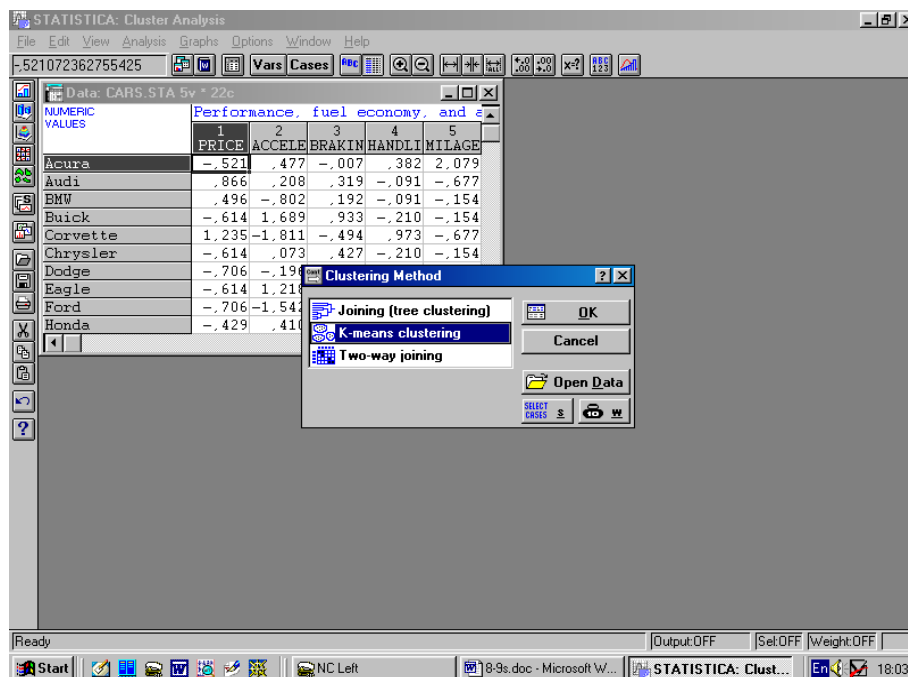
Розглянемо ще ітеративний метод групування k -середніх – *k-means clustering*. Цей метод працює безпосередньо з об'єктами, а не з матрицею подібності. В методі k -середніх об'єкт зараховують до того класу, віддаль до якого мінімальна. Розглядають евклідову віддаль, тобто об'єкти – точки евклідового простору.

Запуск модуля

Виберіть у перемикачі модулів назву модуля – *Cluster Analysis* (Кластерний аналіз), висвітіть його ім'я і натисніть кнопку *Switch To* або просто двічі клацніть на його імені (рис. 11.7). У робочому вікні STATISTICA клацніть на пункт *Analysis* (аналіз). В меню, яке випадає, виберіть *Startup Panel* (стартова панель). На екрані з'явиться стартова панель модуля *Cluster Analysis* (Кластерний аналіз) (рис. 11.8.)

Вибір методу

Подивіться на стартову панель. У головній її частині є список методів

Рис. 11.8. Метод *k-means*

кластерного аналізу, реалізованих у STATISTICA.

У списку методів висвітїть *k-means* (к-середніх) (див. рис. 11.8) і натисніть кнопку ОК в правому верхньому куті панелі.

Діалогове вікно методу *k-means* з'явиться на екрані (рис. 11.9).

Вибір змінних, встановлення початкових значень, запуск обчислювальної процедури методу *k-середніх*

Почніть працювати в цьому вікні. Передусім виберіть змінні для аналізу. Натисніть кнопку *Variables* (змінні) в лівому верхньому куті активного вікна і відкрийте діалогове вікно: *Select variables for analysis* (рис. 11.10).

Нехай потрібно враховувати всі параметри. Тоді натисніть спочатку кнопку *Select All* (вибрати все), а потім – кнопку ОК.

Погляньте далі на поле *Cluster* (Кластер), яке розташоване нижче кнопки *Variables* (змінні). Натисніть стрілку в цьому полі, виберіть пункт меню *Cases* (Випадки). Так діють, коли кластеризують змінні.

У полі *Number of Cases* (Кількість кластерів) потрібно визначити кількість груп, на які ми хочемо розбити змінні. Запишіть у це поле число 3.

Таким чином, ми будемо розбивати дані на три кластери.

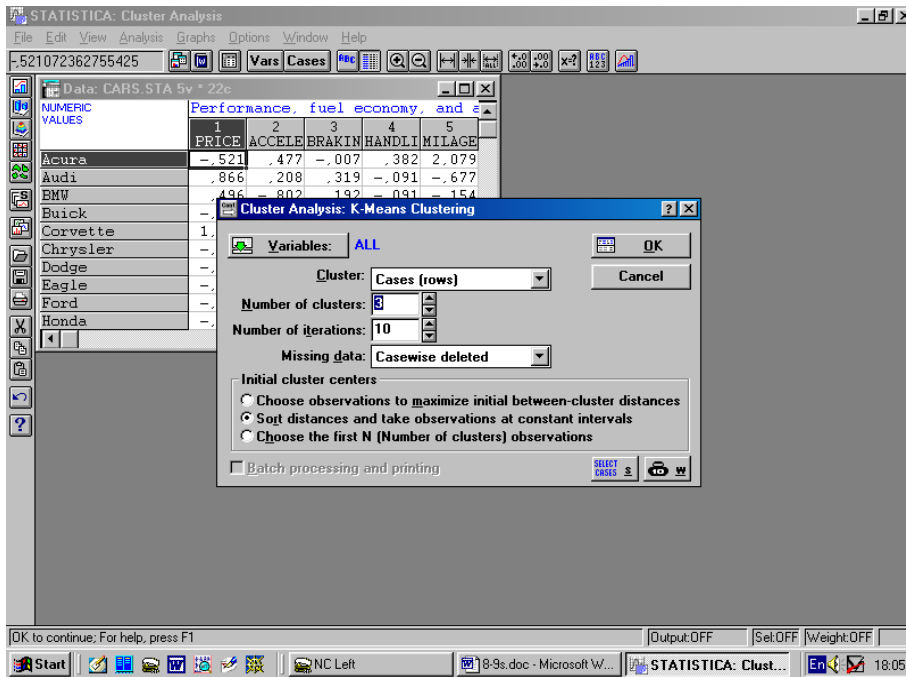
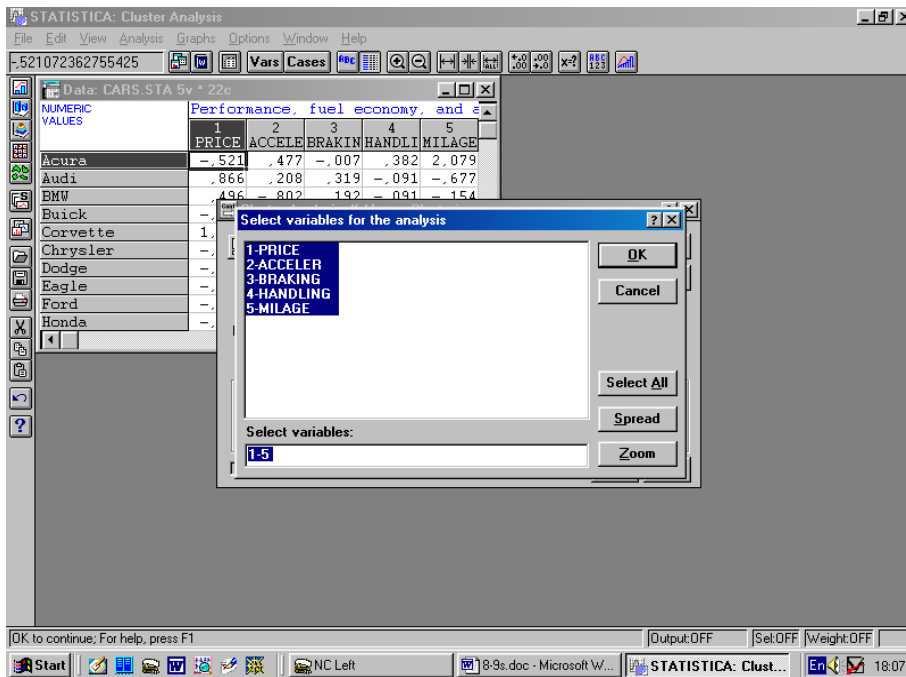
Рис. 11.9. Діалогове вікно методу *k-means*

Рис. 11.10. Вибір змінних для кластерного аналізу

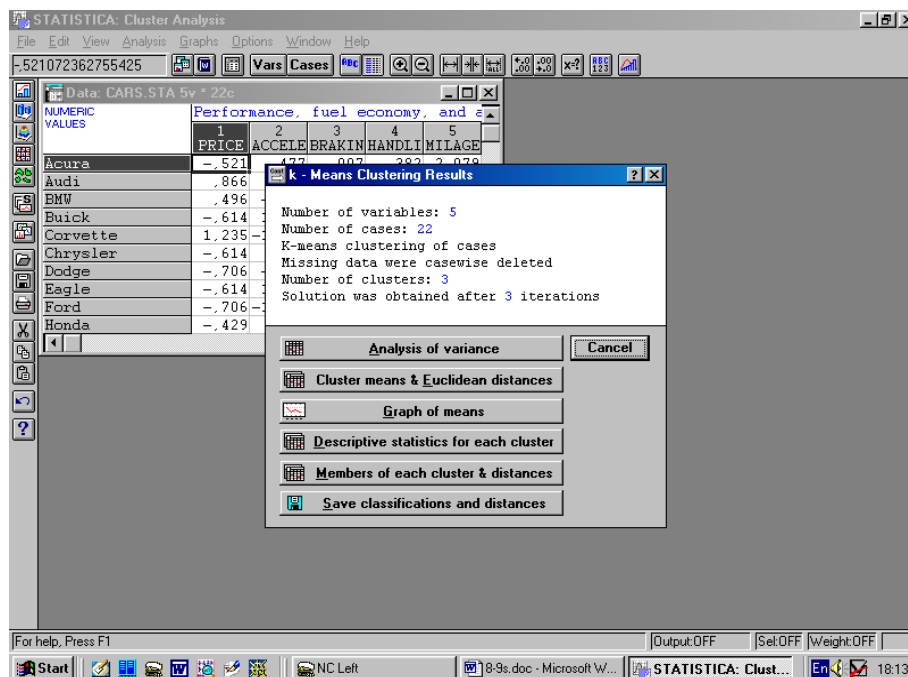


Рис. 11.11. Вікно результатів кластеризації за методом середніх

У рядку *Number of iterations* (кількість ітерацій) задають максимальну кількість ітерацій, які використовуються при побудові класів. Задайте, наприклад, число 11.

У рядку *Missing Data* задають спосіб обробки пропущених значень у даних (наприклад, для якогось об'єкту відсутнє значення деякого параметра). Якщо пропусків в даних немає, то обробка пропущених значень не відбувається.

Група опцій *Initial cluster centers* дозволяє задати початкові центри кластерів.

Після того, як всі установки зроблені, натисніть кнопку ОК у верхньому правому куті вікна *k-means Clustering* і запустіть обчислювальну процедуру.

Перегляд результатів кластеризації

Через декілька секунд після натискання кнопки ОК у вікні *k-means Clustering* вікно результатів з'явиться на екрані (рис. 11.11).

У верхній частині вікна записана інформація: кількість змінних, кількість випадків, метод кластеризації, кількість кластерів, а також повідомлення про те, після скількох ітерацій знайдено рішення: *Solution was obtained after 3 iterations* – Розв'язок знайдено після 3 ітерацій.

Кнопки в нижній частині вікна дозволяють провести аналіз результатів кластеризації.

Кнопка *Analysis of variation* (Дисперсійний аналіз) дозволяє продивитися таблицю дисперсійного аналізу.

Кнопка *Cluster Means&Euclidean Distances* дозволяє вивести таблиці, в першій із яких вказані середні для кожного кластера (знаходження середнього проводиться всередині кластера), в другій вказані евклідові відстані і квадрати евклідових відстаней між кластерами.

Кнопка *Graph of means* дозволяє продивитися середні значення для кожного кластера на лінійному графіку.

Кнопка *Descriptive Statistics for each clusters* відкриває електронну таблицю з описовими статистиками для кожного кластера (середнє, дисперсія і т.д.)

Кнопка *Save classifications and distances* дозволяє зберегти результати класифікації у файлі для подальшого дослідження.

Нам, звичайно, цікаво подивитися, як розподілилися об'єкти за кластерами. Для цього потрібно натиснути кнопку *Member of each cluster&distances*. На екрані з'являться електронні таблиці з назвами об'єктів, віднесені до визначених кластерів. У рядках таблиць вказано відстань від кожного об'єкта до центра кластера.

Натисніть на кнопку *Cluster Means&Euclidean Distances*. На екрані з'явиться таблиця, в якій дані евклідові відстані між середніми кластерів (для кожного із параметрів всередині кластера обчислюють середнє, отримують точки в багатовимірному просторі і між ними знаходять відстані) (рис. 11.12).

Над діагоналлю в таблиці дані квадрати відстаней між кластерами.

За допомогою кнопки *Graph of means* (графік середніх) будують графіки середніх значень характеристик об'єктів для кожного кластера (див. рис. 11.13).

Інші методи кластеризації, реалізовані в системі

У системі реалізовані також інші методи кластеризації, а саме так званий *two way joining*, в якому кластеризуються випадки і змінні одночасно.

На рис. 11.14 показаний результат кластеризації об'єктів методом *two way joining*.

Якщо ви скористаєтеся *Joining (tree clustering)*, то зможете побачити дендрограму, або дерево об'єднання (рис. 11.15), про яке ми говорили раніше.

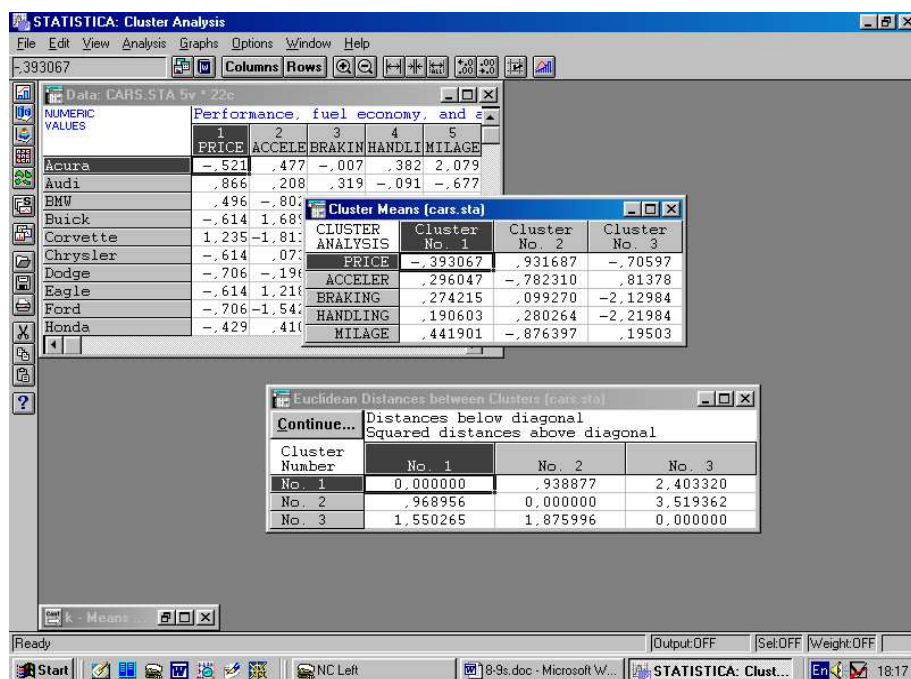


Рис. 11.12. Відстань між кластерами

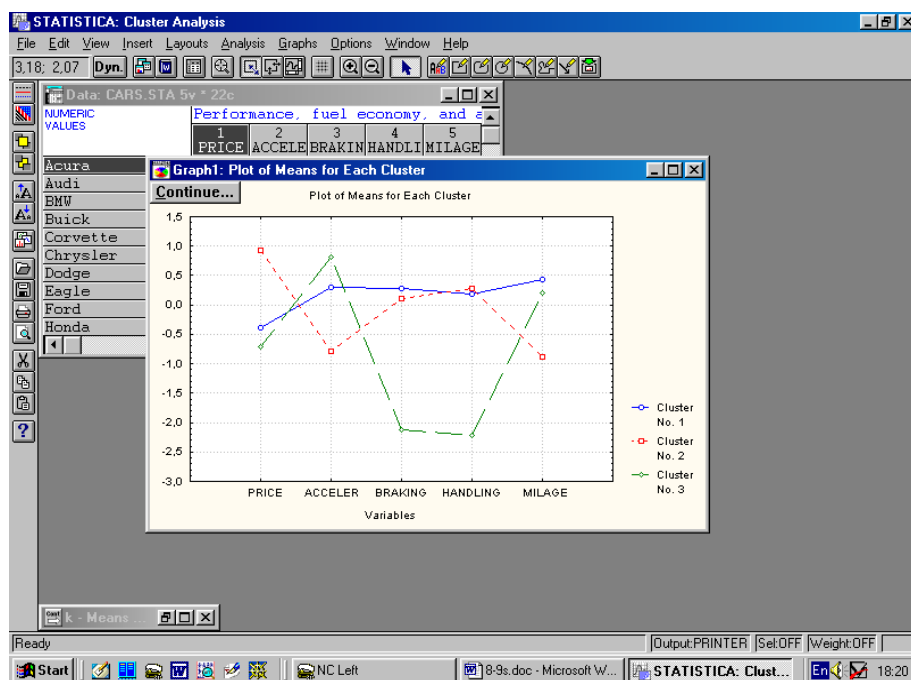


Рис. 11.13. Графік середніх для кожного кластера

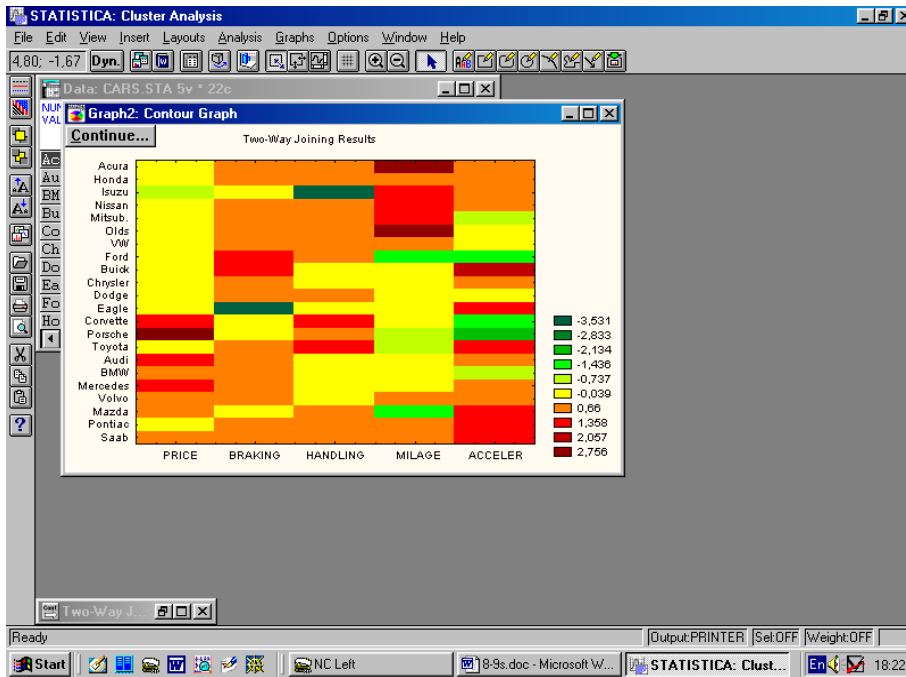


Рис. 11.14. Метод *two way joining*

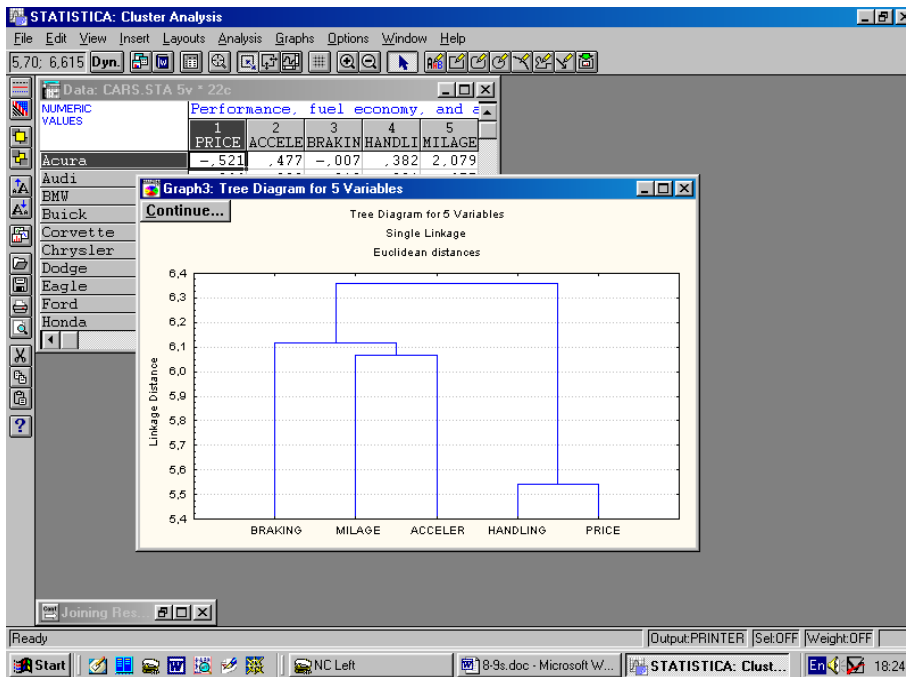


Рис. 11.15. Вертикальна дендрограма

Розділ 12

Дисперсійний аналіз

Дисперсійний аналіз широко використовують як у соціальних, так і технічних дослідженнях. Застосовують його в тих випадках, коли є потреба з'ясувати вплив різних факторів на значення деякої величини. Причому фактори переважно мають якісний характер і можуть мати скінченну 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, у загальній дисперсії свідчить про статистично значущий зв'язок між фактором та досліджуваною ознакою.

12.1 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), ε_{ij} – результат впливу неврахованих факторів. Вважатимемо, що величини ε_{ij} є реалізаціями центрованої нормально розподіленої випадкової величини з дисперсією σ^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, загалом кажучи, різні математичні сподівання a_1, \dots, a_k та однакові дисперсії σ^2 .

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

Нехай \bar{x}_i – вибіркове середнє i -тої вибірки,

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

\bar{x} – вибіркове середнє об'єднаної вибірки,

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

де n – загальна кількість спостережень.

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

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

Позначивши загальну суму квадратів

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

суму квадратів відхилень вибірових середніх \bar{x}_i від загального середнього \bar{x}

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

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

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

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 і мають розподіли χ^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 σ^2 . Оцінка S_1^2 характеризує розсіювання групових середніх, а оцінка S_2^2 – розсіювання всередині груп, яке зумовлене випадковими варіаціями результатів спостережень (впливом неврахованих факторів). Значну перевагу величини S_1^2 над S_2^2 можна пояснити відмінністю середніх у 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 значущості α) результатам спостережень, якщо вибіркове значення F_B 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 $F_B \geq F_{1-\alpha}(k-1, n-k)$, то гіпотезу H_0 відхиляють і потрібно вважати, що серед середніх a_1, a_2, \dots, a_k є хоча б два різні. При цьому в $\alpha \cdot 100\%$ випадків буде допущено помилку (відхилено правильну гіпотезу).

12.2 Виконання в пакеті STATISTICA

У трьох магазинах, що торгують однотипними товарами, зібрано дані про товарообіг за 8 місяців роботи (в тис. грн.)

Магазин	Товарообіг за місяць							
	1	2	3	4	5	6	7	8
1	19	23	26	18	20	20	18	35
2	20	20	32	27	40	24	22	18
3	16	15	18	26	19	17	19	18

Перевірити гіпотезу про рівність середніх значень товарообігу в різних магазинах. Якщо гіпотеза буде відхилена, провести попарне порівняння середніх.

Будемо виконувати дії в модулі *Basic Statistics and Tables* (можна виконувати також у модулі *ANOVA/MANOVA*). Спочатку перевіримо гіпотезу про рівність середніх.

1. Створимо таблицю з двома стовпцями М (магазин) і Т (товарообіг) та 24 (3×8) рядками.

2. У Т занесемо дані про товарообіг, а у М – номери магазинів (рівні фактора) М1, М2, М3 (рис. 12.1).

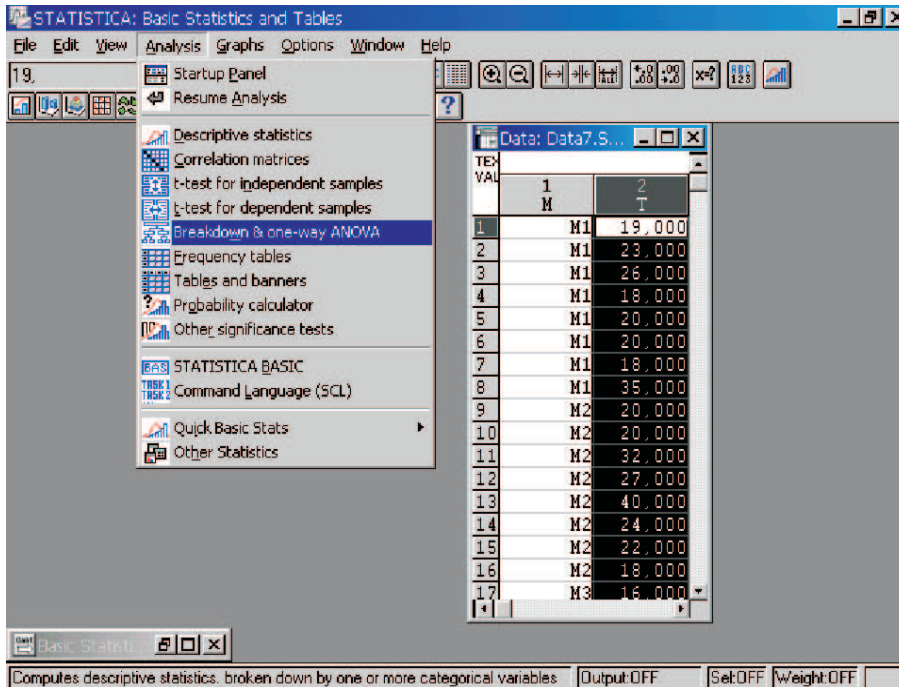


Рис. 12.1. Дані про товарообіг

3. Далі в меню вибираємо *Analysis*; потім вибираємо *Breakdown and one-way ANOVA*; у вікні, що відкрилося (рис. 12.2), вибираємо:
- Analysis: Detailed Analysis Of Individual tables*,
 - Variables*:
 - *Grouping variables* (групуючі змінні): 1-M,
 - *Dependent variables* (залежні змінні – відгуки): 2-P,
 - ОК;
 - Codes for grouping variables: All+OK*;
 - ОК
4. У вікні (*Descriptive Statistics and Correlations by groups – Results*), що відкрилося (рис. 12.3), вибираємо:
- Statistics*:
 - *Number of observations* (кількість спостережень),

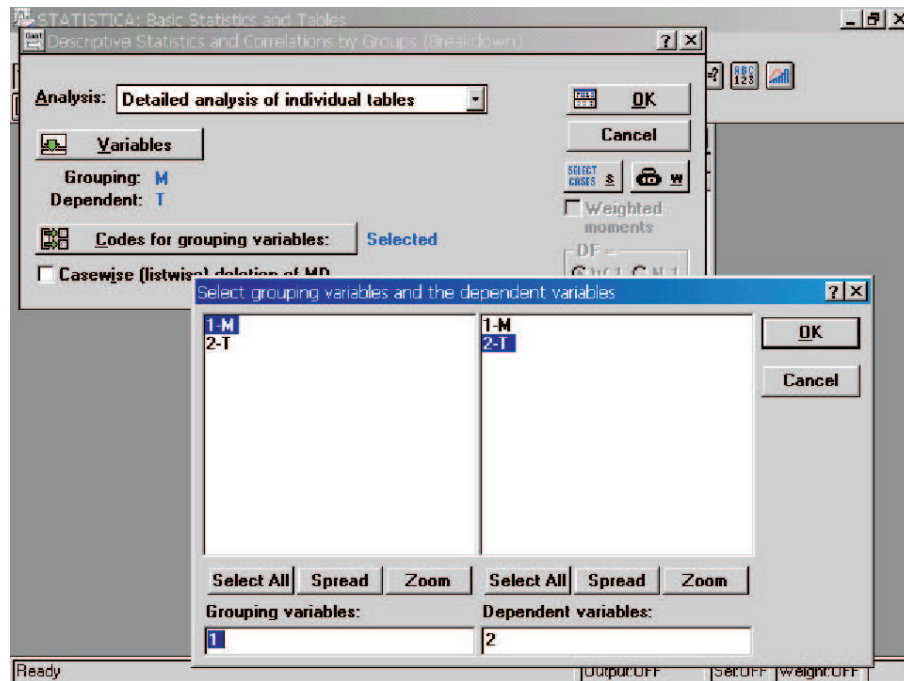


Рис. 12.2. Вибір групуючої і залежної змінних

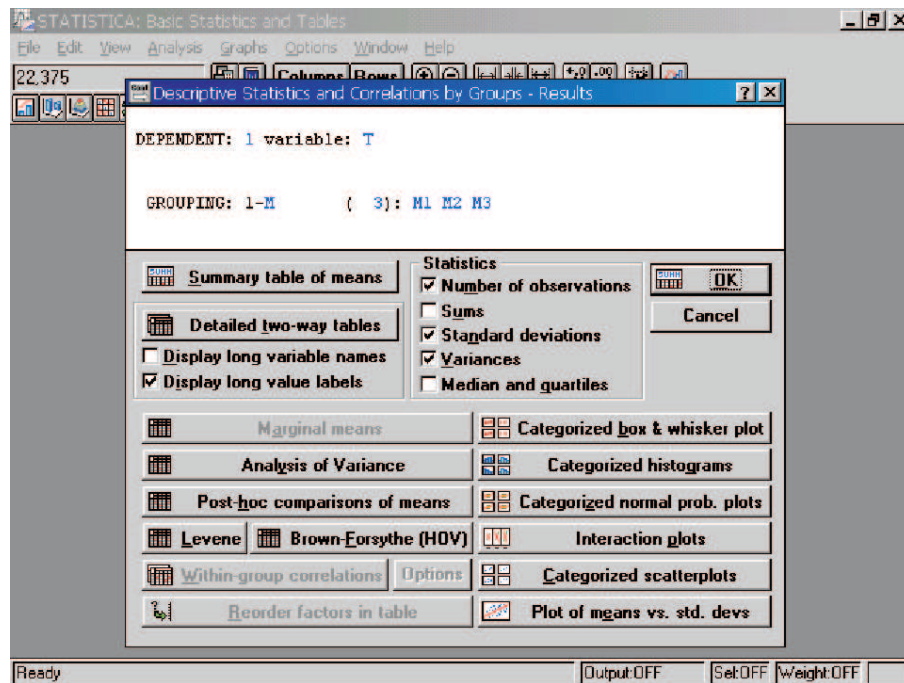
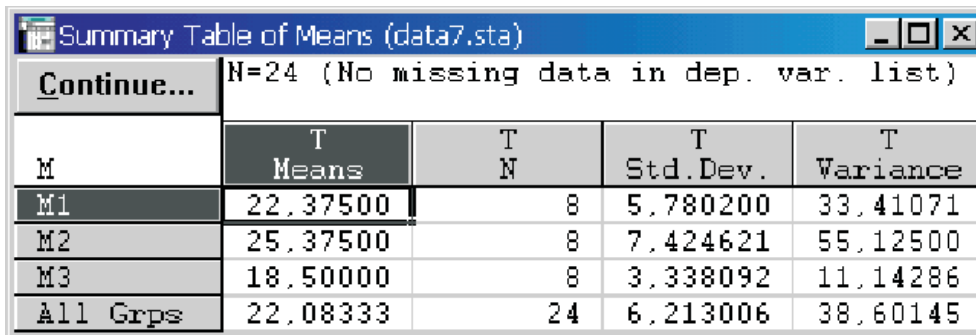
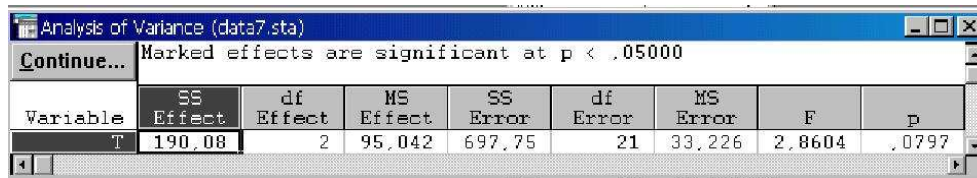


Рис. 12.3. Вікно аналізу результатів



Summary Table of Means (data7.sta)				
N=24 (No missing data in dep. var. list)				
M	T Means	T N	T Std.Dev.	T Variance
M1	22,37500	8	5,780200	33,41071
M2	25,37500	8	7,424621	55,12500
M3	18,50000	8	3,338092	11,14286
All Grps	22,08333	24	6,213006	38,60145

Рис. 12.4. Різниця між середніми



Analysis of Variance (data7.sta)								
Marked effects are significant at p < .05000								
Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
T	190,08	2	95,042	697,75	21	33,226	2,8604	,0797

Рис. 12.5. Таблиця результатів *Analysis of Variance*

- *Standard deviations* (стандартні відхилення),
 - *Variances* (дисперсії);
- (b) ОК. У вікні *Summary Table of Means* (рис.12.4) бачимо, як відрізняються середні в залежності від рівня фактора M.
 - (c) Повертаємось у вікно *Descriptive Statistics and Correlations by groups – Results*, натиснувши на ×;
 - (d) Вибираємо *Analysis of Variance* і отримуємо таблицю результатів *Analysis of Variance* (рис. 12.5).

$p = 0,079683$ – ймовірність, з якою можна стверджувати, що середні рівні між собою. Оскільки p надто мале, щоб вважати середні однаковими, то гіпотезу про рівність середніх відхиляємо. Проведемо попарне порівняння середніх методом лінійних контрастів (метод Шеффе).

1. Повернемо до вікна *Descriptive Statistics and Correlations by groups – Results*;
2. Виберемо *Post-hoc comparisons of means* (рис. 12.6);
3. Виберемо *Sheffe test*;
4. У вікні, що відкрилося, читаємо рівні значущості для гіпотез про рівність усіх пар середніх (рис. 12.7).

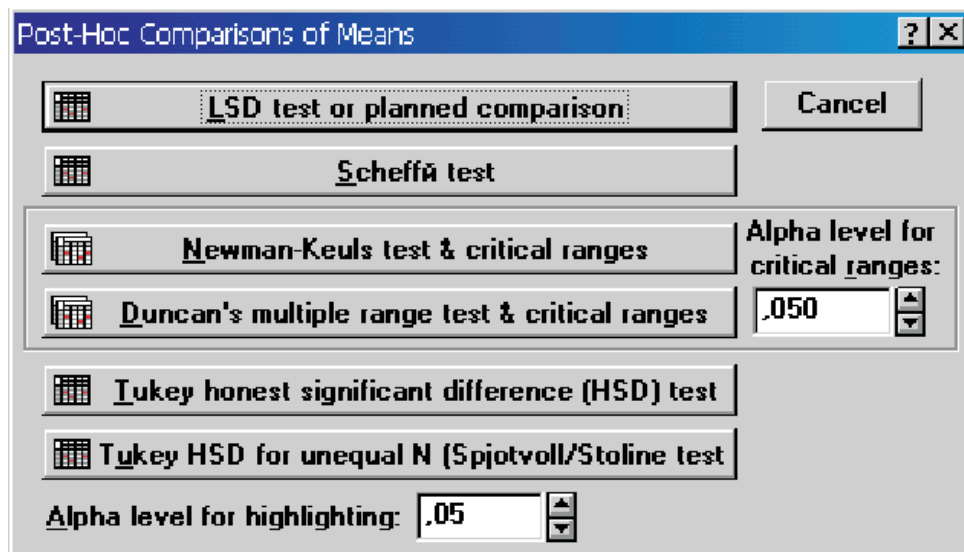


Рис. 12.6. Порівняння середніх

Scheffe Test; Variable: T (data7.sta)			
Marked differences are significant at $p < .05000$			
	{1}	{2}	{3}
M	M=22,375	M=25,375	M=18,500
M1 {1}		,589648	,420194
M2 {2}	,589648		,080652
M3 {3}	,420194	,080652	

Рис. 12.7. Рівні значущості для гіпотез про рівність пар середніх

З результатів можна зробити висновок, що слід вважати різними середні значення товарообігу в другому та третьому магазинах.

12.3 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 значення через x_{ijk} ($i = \overline{1, p}$, $j = \overline{1, q}$, $k = \overline{1, n}$).

Введемо такі характеристики:

1. Середні значення

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

середні значення ознаки при впливі кожної пари рівнів обох факторів;

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

середні значення ознаки при кожному рівні фактора A ;

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

середні значення ознаки при кожному рівні фактора B ;

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

загальне середнє значення ознаки X .

2. Суми квадратів відхилень та виправлені дисперсії

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

зумовлені впливом фактора A на ознаку X ;

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

зумовлені впливом фактора B на ознаку X ;

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

зумовлені впливом на ознаку X обох факторів A і B ;

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

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 значущості α , 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 α . Подібні висновки можна зробити і для впливу фактора B та обох факторів разом.

12.4 Виконання в пакеті STATISTICA

Досліджувався вплив факторів A і B на рейтинг правих політичних партій (у відсотках): фактор A – регіони (A_1 – західний, A_3 – центральний, A_2 – східний); фактор B – вік опитаних (B_1 – 20 – 35 років, B_2 – 35 – 50 років, B_3 – 50 – 70 років). Результати досліджень наведені в таблиці:

Фактор B	Фактор A					
	A_1		A_2		A_3	
B_1	25,2	10,2	4,3	10,5	14,3	10,6
	5,4	13,2	20,3	32,4	28,4	10,8
	18,2	5,2	5,6	12,4	7,4	6,5
	13,4	15,2	6,2	9,8	4,5	26,3
	4,5	19,2	16,8	18,4	30,2	11,8
B_2	10,6	8,4	12,4	4,3	6,2	7,5
	11,2	4,6	13,2	5,6	3,5	12,4
	5,8	18,2	8,9	14,8	13,5	16,4
	16,4	13,2	22,3	6,8	7,9	8,9
	4,8	8,9	7,2	11,4	15,4	10,8
B_3	2,5	6,4	4,5	4,9	14,8	2,9
	12,5	14,8	12,3	15,6	5,9	10,6
	12,3	8,5	7,9	8,9	8,5	13,4
	5,9	8,9	9,8	13,9	2,2	19,5
	15,4	12,8	4,2	6,9	7,9	9,9

Використаємо модуль *ANOVA/MANOVA*.

1. Створимо таблицю з трьома стовпцями X (рейтинг), A (регіони), B (вік) та 90 ($3 \times 3 \times 10$) рядками.
2. У X занесемо дані про рейтинг, у A – індекс регіону, а в B – індекс віку (рис. 12.8).
3. В меню модуля вибираємо *Analysis*, а потім пункт *Resume Analysis*.

TEX	1 X	2 A	3 B
17	16,400	A1	B2
18	13,200	A1	B2
19	4,800	A1	B2
20	8,900	A1	B2
21	2,500	A1	B3
22	6,400	A1	B3
23	12,500	A1	B3
24	14,800	A1	B3
25	12,300	A1	B3
26	8,500	A1	B3
27	5,900	A1	B3
28	8,900	A1	B3
29	15,400	A1	B3
30	12,800	A1	B3
31	4,300	A2	B1
32	10,500	A2	B1
33	20,300	A2	B1
34	32,400	A2	B1

Рис. 12.8. Таблиця з даними

4. У вікні, що відкріється (рис. 12.9), натискаємо:

(a) *Variables* та вводимо інформацію:

- *Independent variables (factors)*: 2-A, 3-B;
- *Dependent variable list*: 1-X;
- ОК.

(b) Натискаємо ОК. У вікні *ANOVA Results*, що з'являється (рис. 12.10) вибираємо *All effects* і читаємо результат у *Summary of all Effects* (для кожного фактора та їх сукупності):

df Effect: $p - 1 = 2$, $q - 1 = 2$, $(p - 1)(q - 1) = 4$;

MS Effect: $S_1^2 = 163,9488$, $S_2^2 = 3,1214$, $S_3^2 = 6,1313$;

df Error: $pq(n - 1) = 81$;

MS Error: $S_4^2 = 38,06314$;

F: $F_A = 4,307285$, $F_B = 0,082007$, $F_{AB} = 0,161082$;

p - level: $p_A = 0,016684$, $p_B = 0,921342$, $p_{AB} = 0,957390$.

З отриманих результатів можна зробити висновок, що фактор *A* (регіон) істотно впливає на рейтинг правих політичних сил (ймовірність помилки не перевищує 0,017), а про наявність впливу фактора *B* (вік) на підставі отриманих результатів статистичних досліджень говорити не можна. Останнє стосується і сукупного впливу цих факторів.

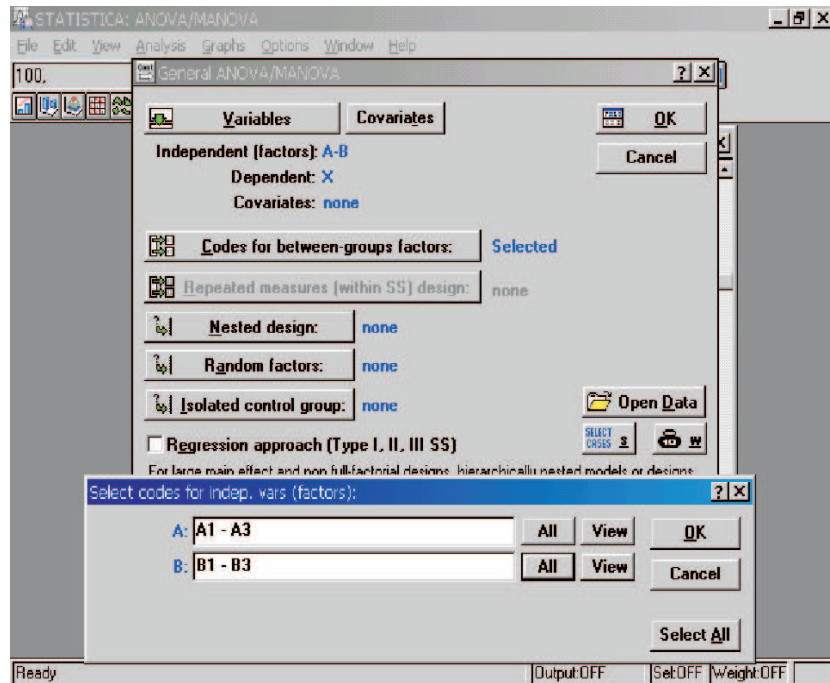


Рис. 12.9. Вибір залежних і незалежних змінних

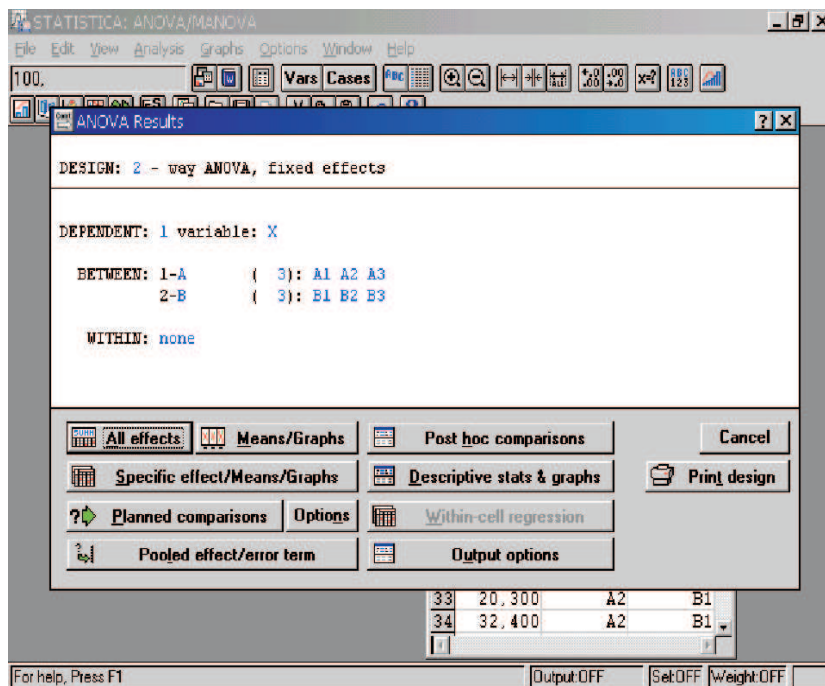


Рис. 12.10. Вікно аналізу результатів

Розділ 13

Категоризовані дані

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 багатой. Так неперервні величини можна штучно перетворити на 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 потрапили в певні категорії чи класи. В цьому випадку для описання категоризованих даних важливу роль відіграє мода – значення з найбільшою частотою.

13.1 Гіпотези про розподіл частот

Адаптуємо тест, розглянутий в § 8.3.1, 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 ν_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 $\hat{\nu}_i$ та використаємо статистику – міру відхилення їх від теоретичних частот

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

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

13.2 Виконання в пакеті STATISTICA

У середовищі кримінологів активно обговорюють питання про зв'язок між порою року та рівнем злочинності. Було вивчено 1361 вбивство та розподілено їх за чотирма порами року. Результати виявились такими:

Пора року			
Зима	Весна	Літо	Осінь
328	334	372	327

Чи можна, на основі цих даних, зробити висновок про наявність згаданої залежності?

Розглянемо гіпотезу про рівність частот вбивств у різні пори року (відсутність зв'язку між злочинністю та порою року). Гіпотетичні частоти всі однакові і дорівнюють $1361 : 4 = 340,25$.

Використаємо модуль *Nonparametric Statistics*, процедуру *Observed versus expected XI*. Створимо два набори даних VAR1 – реальні частоти, VAR2 – гіпотетичні частоти (див. рис. 13.1). Вибравши у вікні *Variables* (рис. 13.2) (*Observed frequencies: 1, Expected frequencies 2*), одержимо результат (після натискування потрібної кількості ОК), наведений на рис. 13.3.

Тому можемо зробити висновок, що гіпотеза про рівність частот вбивств залежно від пори року не суперечить наявним даним (з рівнем значущості меншим, ніж 0,25778). Отже, немає підстав стверджувати про залежність рівня злочинності від пори року.

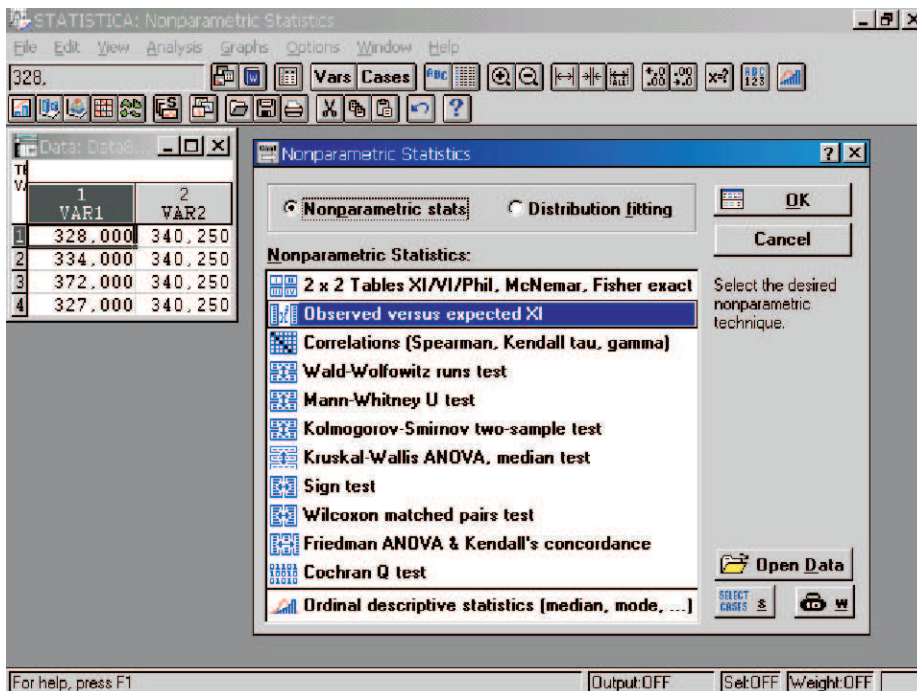


Рис. 13.1. Порівняння частот

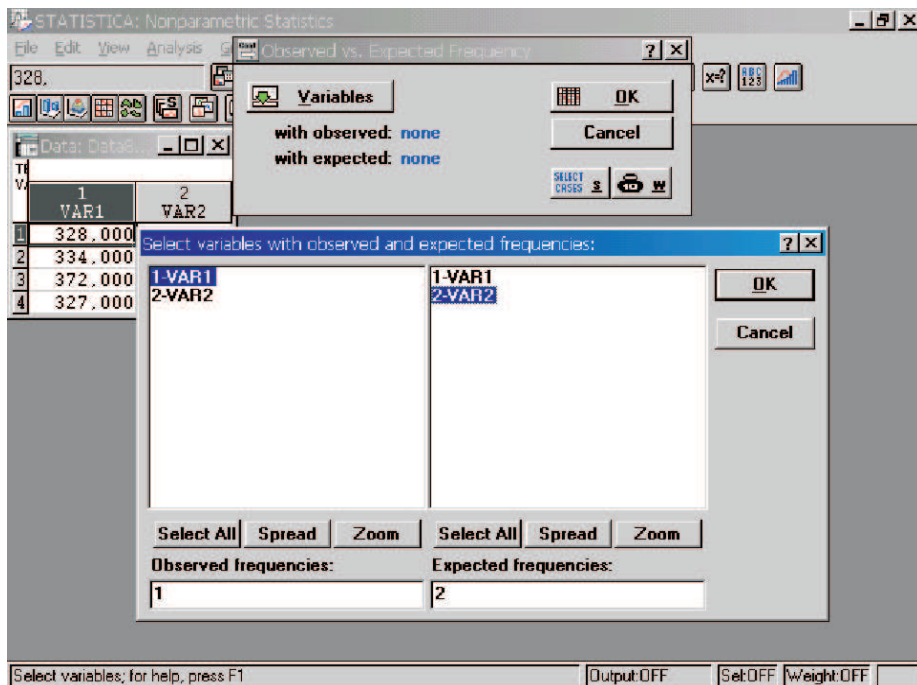


Рис. 13.2. Вибір змінних

Case	observed VAR1	expected VAR2	O - E	(O-E)**2 /E
C: 1	328,000	340,250	-12,2500	,441036
C: 2	334,000	340,250	-6,2500	,114805
C: 3	372,000	340,250	31,7500	2,962711
C: 4	327,000	340,250	-13,2500	,515981
Sum	1361,000	1361,000	0,0000	4,034534

Рис. 13.3. Вікно результатів

13.3 Гіпотези про незалежність ознак

Адаптуємо тест, розглянутий в § 8.5, до категоризованих даних. Розглянемо дві категоріальні змінні з k та l рівнями ознак, відповідно. Нехай проведено n експериментів, у 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, у яких було одержано i -тий рівень першої ознаки, а $n_{\cdot j}$ – кількість експериментів, у яких було одержано j -тий рівень другої ознаки. Обчислимо для кожної пари рівнів ознак число $\tilde{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$. За умови, 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

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

має розподіл χ^2 з $(k-1)(l-1)$ ступенями вільності (за умови, що всі $\tilde{n}_{ij} \geq 4$).

Отже, гіпотезу H_0 про незалежність розглянутих ознак приймають (не суперечить вибірковим даним) на рівні значущості α , 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 χ^2 з $(k-1)(l-1)$ ступенями вільності. В іншому випадку гіпотезу H_0 потрібно відхилити як таку, що не узгоджується з наявними даними.

13.4 Виконання в пакеті STATISTICA

Отримано відповіді 100 студентів перших трьох курсів на запитання: “Чи вважаєте Ви, що куріння заважає навчанню?”

Чи підтверджують ці дані припущення про те, що відношення до куріння у студентів різних курсів різне? Вибрати рівень значущості 0,01.

Для знаходження відповіді використаємо модуль *Correspondence Analysis*. Задамо таблицю даних, увівши частоти, отримані за результатами опитування:

Відповідь	Курс		
	I	II	III
Ні	15	10	0
Не знаю	8	5	7
Так	0	30	25

Вибравши метод *Correspondence Analysis*, тип даних *Frequencies w/out grouping vars*, змінні *Variables with frequencies: All* (рис. 13.4), одержимо (після натискування ОК) результати перевірки гіпотези H_0 (відношення до куріння не залежить від курсу) та можливість отримати деякі додаткові результати (рис. 13.5). Результат (Total chi-square = 44,24; $df = 4$; $p = 0,000$) свідчить про те, що гіпотеза H_0 має бути відхилена.

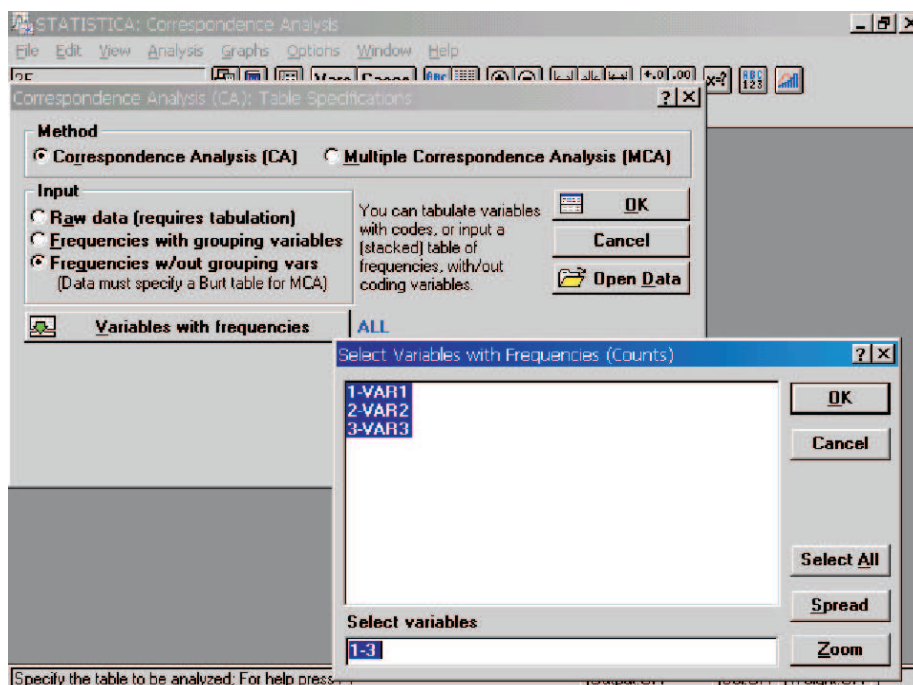


Рис. 13.4. Вибір змінних

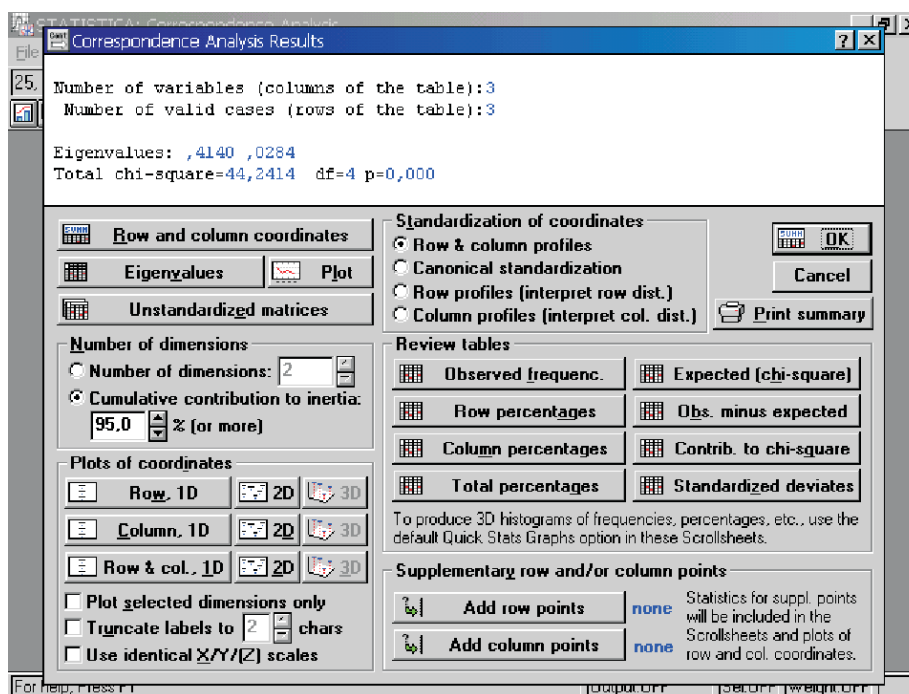


Рис. 13.5. Вікно результатів

13.5 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} разів, а в другій – n_{21} разів. 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 2×2 : Тут

Серія	Подія		Сума
	A	\bar{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 за умови, що найменша з величин $\frac{n_{\cdot i} n_{\cdot j}}{n}$, $i, j = 1, 2$ буде більшою,

ніж 5, як статистику для перевірки гіпотези H_0 можна використати

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

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

$$\tilde{\sigma}_{h_1-h_2}^2 = h(1-h) \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

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

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

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

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

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

$$\chi_B^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_1.n_2.n_{.1}n_{.2}}.$$

Критичну область на рівні значущості α визначають нерівністю $\chi_B^2 \geq \chi_{1-\alpha}^2(1)$, де $\chi_{1-\alpha}^2(1)$ – квантиль порядку $1 - \alpha$ розподілу χ^2 з одним ступенем вільності.

Критерій χ^2 можна використовувати за умови, що всі значення $\frac{n_{.i}n_{.j}}{n} > 3$ і $n > 20$. Для малих n при обчисленні χ_B^2 потрібно n замінити на $n - 1$; при цьому повинно бути $n_1. > 5$ $n_2. > \frac{n_1.}{3}$.

13.6 Виконання в пакеті STATISTICA

Під час епідемії грипу вивчався вплив щеплень проти цієї хвороби. Отримали такі результати:

	Після щеплення	Без щеплення
Захворіли	4	34
Не захворіли	192	111

Чи вказують ці результати на ефективність щеплень?

Застосуємо процедуру 2×2 Tables XI/VI/Phil, McNemar, Fisher exact модуля *Nonparametric Statistics*. Завантажимо пакет та виберемо відповідну процедуру (рис. 13.6). У вікні, що відкриється (рис. 13.7), введемо значення таблиці спряженості ознак. Натиснувши ОК, отримуємо результат (рис. 13.8).

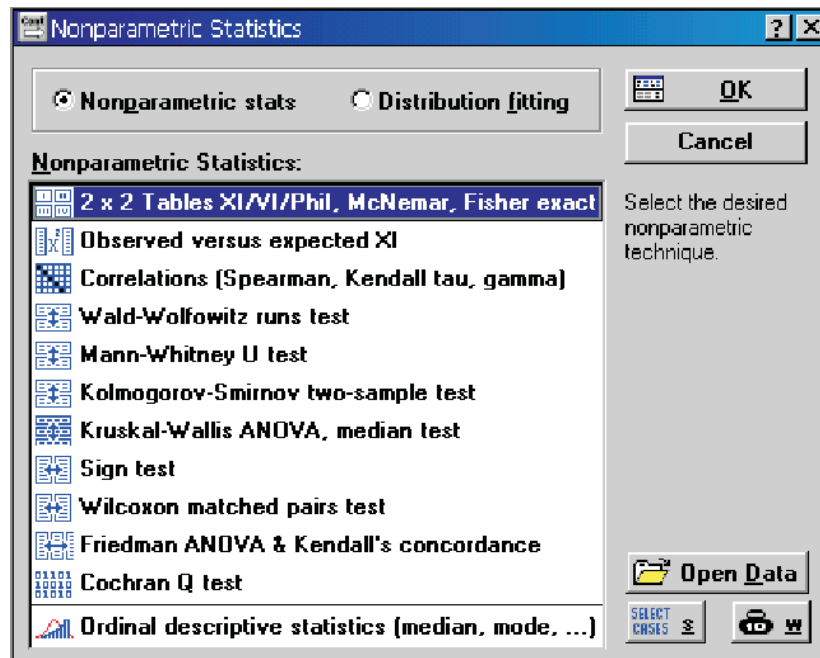


Рис. 13.6. Вікно *Nonparametric Statistics*

Бачимо, що гіпотезу про відсутність впливу щеплень на рівень захворюваності потрібно відхилити.

Крім стандартного критерію χ^2 Пірсона (*Chi-square*) і скоректованого χ^2 (*V-square*), STATISTICA обчислює: χ^2 з поправкою Йетса (*Yates corrected Chi-square*) для випадку малих значень частот; статистику F^2 (*F-square*), що є мірою зв'язку між номінальними чи категоріальними змінними, значення яких неможливо впорядкувати; рівні значущості одностороннього (*one-tailed*) та двостороннього (*two-tailed*) критерію

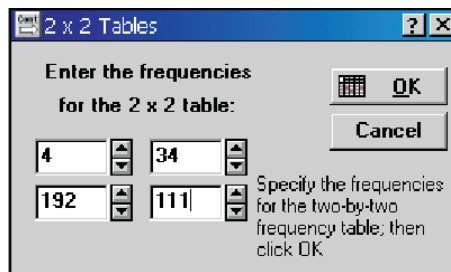


Рис. 13.7. Таблиця спряженості ознак

Continue...	Column 1	Column 2	Row Totals
Frequencies, row 1	4	34	38
Percent of total	1.173%	9.971%	11.144%
Frequencies, row 2	192	111	303
Percent of total	56.305%	32.551%	88.856%
Column totals	196	145	341
Percent of total	57.478%	42.522%	
Chi-square (df=1)	38.57	p= .0000	
V-square (df=1)	38.46	p= .0000	
Yates corrected Chi-square	36.44	p= .0000	
Phi-square	.11312		
Fisher exact p, one-tailed		p= .0000	
two-tailed		p= .0000	
McNemar Chi-square (A/D)	97.70	p= .0000	
Chi-square (B/C)	109.07	p= .0000	

Рис. 13.8. Результати аналізу

Фішера (*Fisher exact*); значення та рівні значущості критерію χ^2 Макнемара (*McNemar Chi-square*), який застосовують для дослідження залежних вибірок. Якщо сума частот невелика, то краще використовувати критерій Фішера замість χ^2 .

Розділ 14

Канонічний аналіз

14.1 Загальні положення

Метод канонічної кореляції призначений для аналізу залежностей між двома наборами змінних. Якщо обчислення попарних кореляцій між змінними дозволяє встановити залежності між окремими парами змінних, то 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 змінними. Частку обчислюють відносно дисперсії канонічних змінних, тобто зважених сум за двома наборами змінних. Знаходять стільки власних значень, скільки є змінних у найменшому з двох наборів даних.

Послідовне обчислення власних значень. У результаті виконання процедури послідовно обчислюються власні значення. Спочатку обчислюють ваги, які максимізують кореляцію між зваженими сумами по двох множинах змінних, і знаходять 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 значущості проводиться саме так.

Канонічні ваги. Після встановлення значущості 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 більша 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 використовуватися 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 значною мірою “пояснюються” нею. Таке пояснення канонічних змінних схоже на метод факторного аналізу.

Факторна структура та канонічні ваги. Канонічні значення відповідають унікальному внескові 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 за 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 виявитися 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. За великих обсягів вибірки невеликі канонічні кореляції (наприклад, 0,3) можуть виявитися статистично значущими. Для обчислення надлишковості цей коефіцієнт підносять до квадрату. 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 низку припущень, врахування яких важливе для отримання достовірних результатів.

Застосування критеріїв для перевірки значущості 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 обсяги вибірки (наприклад, $n = 50$) дозволяють у більшості випадків виявити ці

кореляції.

Викиди. Наявність викидів може мати значний вплив на величину коефіцієнтів кореляції. При збільшенні обсягу вибірки вплив невеликої кількості викидів нівелюється. Рекомендують перед проведенням процедури виявити значні викиди, наприклад, на діаграмі розсіювання.

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 кореляційні 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.

14.2 Виконання в пакеті STATISTICA

Розглянемо приклад дослідження залежності між обсягами валового збору пшениці, цукрового буряка та овочів (перша група показників) 15 господарств України та базовими показниками виробництва (друга група показників).

Введемо такі змінні:

Y_1 – валовий збір пшениці (тис. т.);

Y_2 – валовий збір цукрового буряка (тис. т.);

Y_3 – валовий збір овочів (тис. т.);

X_1 – загальна площа посівів (га);

X_2 – кількість внесених добрив (т.);

X_3 – фондозабезпеченість працівників (тис. т.);

X_4 – механізованість праці (млн. квт-год./чол.)

Попередньо проведений аналіз показників дозволив висунути гіпотезу про те, що ознаки, Y_1 , Y_2 , Y_3 є результативними, залежними, а ознаки X_1 , X_2 , X_3 , X_4 є факторними, незалежними. При цьому не заперечується наявність і зворотних зв'язків між групами змінних.

Будемо розглядати такі задачі:

- оцінити величину канонічної кореляції між першою та другою групами показників;
- перевірити статистичну значущість;

- виявити внутрішні латентні властивості досліджуваних господарств (канонічні змінні), дати їм економічне тлумачення та кількісну оцінку;
- виявити можливі практичні напрямки використання знайдених кількісних оцінок канонічних змінних для кожного об'єкта.

Для аналізу використано дані, наведені у таблиці:

№	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4
1	47,1	717,4	12,386	3317	308,5	44	0,85
2	66,6	1124	12,833	4167	304,4	46	1,30
3	38,6	593,7	9,786	2410	244,2	28	1,50
4	87,0	1854,5	15,266	5332	378,7	48	1,60
5	53,8	971	12,577	3763	290,3	37	1,10
6	106,0	1187,9	14,282	4094	378,7	54	0,02
7	74,2	870,0	15,048	4892	357,9	43	0,07
8	47,0	1024,0	15,171	3031	381,2	37	1,08
9	56,6	1009,1	16,190	3596	366,7	50	5,02
10	62,8	859,2	11,654	4356	304,8	44	0,14
11	90,2	1285,8	20,871	5750	458,9	53	4,70
12	48,7	608,9	9,969	3137	237,6	38	0,45
13	24,1	487,3	7,000	1477	181,5	44	0,10
14	66,7	908,6	15,820	4335	364,4	48	0,18
15	40,0	710,0	9,039	2720	205,7	51	2,00

Після переходу до модуля “Канонічний аналіз” (*Canonical Analysis*) потрібно створити новий файл даних, активізувати на панелі клавішу “Аналіз” (*Analysis*) і відкрити стартову панель, яка показана на рис.14.1.

Зазвичай, роботу в модулі починають з вибору змінних для дослідження за допомогою клавіші “Змінні” (*Variables*), яка відкриває список усіх ознак створеного файлу даних. Оберемо для канонічного аналізу усі 7 змінних, натиснувши на клавішу “Обрати усе” (*Select All*).

У вікні “Файл, що вводиться” (*Input file*) потрібно відмітити “Вхідні дані” (*Raw Date*), а у вікні “Видалення пропущених даних” (*MD deletion*) – “Пропущені випадки” (*Casewise*).

Якщо файл, що вводиться, потрібно задати у вигляді кореляційної матриці (*Correlation Matrix*), то її необхідно попередньо створити у якому-небудь модулі системи, наприклад, “Множинна регресія” (*Multiple Regression*), зберегти і потім відкрити в модулі “Канонічний аналіз”.

При активізації опції “Перегляд описових статистик і кореляційної матриці” (*Review descriptive stats and correlation matrix*) натисне-

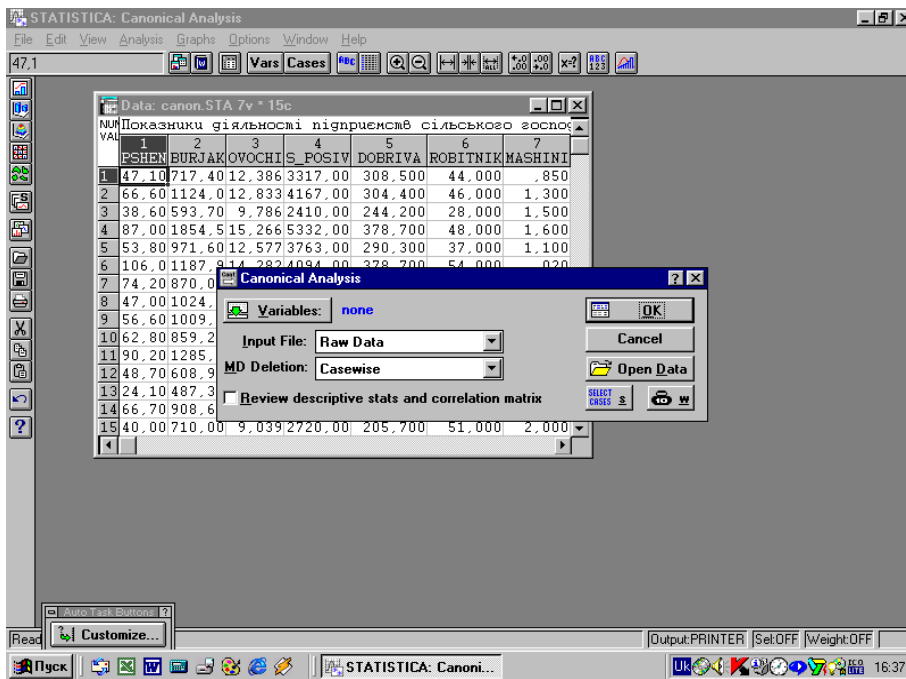


Рис. 14.1. Стартова панель модуля “Канонічний аналіз”

ння клавіші *OK* викликає появу вікна, за допомогою якого виводяться середні значення та значення стандартних відхилень усіх змінних, що вивчаються, матриця коефіцієнтів парної кореляції між ними, матриця дисперсій – коваріацій, різні графіки і т.д. (рис.14.2.)

Натиснення на клавішу *OK* викликає появу панелі “Вибір моделі” (*Model Definition*), за допомоги якої необхідно вказати перелік змінних у першій і другій групах *Variables for canonical analysis* (рис. 14.3).

У першому списку вкажемо *VAR1 – VAR3*, які відповідають змінним $Y_1 – Y_3$, а у другому – *VAR4 – VAR7*, які відповідають ознакам $X_1 – X_4$. Натиснення на клавішу *OK* викликає появу вікна результатів канонічного аналізу, яке складається з двох частин: інформаційної та функціональної (рис. 14.4).

У верхній інформаційній частині містяться дані про величину першого канонічного коефіцієнта кореляції, розрахункове значення статистики χ^2 для r_1 , кількість ступенів вільності і p – значущість, кількість повноцінних спостережень і змінних, відсоток виділеної дисперсії і залишку в кожній групі. Іншими словами, тут вказуються загальні результати канонічного аналізу.

У нижній, функціональній, частині вікна розміщені опції та клавіші, які необхідні для поглибленого перегляду всіх найважливіших резуль-

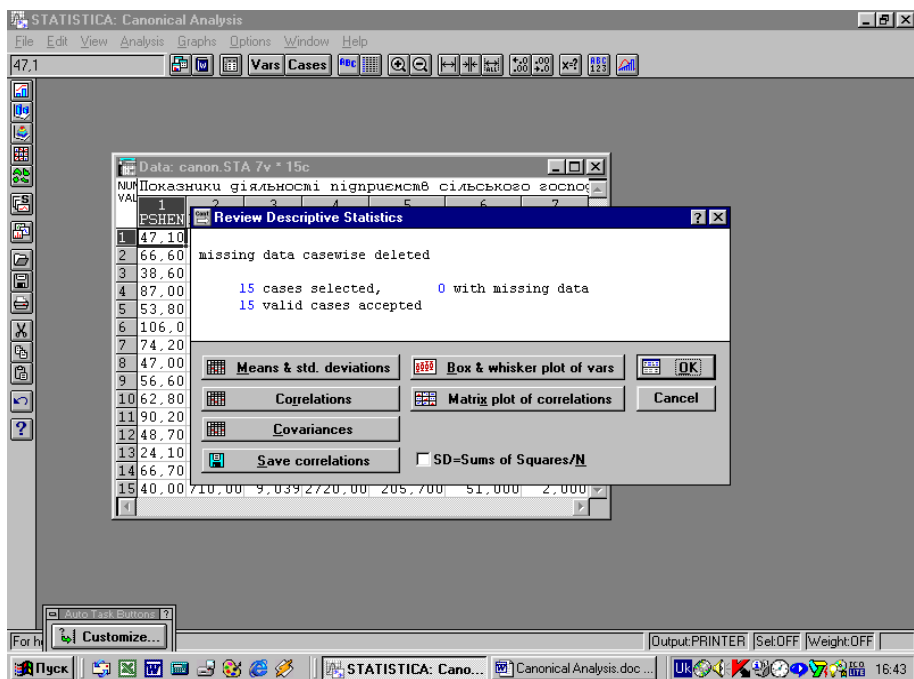


Рис. 14.2. Вікно перегляду описових статистик

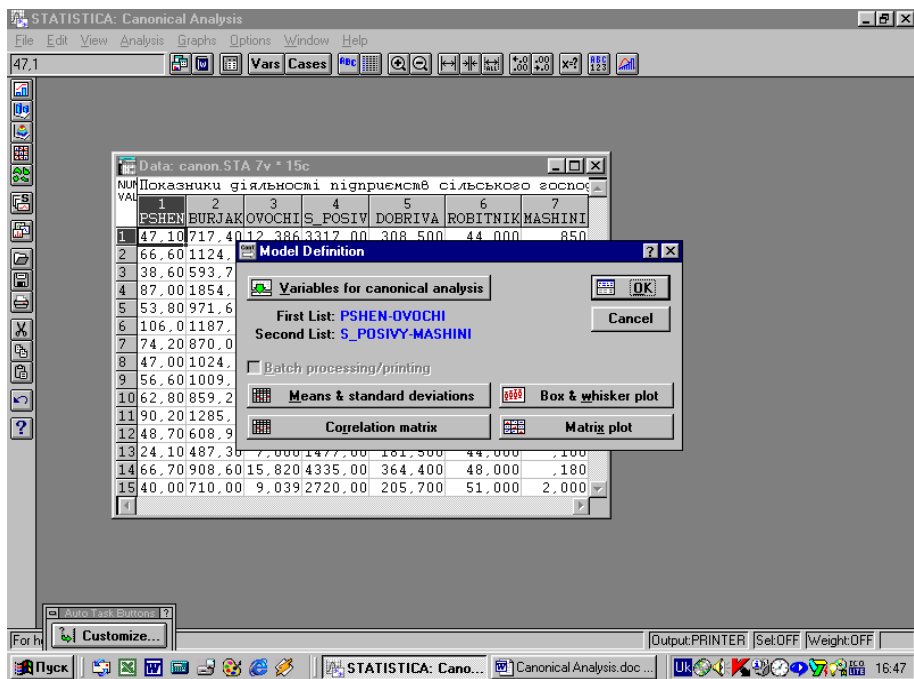


Рис. 14.3. Панель вибору канонічної моделі

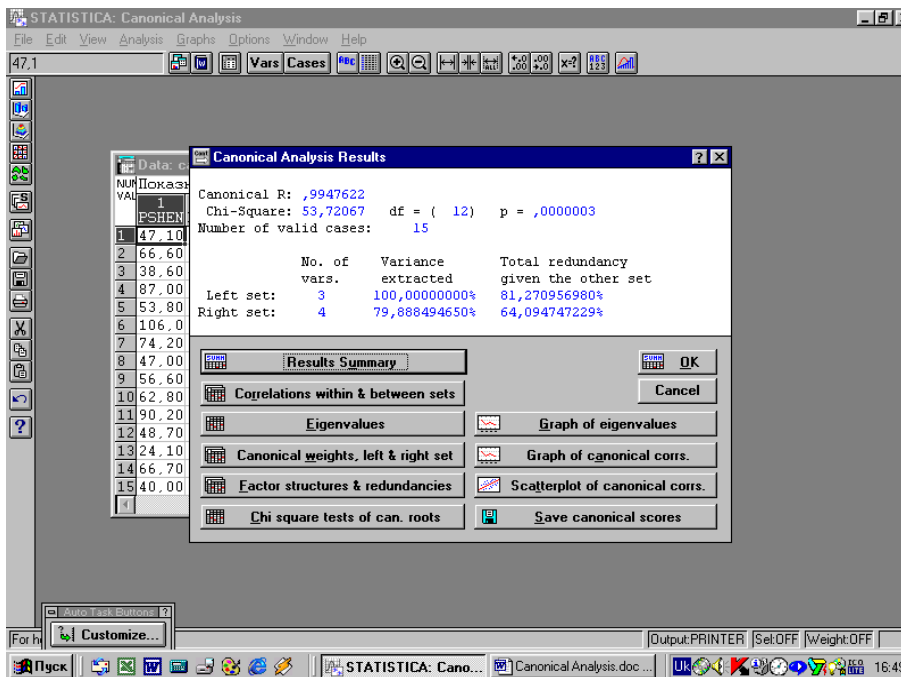


Рис. 14.4. Вікно результатів канонічного аналізу

татів канонічного аналізу.

Так, клавіша “Підсумкові результати” (*Results Summary*) дозволяє вивести на екран таблицю загальних підсумків дослідження (рис. 14.5), які стосуються r_1 , y_1 , x_1 , що наводилися вище в інформаційній частині вікна результатів канонічного аналізу.

Аналіз даних, який наведений у таблиці на рис. 14.5, показує, що в результаті проведеного канонічного аналізу загальна частина дисперсії ознак першої малої групи ($Y_1 - Y_3$), яка була виділена за допомогою першої канонічної змінної y_1 , становить 100%. А загальна частина дисперсії ознак другої великої групи ($X_1 - X_4$), виділеної канонічної змінної x_1 , становить 79,9%.

Загальний збиток для змінних першої групи дорівнює 81,3%, а загальний збиток для змінних другої групи – 64,1%. Це означає, що 81,3% варіації валового збору пшениці, цукрового буряка та овочів визначається змінами чотирьох агротехнічних факторів виробництва $X_1 - X_4$ на виробництвах, що вивчають. Водночас самі обсяги валових зборів вказаних культур детермінують 64,1% варіації ресурсних та агротехнічних можливостей сільськогосподарських об’єктів.

Наведені результати свідчать про достатньо високу точність побудованої канонічної моделі: менше 19% дисперсії змінних $Y_1 - Y_3$ залежить

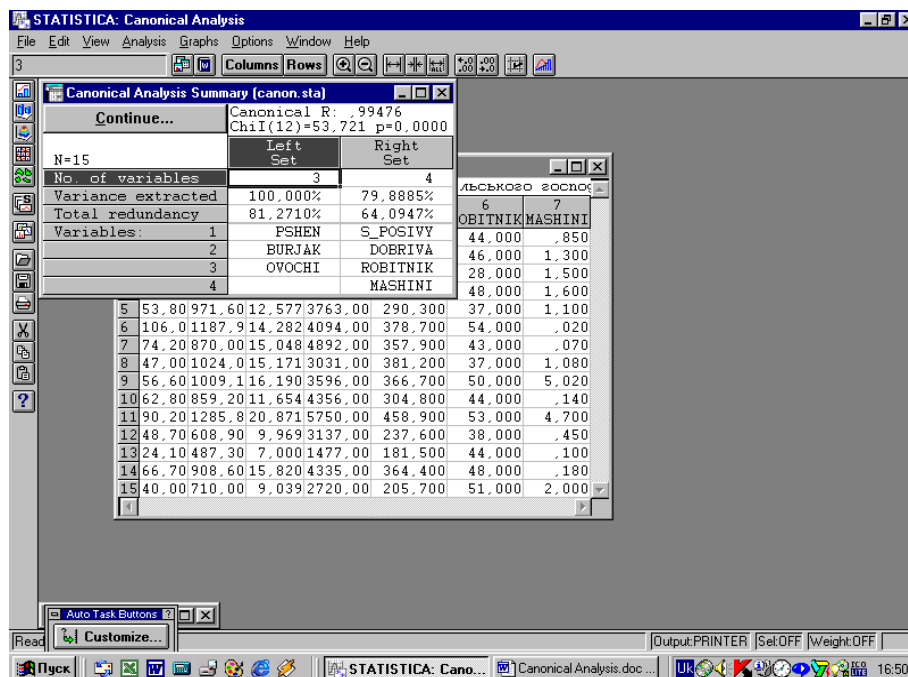


Рис. 14.5. Вікно загальних підсумків канонічного аналізу

від інших, не врахованих в аналізі факторів виробництва.

Клавіша “Кореляція всередині та між групами” (*Correlations within and between sets*) дозволяє вивести на екран кореляційні матриці r_y , r_x , r_{yx} , на базі яких потім будують матрицю r' (рис. 14.6.)

Так, аналіз останньої матриці показує, що результативні ознаки $Y_1 - Y_3$ тісно пов'язані з факторами X_1, X_2 ($r_{yx} > 0,7$) і слабше – із змінними X_3 та X_4 .

Натиснення на клавішу “Характеристичні корені” (*Eigenvalues*) відкриває для перегляду таблицю квадратів коефіцієнтів канонічної кореляції r_1^2, r_2^2, r_3^2 (рис.14.7.) Зауважимо, що кількість характеристичних коренів дорівнює $\min(s, m) = \min(3, 4) = 3$.

Клавіша “Канонічні ваги, ліва та права групи” (*Canonical weights, Left and right set*) дозволяє вивести на екран три пари канонічних змінних y_1 і x_1, y_2 і x_2, y_3 і x_3 (стовпці таблиць на рис. 14.8). У подальшому будемо аналізувати лише першу пару канонічних змінних, яка має найбільш тісний взаємозв'язок: $r_1 = 0,99476$.

Аналіз даних таблиці показує, що всі статистичні ваги перших канонічних змінних додатні й знаходяться в межах від $0,0377$ до $0,8100$.

Мінімальне значення статистичної ваги відповідає Z_{x3} ($\beta_3 = 0,037741$), тому фактор X_3 (VAR6) дає найменший внесок у пояснення варіації ре-

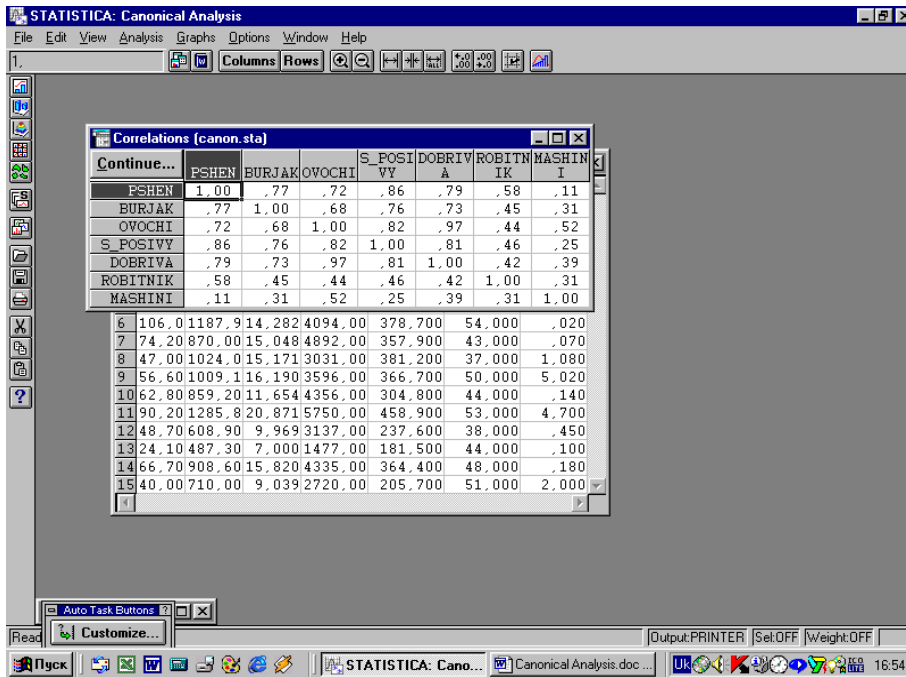


Рис. 14.6. Кореляційна матриці

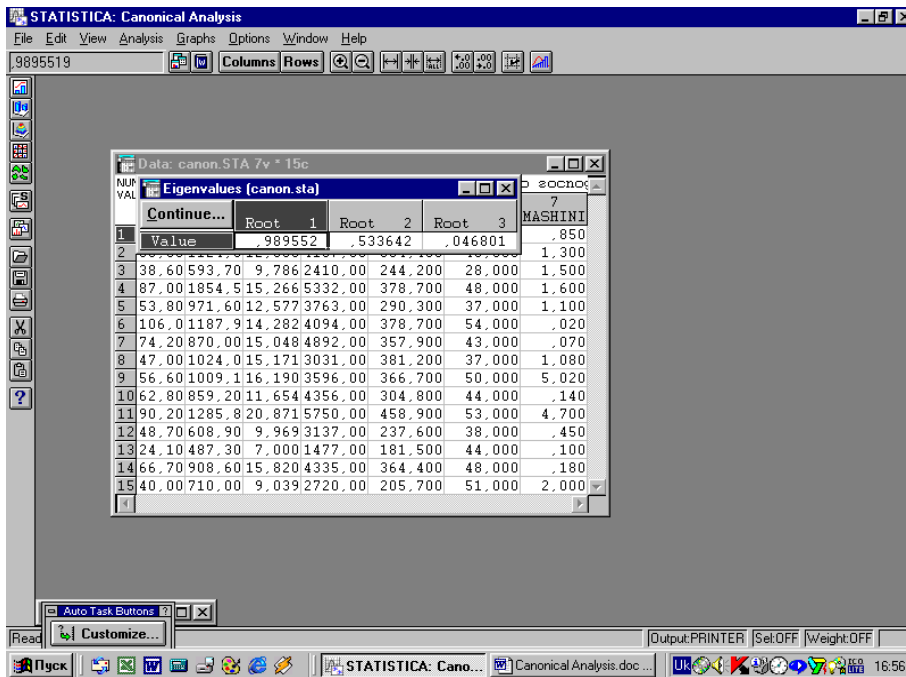


Рис. 14.7. Вікно характеристичних коренів

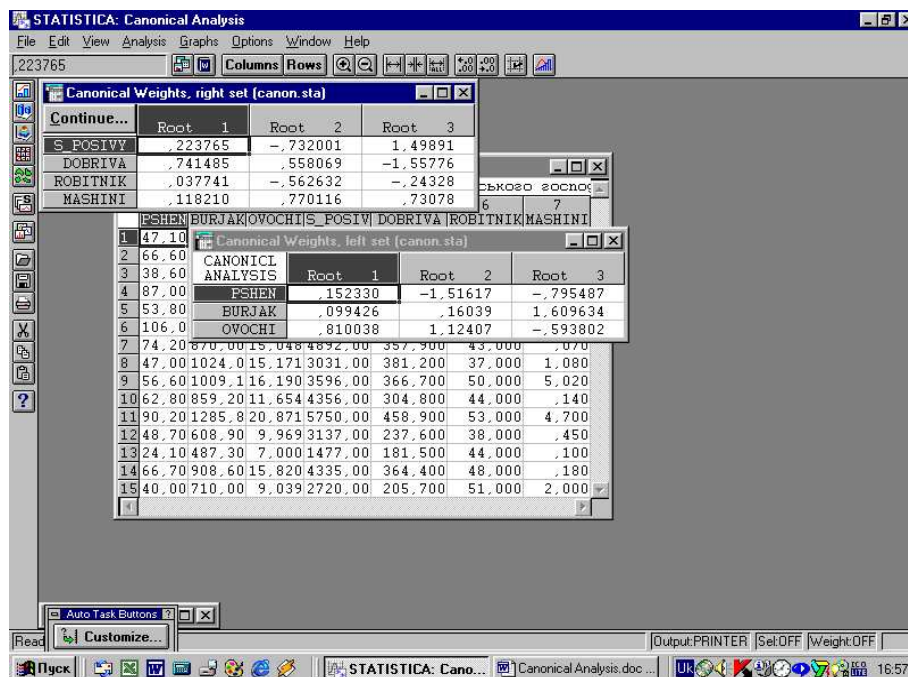


Рис. 14.8. Канонічні змінні двох груп ознак

зультативних ознак першої групи і є першим кандидатом на виключення з подальшого аналізу. Застосуємо до нього алгоритм послідовного відсіву змінних з відкиданням найменш значущих (з погляду внеску у величину r_1) ознак. Для цього знову звернемося до системи STATISTICA і розрахуємо нову канонічну модель, яка не має фактора X_3 (рис. 14.9).

У верхній інформаційній частині вікна на рис. 14.9. знаходимо $r_{t-1} = 0,9943498$ ($r_1 = 0,9947622$ – перший канонічний коефіцієнт кореляції до виключення X_3). Тому перетворення Фішера мають такий вигляд:

$$F_t = 0,5 \ln[(1 + 0,9947622)/(1 - 0,9947622)] = 2,97118928;$$

$$F_{t-1} = 0,5 \ln[(1 + 0,9943498)/(1 - 0,9943498)] = 2,93319121.$$

Вони використовуються для знаходження статистики:

$$N = (2,97118928 - 2,93319121)[(15 - 3)/2]^{1/2} = 0,09307589.$$

Перевіряємо нульову гіпотезу $H_0 : r_t = r_{t+1}$. Оскільки $p > \alpha$ ($0,463 > 0,05$), то розрахункове значення p потрапляє в допустиму область і нульову гіпотезу не відхиляємо. Це значить, що виключення із дослідження змінної X_3 привело до несуттєвої зміни першого канонічного коефіцієнта r_1 .

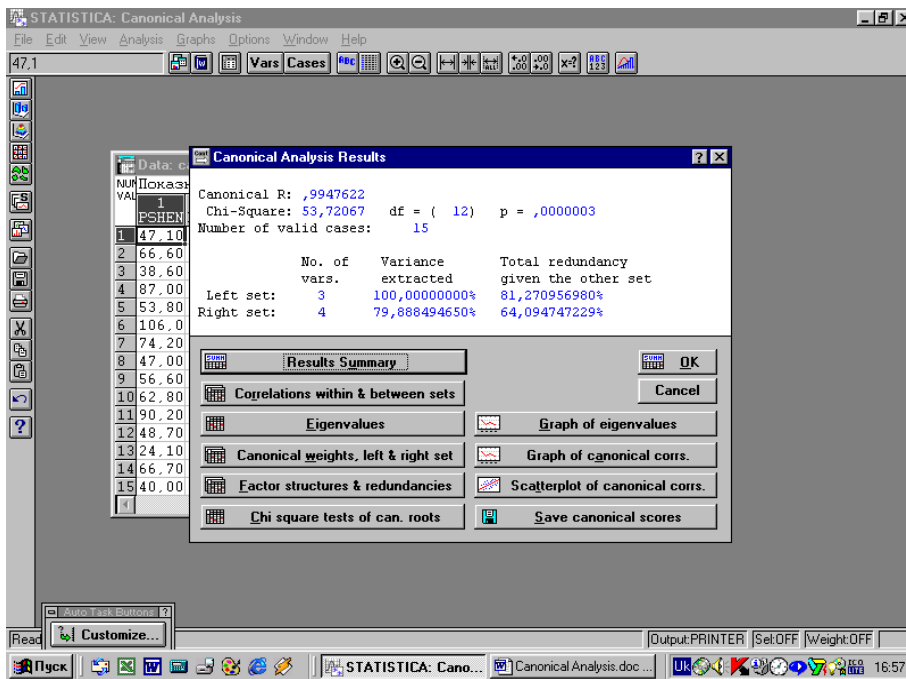


Рис. 14.9. Результати після виключення змінної X_3

Знову розглянемо статистичні ваги перших канонічних змінних після виключення X_3 (рис. 14.10).

Аналіз даних таблиці на рис. 14.10 показує, що всі статистичні ваги перших канонічних змінних додатні і достатньо великі за абсолютною величиною. Тому подальший покроковий відсів несуттєвих ознак вирішено було припинити і для подальшого аналізу використовувати канонічну модель з трьома змінними у кожній групі: $Y_1 - Y_3$, X_1 , X_2 , X_4 .

Клавіша “Факторні структури та втрати” (*Factor structures and redundancies*) вікна результатів канонічного аналізу дозволяє передивитися факторні навантаження для кожної групи ознак (рис.14.11), а також виділену варіацію (пропорції) і відповідні втрати, знайдені на їх основі для кожної канонічної змінної (рис. 14.12).

Підсумкові значення виділеної варіації (пропорції) і втрат за першою канонічною змінною наведені у верхній інформаційній частині вікна результатів канонічного аналізу на рис. 14.5, а також можуть бути виведені на екран за допомогою клавіші “Підсумкові результати” (*Results Summary*).

Порівняння результатів канонічного аналізу показує, що після відсіву змінної X_3 відбулось незначне зниження першого канонічного коефіцієнта кореляції ($r_t - r_{t+1} = 0,9947622 - 0,9943498 = 0,0004124$), а також

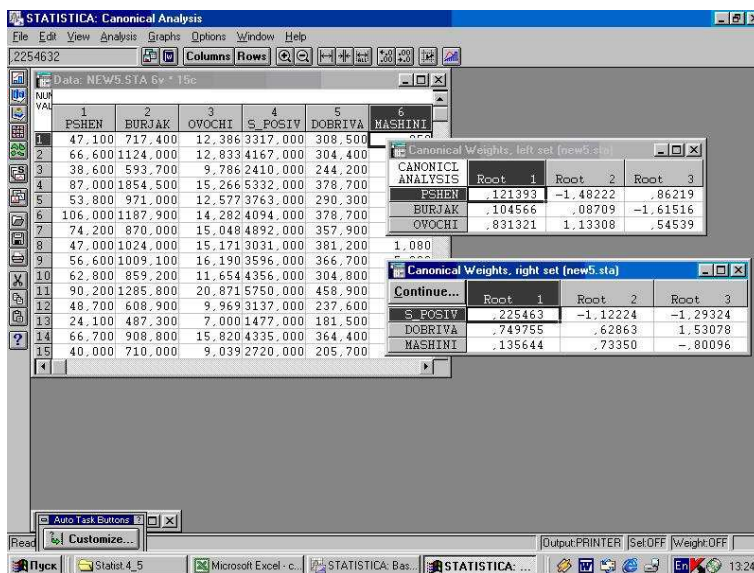


Рис. 14.10. Канонічні змінні після виключення X_3

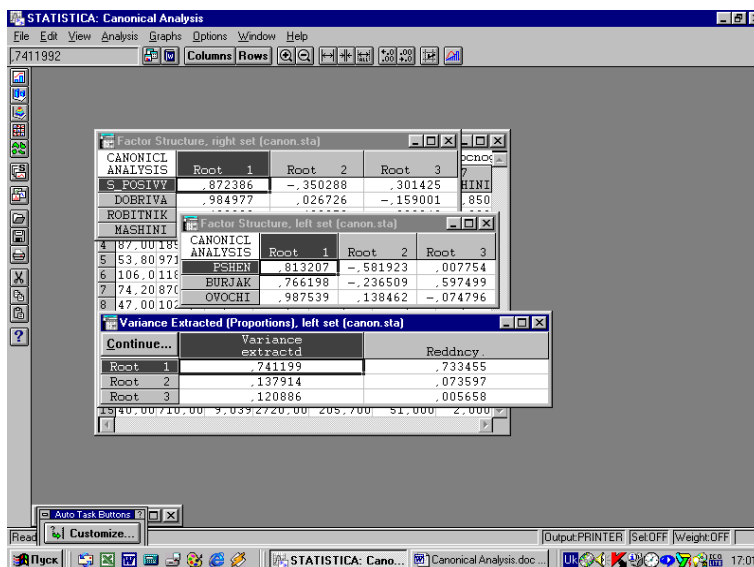


Рис. 14.11. Виділена варіація та втрати

деяке зменшення загальної втрати для змінних першої групи (з 81,3% до 79,2%). Водночас відбувся зріст загальної втрати для змінних другої групи (з 64,1% до 74,3%), тобто більше, ніж на 10 відсоткових пунктів.

Це означає, що 79,2% варіації валового збору пшениці, цукрового буряка та овочів визначається зміною трьох агротехнічних факторів виробництва (X_1, X_2, X_4). Водночас самі обсяги валових зборів указаних культур детермінують 74,3% варіації ресурсних та агротехнічних можливостей об'єктів сільського господарства.

Наведені результати свідчать про достатньо високу точність канонічної моделі, побудованої на другому кроці (після виключення X_3): менше 21% дисперсії змінних ($Y_1 - Y_3$) залежить від інших, неврахованих в аналізі факторів виробництва.

Клавіша “ χ^2 критерії для канонічних коренів” (*Chi square tests for can. roots*) у вікні результатів канонічного аналізу дає можливість отримати розрахункові значення статистики χ^2 .

На рис. 14.12 наведено таблицю, в якій у першому стовпці вказані канонічні коефіцієнти кореляції r_1, r_2, r_3 , у другому – їх квадрати, тобто характеристичні корені $\lambda_1, \lambda_2, \lambda_3$ матриці, а в третьому – розрахункові значення χ^2 статистики для послідовно виділених канонічних коренів, у четвертому – кількість ступенів вільності статистики χ^2 , у п'ятому – p -значущість χ^2 критерію, у шостому – значення Λ' -статистики Уїлкса.

Скористаємося наведеними даними для перевірки статистичної значущості для канонічних коефіцієнтів кореляції r_1, r_2, r_3 . Автоматичний розрахунок p – значущості розрахункових величин χ^2 у STATISTICA дозволяє використати сучасну схему процедури перевірки гіпотез: достатньо порівняти p з допустимим рівнем значущості (зазвичай $\alpha = 0,05$).

Оскільки для $r_1 : p \leq \alpha$ ($0 < 0,05$), а для $r_2, r_3 : p > \alpha$ ($0,492 > 0,200 > 0,05$), то можна зробити висновок про те, що для першого канонічного коефіцієнта кореляції нульову гіпотезу $H_0 : r_1 = 0$ відхиляємо практично із 100% надійністю і справедлива альтернатива $H_a : r_1 > 0$. Для другого і третього канонічних коефіцієнтів кореляції нульову гіпотезу $H_0 : r_1 = 0$ не відхиляємо і можна стверджувати, що вони статистично незначущі.

Навпаки, перший канонічний коефіцієнт кореляції $r_1 = 0,99435$ є значущим, а відповідні йому канонічні змінні:

$$Z_y = 0,121419Z_{y1} + 0,104535Z_{y2} + 0,831331Z_{y3},$$

$$Z_x = 0,225475Z_{x1} + 0,749734Z_{x2} + 0,135645Z_{x3}$$

є статистично значущими.

Останні являють собою, як було зазначено вище, латентні показники, якісне тлумачення яких здійснюють аналогічно поясненню виділених го-

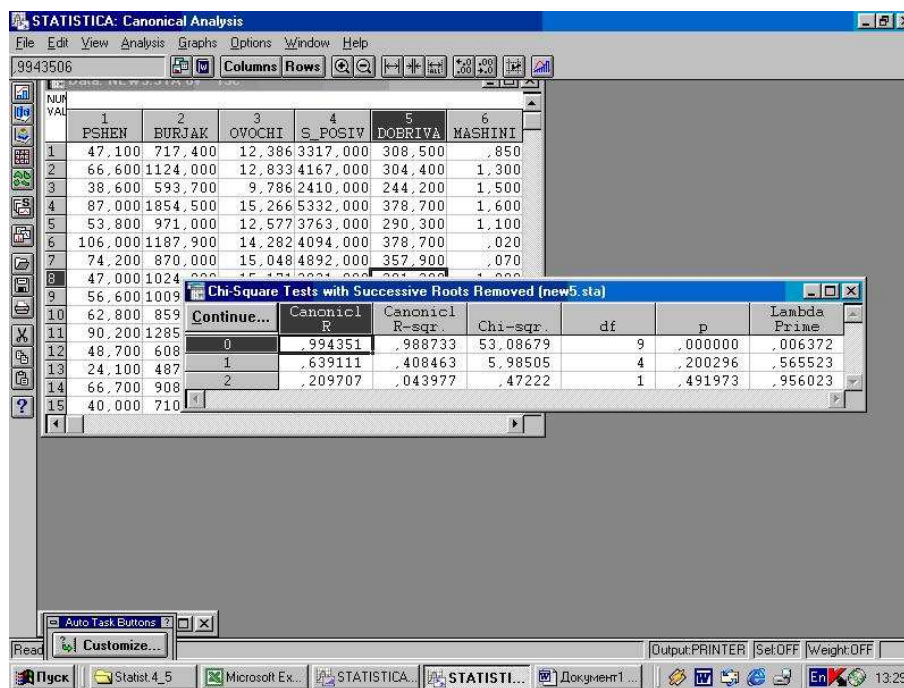


Рис. 14.12. χ^2 критерії для послідовно виділених канонічних коренів

ловних компонент або загальних факторів. Тому стандартизована змінна Z_y , що включає валові збори пшениці, цукрового буряка та овочів, може розглядатись як показник “валового збору сільськогосподарських культур”. Він дає загальну характеристику річних результатів виробничої діяльності господарств, які вивчають.

Канонічна змінна Z_x , яку “навантажують” ознаки X_1 – загальна площа посіву (га), X_2 – кількість внесених добрив (т), X_4 – машиноозброєність праці (млн. кВт-год/люд), є загальним показником ресурсно-агротехнічних факторів виробництва, які використовували господарства, які вивчають у звітному році для отримання валового збору сільськогосподарських культур.

Підставляючи в наведені вище рівняння для Z_y , Z_x конкретні значення стандартизованих ознак-симптомів, дуже легко отримати кількісні оцінки латентних показників “валовий збір сільськогосподарських культур” і “ресурсно-агротехнічний фактор виробництва” для кожного окремого господарства. Ці стандартизовані оцінки можна використовувати для ранжування досліджуваних об’єктів за величиною виявлених латентних показників, для виділення груп лідерів, посередніх та аутсайдерів.

Розділ 15

Дискримінантний аналіз

15.1 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, передбачають отримання 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 ма-

ють вимірюватися за інтервальною шкалою, чи шкалою відношень. Як правило, кількість об'єктів переважає кількість дискримінантних змінних хоча б на два.

Обмеження для дискримінантних змінних

1. Жодна змінна не має бути лінійною комбінацією інших змінних.
2. Недопустимо, щоб для будь-якої пари змінних коефіцієнт кореляції дорівнював 1 чи -1 .
3. Часто використовують припущення про збіг коваріаційних матриць різних груп.
4. Закон розподілу для кожної групи багатовимірний, нормальний (це дозволяє отримати значення ймовірностей належності до груп).

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 віднесенень. Тоді сумарні втрати при класифікації n об'єктів та k класах дорівнюють

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

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

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

(використовуємо збіжність за ймовірністю).

Тут $n_i(n)$ – частота i -го класу; π_i – ймовірність вибору об'єкта i -го класу із загальної сукупності (так звана апіорна 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 $1 - \sum_{i=1}^k \pi_i p(j|i)$ визначає ймовірність неправильної класифікації.

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, $\overline{X}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)})$ – це i -тий об'єкт, який потрібно класифікувати.

z_1, z_2, \dots, 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 зі щільністю ймовірності $f(x) = \sum_{j=1}^k \pi_j f_j(x)$, де

π_j – апіорна 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 $1, 2, \dots, k$, 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 $S = (s_1, s_2, \dots, s_k)$. Отже, і вирішальна процедура може бути задана розбиттям.

Процедуру класифікації (дискримінантну функцію) 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:

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

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

тобто спостереження X_r ($r=1, 2, \dots, m$) 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 π_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 $\hat{\pi}_j$ 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 як 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:

$$f_{ij} = u_0 + u_1 x_{1ij} + u_2 x_{2ij} + \dots + u_m x_{mij}, \quad (15.1)$$

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 $\min(k - 1, m)$.

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 вимірністю на одиницю меншою від кількості класів. Отже, задачу класифікації тепер можна розглядати в $(k - 1)$ -вимірному просторі, натягнутому на центроїди.

Початок координат поміщають у точку нульових значень показників. Першу вісь направляють так, щоб середні значення класів розділялись більшою мірою, 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 (15.1) задає перетворення m -вимірного простору дискримінантних змінних у q -вимірний простір канонічних дискримінантних функцій. Кожній осі відповідає своє співвідношення виду (15.1). Для даного спостереження f_{ij} тлумачать як координату об'єкта в просторі канонічних дискримінантних функцій.

У разі, коли кількість дискримінантних змінних m менша від кількості класів, максимальна кількість функцій q дорівнює m . Тоді вже не відбувається перетворення з простору з більшою вимірністю в простір з меншою вимірністю, проводиться тільки заміна координат.

15.2 Параметричний дискримінантний аналіз (випадок нормального розподілу класів)

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

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

$$\hat{a}_j^l = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}^{(l)}, \quad l = \overline{1, m}, \quad j = \overline{1, k}; \quad (15.2)$$

$$\hat{\sigma}_{lp} = \frac{1}{n-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ji}^{(l)} - \hat{a}_j^{(l)})(x_{ji}^{(p)} - \hat{a}_j^{(p)}), \quad (15.3)$$

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

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

$$\left[X - \frac{1}{2}(\hat{a}_{j0} + \hat{a}_j) \right]^T \sum_{j=1}^{\wedge-1} (\hat{a}_{j0} - \hat{a}_j) \geq \ln \frac{\pi_j}{\pi_{j0}} \quad (15.4)$$

для всіх $j = \overline{1, k}$.

Правило, задане співвідношенням (15.4), має місце для випадку однакових значень втрат.

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

$$\left[X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) \right]^T \sum_{j=1}^{\wedge-1} (\hat{a}_1 - \hat{a}_2) \geq 0, \quad (15.5)$$

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

$$\left(X - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) \right) (\hat{a}_1 - \hat{a}_2) \geq 0. \quad (15.6)$$

Приклад.[1] Спеціальне дослідження показало, що схильність фірм до приховування частини своїх доходів (і, відповідно, до ухилю від сплати частини податків) по суті визначається двома показниками:

$x^{(1)}$ – співвідношення “швидких активів” і поточних пасивів;

$x^{(2)}$ – співвідношення прибутку і прострочених платежів (обидва показники оцінюють за певною методикою за шкалою від 300 до 900 балів).

У таблиці наведені значення цих показників (дані податкової інспекції) по 10 фірмах, які в тій чи іншій формі не сплачували податки ($n_1 = 10$), і по 13 фірмах, які стабільно сплачували податки ($n_2 = 13$):

№	Навчаюча вибірка (фірми, які ухиляються від сплати податків)		Навчаюча вибірка (фірми, які не ухиляються від сплати податків)	
	$x_{1i}^{(1)}$	$x_{1i}^{(2)}$	$x_{1i}^{(1)}$	$x_{1i}^{(2)}$
1	740	680	750	590
2	670	600	360	600
3	560	550	720	750
4	540	520	540	710
5	590	540	570	700
6	590	700	520	670
7	470	600	590	790
8	560	540	670	700
9	540	630	620	730
10	500	600	690	840
11	–	–	610	680
12	–	–	550	730
13	–	–	590	750
середнє	$\hat{a}_1^{(1)} = 576,0$	$\hat{a}_1^{(2)} = 596,0$	$\hat{a}_2^{(1)} = 598,5$	$\hat{a}_2^{(2)} = 710,8$

Крім того, статистика і спеціальні дослідження говорять про те, що відсоток фірм, які в тій чи іншій формі не сплачують податки, становить 50% (тобто $\pi_1 = \pi_2 = 0,5$).

Статистична перевірка гіпотез про нормальний характер розподілу двовимірної ознаки $X = (x^{(1)}, x^{(2)})^T$ всередині кожної сукупності фірм, які аналізують, і про рівність їх коваріаційних матриць $\Sigma_{(1)}$ і $\Sigma_{(2)}$ дала позитивний результат, тобто можна вважати, що вибірки (1) і (2), які ми маємо, взято з нормальних генеральних сукупностей з однаковими коваріаційними матрицями.

На фірмі, яка не пройшла перевірку податкової інспекції, зареєстровані такі значення змінних $X = (x^{(1)}, x^{(2)})^T$: $x_0^{(1)} = 740$, $x_0^{(2)} = 590$. Потрібно визначити, до якої сукупності (1 чи 2) можна віднести цю фірму (тобто спостереження $X_0 = (x_0^{(1)}, x_0^{(2)})^T$), використовуючи метод параметричного дискримінантного аналізу (з урахуванням нормальності спостережень, які аналізують).

З умови задачі випливає, що потрібно використати правило класифікації, яке визначене співвідношенням (15.4).

Необхідні обчислення дають:

1. Оцінки середніх значень по кожній навчаючій вибірці, підраховані

за формулою (15.2):

$$\hat{a}_1 = (576, 0; 596, 0)^T; \quad \hat{a}_2 = (598, 46; 710, 77)^T;$$

відповідно:

$$\hat{a}_1 - \hat{a}_2 = (-22, 46; -114, 77)^T; \quad (\hat{a}_1 + \hat{a}_2)/2 = (587, 23; 653, 39)^T.$$

2. Загальну коваріаційну матрицю $\hat{\Sigma}$ оцінюють за формулою (15.3) або, що те ж саме,

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^2 \sum_{i=1}^{n_j} (X_{ji} - \hat{a}_j)(X_{ji} - \hat{a}_j)^T = \frac{1}{10 + 13 - 2} \times$$

$$\left[\begin{pmatrix} 56240 & 17240 \\ 17240 & 33240 \end{pmatrix} + \begin{pmatrix} 121569 & 25615 \\ 25615 & 56492 \end{pmatrix} \right] = \begin{pmatrix} 8467 & 2041 \\ 2041 & 4273 \end{pmatrix}.$$

Отримуємо таку обернену матрицю:

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 0,0001335 & -0,0000637 \\ -0,0000637 & 0,0002645 \end{pmatrix}.$$

$$3. \quad \hat{\Sigma}^{-1}(\hat{a}_1 - \hat{a}_2) = \begin{pmatrix} 0,0043 \\ -0,0289 \end{pmatrix},$$

$$X_0 - \frac{1}{2}(\hat{a}_1 + \hat{a}_2) = \begin{pmatrix} 740 - 587, 23 \\ 590 - 653, 39 \end{pmatrix} = \begin{pmatrix} 152, 77 \\ -63, 39 \end{pmatrix}.$$

4. Оскільки числове значення виразу, який визначається співвідношенням (15.5), у нашому випадку дорівнює

$$\begin{pmatrix} 152, 77 \\ -63, 39 \end{pmatrix}^T \begin{pmatrix} 0,0043 \\ -0,0289 \end{pmatrix} = 2,489 > 0,$$

то спостереження X_0 має бути віднесене до класу 1, а це значить, що є підстави для того, щоб діагностувати фірму, яку аналізують, як фірму, що в тій чи іншій формі ухиляється від сплати податків.

Для класифікації нового об'єкту можна також використовувати класифікуючі функції (у цьому випадку алгоритм класифікації навіть простіший). Класифікуючі функції визначають для кожної з існуючих груп у вигляді лінійної функції від дискримінантних показників. Для розпізнавання нового об'єкту його координати підставляють у всі знайдені класифікуючі функції. Для якої з них отримують найбільше значення, до тієї групи і слід віднести новий об'єкт.

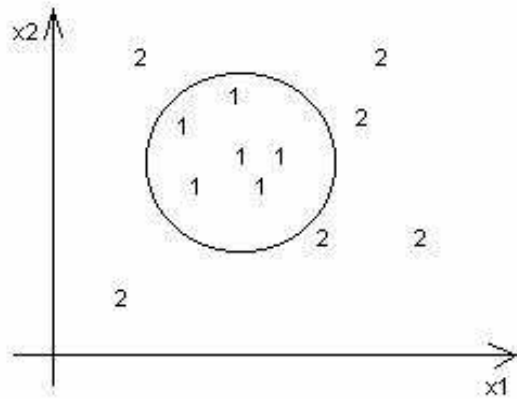


Рис. 15.1. Розташування навчальних вибірок

15.3 Нелінійний дискримінантний аналіз.

На практиці трапляються випадки, коли лінійні дискримінантні функції недостатні для проведення класифікації. Наприклад, у випадку двох дискримінантних ознак та двох груп, навчальні вибірки розташовані, як на рисунку 15.1.

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 як вирішальне правило таке: об'єкти, які містяться всередині кола, зараховують до першої групи, а об'єкти, які містяться зовні кола – до другої.

Описана ситуація досить часто спостерігається 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. 15.2.

Для цього обчислюють всі відстані від об'єкта X до елементів навчальних 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 ймовірність помилки через малу кількість відібраних елементів; а при надто великих k на класифікацію

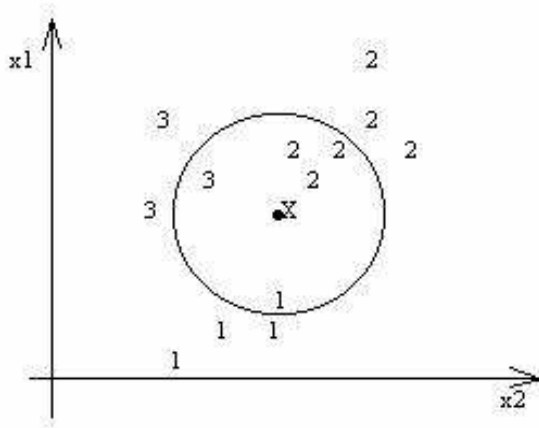


Рис. 15.2. Об'єкт X і елементи навчаючих вибірок

можуть впливати елементи, що розташовані дуже далеко і не мають ніякого відношення до цього спостереження. Рекомендована величина $k \approx \ln(n)$, де n – загальний обсяг навчаючих вибірок.

Як відстані між об'єктами використовують відстань $d = \sum_{j=1}^m |x_j^i - x_j^k|$, евклідову відстань $d(X^i, X^k) = \sqrt{\sum_{j=1}^m (x_j^i - x_j^k)^2}$ або відстань Махалано-біса $d = \sqrt{\sum_{j=1}^m \frac{(x_j^i - x_j^k)^2}{D_j}}$, де D_j – вибіркова дисперсія ознаки. Останню відстань використовують для врахування різниці у розподілі окремих показників (компонент).

15.4 Виконання в пакеті STATISTICA

Застосуємо модуль “Дискримінантний аналіз” до розглянутого вище прикладу.

Початок роботи

У вікні переключення модулів *STATISTICA Module Switcher* виберемо зі списку *Discriminant Analysis* і натиснемо на кнопку *Switch to*.

Вхідні дані у вигляді таблиці потрібно розмістити у відповідному файлі. Пункт меню *File / New Data* дозволяє його створити.

Якщо у створюваній таблиці має бути більше стовпців чи рядків, ніж пропонується, то створити відповідні елементи можна за допомогою

STATISTICA: Discriminant Analysis

File Edit View Analysis Graphs Options Window Help

740

Data: DISCR.STA 10v * 40c

NUR VAL	1 VAR1	2 VAR2	3 VAR3	4 VAR4	5 VAR5	6 VAR6	7 VAR7	8 VAR8	9 VAR9	10 VAR10
1	740,0	680,000	1,000							
2	670,0	600,000	1,000							
3	560,0	550,000	1,000							
4	540,0	520,000	1,000							
5	590,0	540,000	1,000							
6	590,0	700,000	1,000							
7	470,0	600,000	1,000							
8	560,0	540,000	1,000							
9	540,0	630,000	1,000							
10	500,0	600,000	1,000							
11	750,0	590,000	2,000							
12	360,0	600,000	2,000							
13	720,0	750,000	2,000							
14	540,0	710,000	2,000							
15	570,0	700,000	2,000							
16	520,0	670,000	2,000							
17	590,0	790,000	2,000							
18	670,0	700,000	2,000							
19	620,0	730,000	2,000							
20	690,0	840,000	2,000							
21	610,0	680,000	2,000							
22	550,0	730,000	2,000							
23	590,0	750,000	2,000							
24										
25										

Auto Task Buttons

Read Customize... Output:OFF Set:OFF Weight:OFF

Пуск @MA... Micro... Мой... 1815 (... Stat ST... 11:06

Рис. 15.3. Дані у вигляді таблиці

контекстного меню (права клавіша миші) командою *Modify Variable(s) / Add*, або *Modify Case(s) / Add*.

У створеній таблиці (див. рис. 15.3) змінна Var3 набуває значення “1”, якщо дана фірма ухилялася від податків, а “2” – не ухилялася.

Далі потрібно запустити на виконання саму процедуру дискримінантного аналізу (пункт меню *Analysis* – див. рис. 15.4).

У вікні, що з’являється при цьому, слід натиснути на кнопку *Variables* (Змінні), щоб визначити, за якими змінними має проводитись класифікація (*independent variables*) та яка змінна задає розбиття на групи (*grouping variable*).

У нашому випадку змінні, за якими має проводитись класифікація (*independent variables*) – це var1, var2, а змінна, яка задає розбиття на групи (*grouping variable*) – це var3 (див. рис. 15.5).

Після натиснення на кнопку ОК відбувається повернення до вікна *Model Definition*. Варто залишити встановлений за замовчуванням метод Стандартний (*Standard*) (див. рис. 15.6). Кнопка ОК запускає на виконання процедуру.

Вигляд класифікуючих функцій для обох груп можна встановити, використавши пункт меню *Classification functions*.

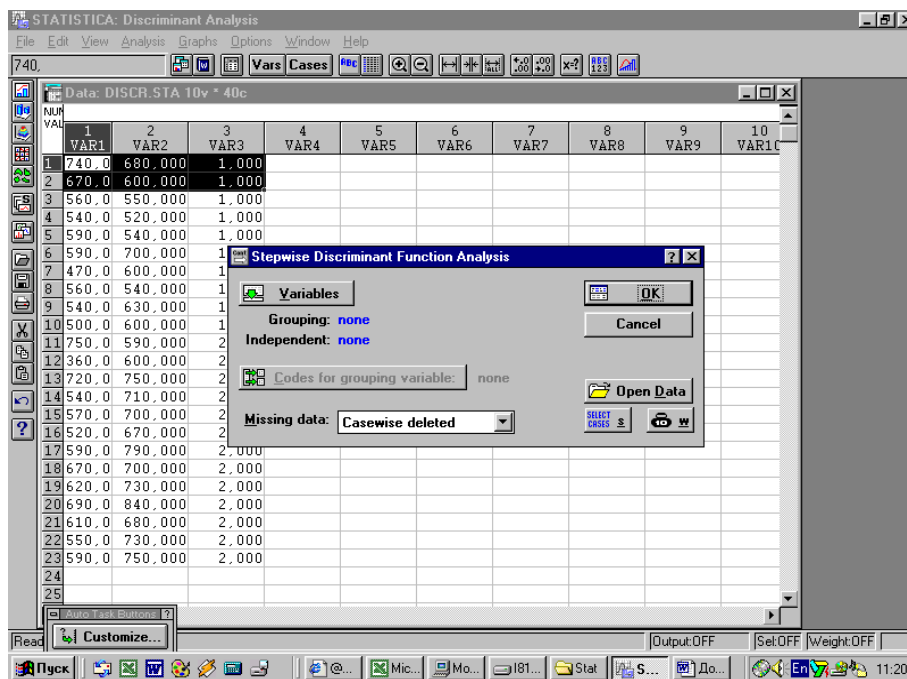


Рис. 15.4. Вікно модуля дискримінантного аналізу

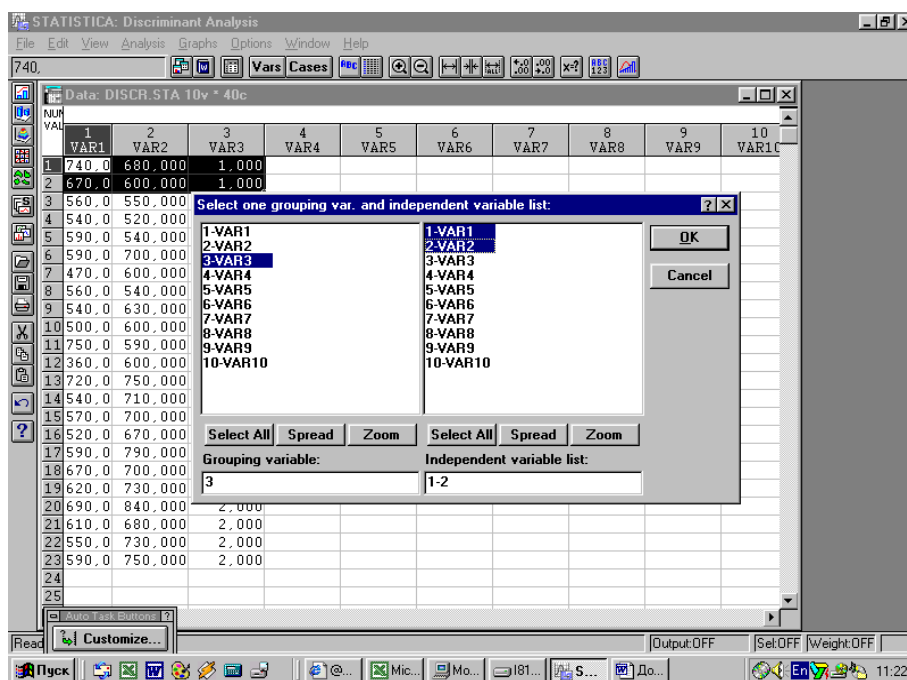


Рис. 15.5. Вибір групуючої і незалежних змінних

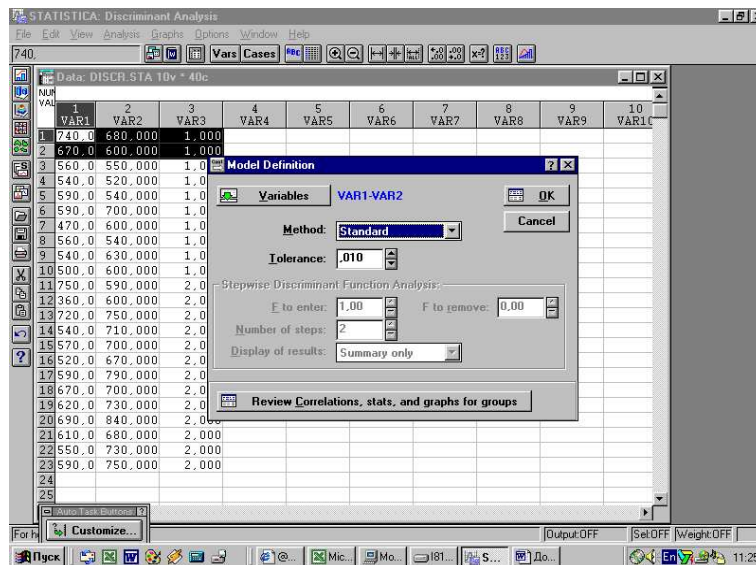


Рис. 15.6. Вибір методу

У нашому випадку отримуємо (див. рис. 15.7) для першої групи

$$f_1 = 0,0389 * var1 + 0,1209 * var2 - 48,0633;$$

для другої

$$f_2 = 0,0346 * var1 + 0,1490 * var2 - 64,1622.$$

Ці функції можна використати, щоб встановити, до якої з виділених груп слід віднести нове спостереження. Для нового спостереження обчислюють значення класифікуючих функцій. Для якої з груп таке значення більше, до тієї групи і зараховують.

У нашому випадку нове спостереження – (740; 590). Підставивши у функції, отримуємо:

$$f_1 = 0,0389 * 740 + 0,1209 * 590 - 48,0633 = 52,0537;$$

$$f_2 = 0,0346 * 740 + 0,1490 * 590 - 64,1622 = 49,3518.$$

Отже, це спостереження слід віднести до першої групи. Це означає, що є підстави для того, щоб діагностувати фірму, яку аналізують як фірму, яка в тій чи іншій формі ухиляється від сплати податків.

У вікні *Posterior Probabilities* (див. рис. 15.8) можна побачити ймовірності того, що нове спостереження належить до існуючих груп (попередньо це спостереження занесли в таблицю останнім, без зазначення групи). Як бачимо, ймовірність того, що воно належить до першої групи – 0,903, а до другої – 0,097, що теж свідчить на користь гіпотези, що ця фірма, очевидно, ухиляється від сплати податків.

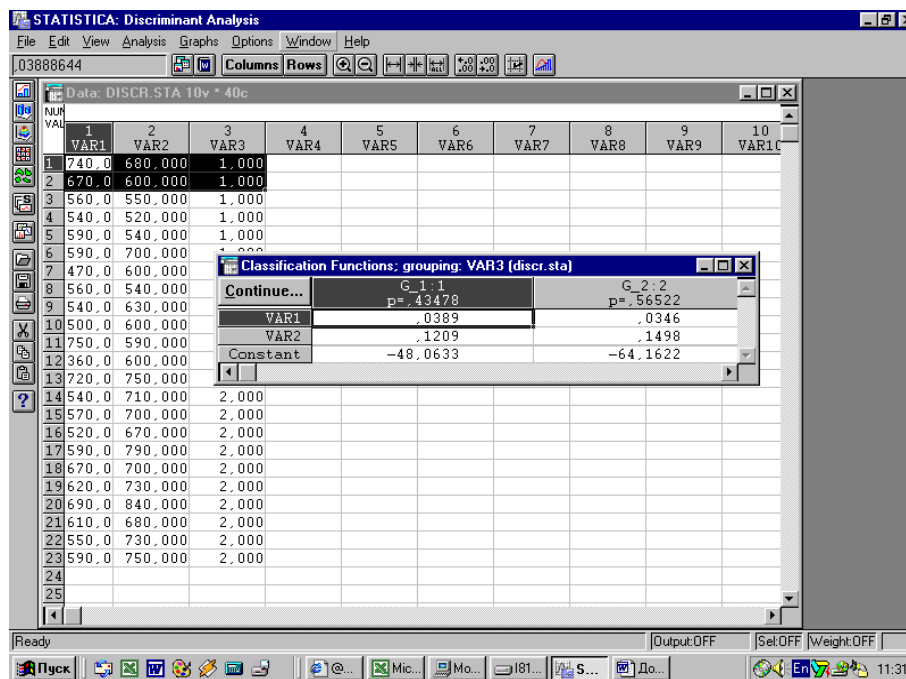


Рис. 15.7. Класифікуючі функції

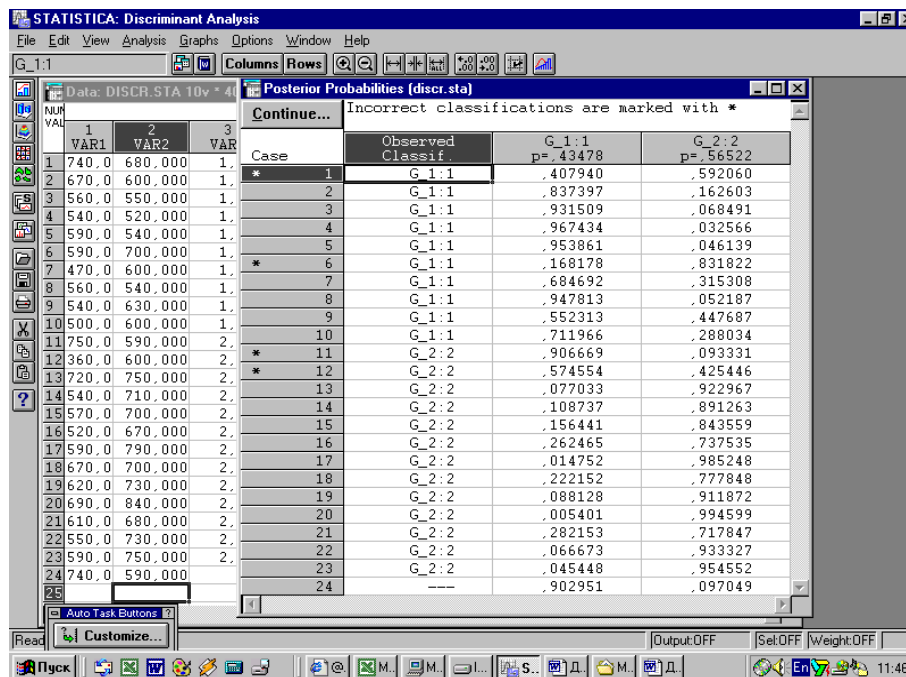


Рис. 15.8. Апостеріорні ймовірності

Розділ 16

Факторний аналіз

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

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

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, \dots, 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, \dots, 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, \dots, f_k .

Метою застосування цього методу є виділення та змістовне тлумачення латентних загальних факторів, кількість яких суттєво менша від

кількості спостережених змінних, і одночасно прагнення мінімізувати залежність змінних від своїх специфічних компонент. Хотілося б виділити кілька факторів, які б досить повно описували модель.

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

Поява головних ідей методу факторного аналізу датується початком 20 ст. (Ч.Спірмен, Л.Терстоци, Г.Томсон), але тільки в 50-х роках ці ідеї отримали теоретичне обґрунтування (наприклад, роботи Т.Андерсона та Г.Рубіна).

16.1 Лінійна модель

Як і раніше, розглядаємо прямокутну матрицю емпіричних спостережень

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

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

16.2 Припущення

Вважаємо, що змінні X_1, X_2, \dots, X_m – центровані (від кожного значення, яке спостерігають, відняли середнє, що означає перенесення початку координат у “центр” набору даних). Вважаємо, що змінні X_1, X_2, \dots, X_m залежать від факторів f_1, f_2, \dots, f_k лінійно:

$$X_i = q_{i1}f_1 + q_{i2}f_2 + \dots + q_{ik}f_k + u_i, \quad i = \overline{1, m}. \quad (16.1)$$

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)

$$X_j = QF_j + U_j. \quad (16.2)$$

(16.2) можна записати як:

$$X = QF + V, \quad (16.3)$$

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

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

У факторній моделі (16.1) вектор спостережень X – нормально розподілений m -вимірний (як лінійна комбінація двох нормально розподілених випадкових векторів F та U).

З рівностей (16.2) і (16.3) отримуємо:

$$\mathbf{E}X_i = 0, \quad \begin{cases} \sigma_{ii} = \sum_{l=1}^k q_{il} + v_{ii} \\ \sigma_{ij} = \sum_{l=1}^k q_{il}q_{jl} \end{cases}, \quad i, j = \overline{1, m} \quad (16.4)$$

або в матричній формі

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

Тут $\sum = (\sigma_{ij})_{i,j=\overline{1,m}}$ позначає коваріаційну матрицю.

(16.2) і (16.3) 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.

Зв'язок факторного аналізу з методом головних компонент розглянуто в [1] та [14].

Існування та однозначність моделі

Виявляється, що не для будь-якого набору вихідних показників $X = (X_1, X_2, \dots, X_m)$ можна вказати задану кількість загальних факторів f_1, f_2, \dots, f_k , які б пояснювали наявну кореляцію між показниками. Не кожна коваріаційна матриця \sum допускає зображення у вигляді (16.4), 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 $F = (f_1, f_2, \dots, f_k)$ (а відповідно, і матриці $Q = (q_{ij})$) не єдине.

У роботі [1] запропоновані окремі випадки апріорних співвідношень, за яких модель однозначно ідентифікується:

1. Розв'язок (Q, V) системи (16.4) належить класові 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.

16.3 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 значущості розбіжності між певною моделлю та набором даних. За відсутності апріорних даних звертаються до однофакторної моделі з наступною перевіркою значущості розбіжності. Якщо розбіжність статистично значуща, то оцінюють модель із ще одним додатковим фактором і знову застосовують критерій. І так процес продовжують доти, доки розбіжність буде визнана незначною, тобто розбіжність буде пояснюватись випадковістю вибірки.

Самі фактори конструюють за коваріаційною (кореляційною) матрицею. Основна математична ідея ґрунтується на відшуканні власних чисел та власних векторів редукованої коваріаційної (кореляційної) матриці. Редукованою кореляційною (adjusted correlation) матрицею називають кореляційну (коваріаційну) матрицю із загальностями на головній діагоналі (як загальності 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, λ – шукане власне число. Сума власних чисел дорівнює кількості змінних, а добуток дорівнює детермінантові 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. Як фактори 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 відхилень) коефіцієнтів кореляції, які обчислені та спостерігають.

Алгоритм у 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 як таку оцінку використовують квадрат множинного коефіцієнта кореляції між відповідною змінною та сукупністю всіх 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 якнайкраще (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 покращання стане неможливим. Описаний алгоритм відомий під назвою “Метод головних факторів з ітераціями по загальностях”.

Може використовуватися також метод мінімальних залишків Хармана, який є теж ітераційний. У цьому методі критерієм зупинки слугує критерій χ^2 .

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

функціонує в припущенні, що дані, які спостерігають, – це вибірка з генеральної сукупності, яка точно відповідає k -факторній моделі.

Може також використовуватися критерій знаходження факторних навантажень, за 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 здебільшого використовують ітераційні схеми.

Усі варіанти методу максимальної вірогідності зводяться до розв'язку характеристичного рівняння $\det(R'' - \lambda I) = 0$, 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 зі змінними, що мають меншу специфічність (u_i).

16.4 Критерії визначення кількості факторів

1. З методами максимальної вірогідності та найменших квадратів найчастіше використовують критерій χ^2 . 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 ґрунтується 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, λ_k) , $k = \overline{1, m}$, де λ_k – власні числа кореляційної матриці,

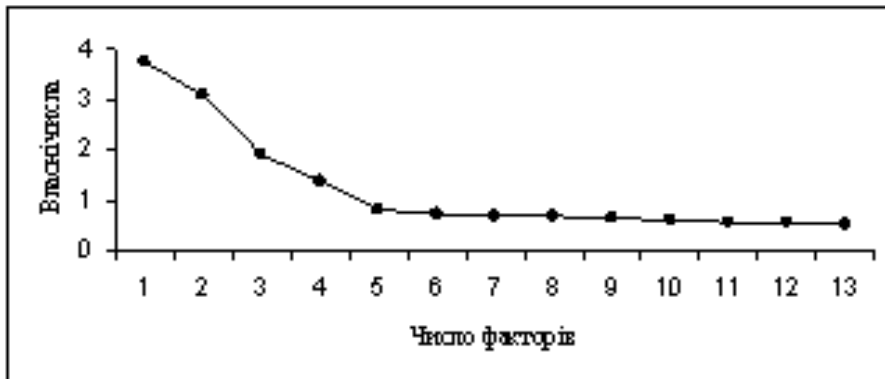


Рис. 16.1. Критерій відсіювання Каттелла

впорядковані за спаданням. Виділення закінчують на факторі, після якого досліджувана залежність наближається до прямої, майже горизонтальної лінії (див. рис. 16.1).

- Критерій інтерпретовності та інваріантності. Пропонують використовувати таку кількість факторів, що узгоджується з усіма наведеними критеріями. Остаточне рішення повинне ґрунтуватися на змістовному тлумаченні в предметній області.

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. Хотілося б позбутись цих недоліків, що утруднюють змістовне тлумачення моделі.

Математично розв'язок (Q, V) системи (16.4) можна шукати лише з точністю до обертання системи координат. Тож виникає питання: чи не можна досягти прозорого тлумачення для іншого набору факторів, отриманого обертанням. Метою всіх обертань є отримання найбільш простої факторної структури.

Зауважимо, що 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 у 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.

Є 3 різні підходи до проблеми обернання:

Перший підхід – графічний. Якщо у k -вимірному просторі факторів спостерігають яскраво виражені групи змінних, скупчення, то є сенс провести нові осі через ці скупчення.

Другий підхід використовує аналітичні методи. Проводять ортогональне або косокутне обернання згідно з певним критерієм.

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

Методи ортогонального обернання

Якщо факторна складність змінної більша 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 із цих квадратів навантажень, а всі інші квадрати – нулі (що означає залежність змінної лише від одного фактора).

Критерій *квартимакс* Q спрямований на обернання 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. У ньому максимізують дисперсію квадратів навантажень фактора

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

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

Можна отримати узагальнений критерій $\alpha Q + \beta V = M$, де α та β – 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 *кватримакс* Q ; при $\gamma = 1$ – *варімакс* V . Критерій, який отримують при $\gamma = k/2$ називають *еквімакс*, при $\gamma = 0,5$ – *бікватримакс*.

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 найпоширеніший *кватримін*, який подібний до ортогонального методу *кватримакс*, тільки не висувається вимога ортогональності факторів.

Вибіркова адекватність факторної моделі

Для вирішення питання адекватності факторної моделі по відношенню до заданого набору змінних було запропоновано (Конізер) спеціальний критерій – “міру вибіркової адекватності” (*МВА*)

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

де r_{ij} – коефіцієнти кореляції, які спостерігають,

g_{ij} – елементи матриці $Q = SR^{-1}S$, тут R – кореляційна матриця, $S = (\text{diag} R^{-1})^{1/2}$.

Коефіцієнт *МВА* 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 *МВА* збільшується при збільшенні кількості змінних, зменшенні кількості загальних факторів, збільшенні обсягу спостережень, збільшенні середнього значення коефіцієнтів кореляції.

Для факторної моделі, отриманої методом максимальної вірогідності, запропоновано (Такер, Левіс) коефіцієнт надійності ρ . На практиці частіше використовують його асимптотичне наближення

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

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

16.5 Виконання в пакеті STATISTICA

Для прикладу розглянемо факторний аналіз шести річних показників господарської діяльності восьми підприємств. Досліджувані дані занесено в таблицю:

№	X_1	X_2	X_3	X_4	X_5	X_6
1	47,1	1285,8	458,9	1,9	576	17,3
2	66,6	608,9	237,6	1,0	764	15,6
3	38,6	487,3	181,5	2,7	649	19,4
4	87,0	908,6	364,4	2,0	868	16,9
5	53,8	710,0	205,7	1,4	970	17,8
6	106,0	1023,7	394,9	1,3	521	18,0
7	74,2	743,8	260,1	1,2	493	9,1
8	47,0	1806,8	355,3	0,4	475	10,0

- X_1 – обсяг товарної продукції, млн. грн.;
 X_2 – втрати від браку, кг;
 X_3 – середньомісячна оплата праці, грн.;
 X_4 – загальні простой технологічного устаткування, тис. днів;
 X_5 – продуктивність праці, тис. грн.;
 X_6 – енергоозброєність, млн. квт-год./чол.

Головні завдання:

- понизити вимірність простору ознак;
- виявити внутрішні латентні властивості підприємств.

Відкрити модуль Факторний аналіз за допомогою перемикача модулів.

На екрані з'явиться стартова панель модуля *Factor Analysis*. В рядку *Input File* – Файл вхідних даних вказуємо тип вхідного файлу. У цьому модулі використовують такі типи вхідних файлів:

Correlation Matrix – Кореляційна матриця;

Raw Data – Вхідні дані. Це звичайний файл, у якому в рядках записані значення змінних. У рядку *Missing Data* – Пропущені дані потрібно задати спосіб обробки пропущених значень:

Casewise – Спосіб обробки пропущених випадків полягає в тому, що виключають з обробки всі випадки (рядки, записи), в яких є хоча б одне пропущене значення (для будь-якої змінної). Залишають лише випадки без жодного пропуску.

Pairwise – Парний спосіб виключення пропущених значень ігнорує пропущені значення тільки для пари фіксованих змінних. При такому способі обробки залишається більше випадків, ніж при попередньому. Правда, можуть з'явитись неузгодженості у зв'язку з тим, що різні коефіцієнти можуть бути обчислені по різних випадках, а також з різною кількістю вхідних даних.

Mean Substitution – Підстановка середнього замість пропущених значень.

Відкриємо файл. Для цього використовуємо кнопку *Open Data* – Відкрити дані.

У вікні *Open Data File* – вибираємо файл з даними.

Після цього відкриваємо стартову панель модуля *Factor Analysis* – Факторний аналіз. Вибираємо змінні у відкритому файлі для аналізу за допомогою кнопки *Variables* – Змінні.

У вікні вибору змінних *Select the variables for the factor analysis* – Вибрати змінні для факторного аналізу виконуємо вибір. Вибирати змінні можна за допомогою курсора миші, позначаючи назви або номери відповідних змінних. Кнопка *Select All* – Вибрати все дозволяє вибрати всі змінні одразу.

Використавши *Spread* – Розкрити, можемо продивитись розширений опис змінних.

Після натиснення у стартовому вікні модуля кнопки *OK* починаємо виконання процедури факторного аналізу для вибраних змінних.

На першому кроці обчислюємо кореляційну матрицю (якщо тільки кореляційна матриця не була задана).

Потрібно вибрати метод виділення факторів у вікні *Define Method of Factor Extraction*, див. рис. 16.2.

Можливі такі варіанти:

Principal components – Метод головних компонент.

Principal factor analysis – група методів, об'єднаних під назвою Аналіз головних (загальних) факторів:

*Communalities=multiple R**2* – Загальності дорівнюють квадратові коефіцієнта множинної кореляції.;

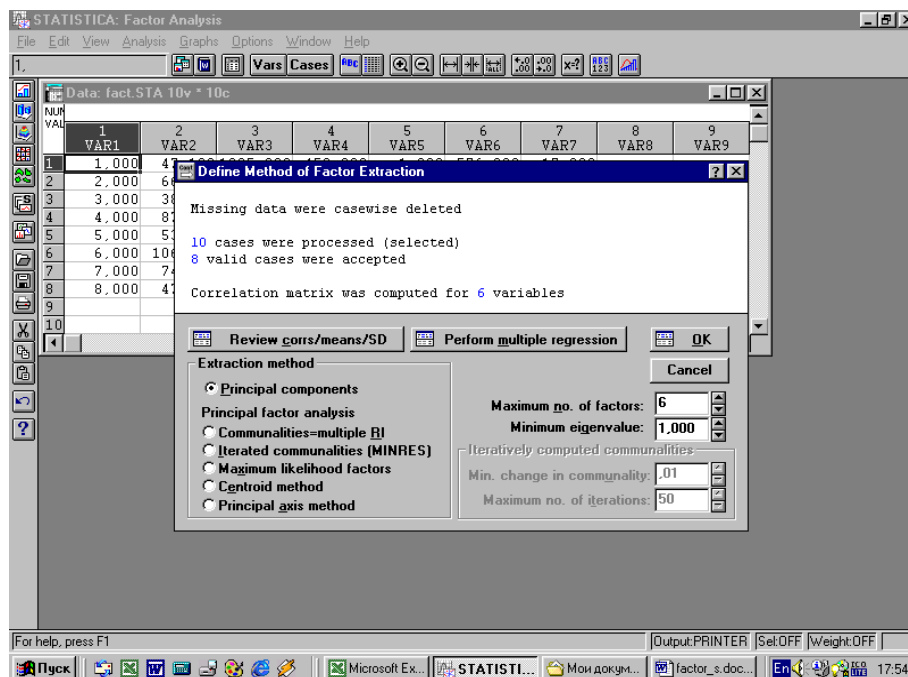


Рис. 16.2. Вибір методу

Iterated communalities (MINRES) – Метод ітераційних залишків (мінімальних залишків);

Centroid method – Центроїдний метод;

Principal axis method – Метод головних осей.

У правій частині вікна потрібно задати *Maximum no. of factor* – Максимальну кількість факторів, які будуть виділені, та *Minimum eigenvalue* – Мінімальне власне значення, всі власні значення, менші від заданого, будуть ігноруватися.

Кнопка *Cancel* повертає до стартового вікна модуля.

Кнопка *Rewier corr/s/means/SD* – Продивитись кореляції/середні/стандартні відхилення відкриває вікно, у якому можна проглянути середнє значення, стандартні відхилення, кореляції, коваріації, побудувати графіки. Наприклад, для обчислення кореляційної матриці для вибраних змінних потрібно натиснути кнопку *Correlation*. Після обчислення кореляцій повернутися до вікна *Define Method of Factor Extraction* можна за допомогою кнопки *Continue*.

За допомогою кнопки *Perform multiple regression* – Виконати множинну регресію можна виконати множинну регресію без виходу з модуля.

Після вибору методу факторизації (*Principal components* – Метод головних компонент) і натиснення кнопки *OK* з'являється вікно *Factor*

Analysis Results (див. рис. 16.3), яке містить таку інформацію:

Number of variables – Кількість аналізованих змінних;

Method – Метод аналізу;

log(10) determination of correlation matrix – Десятковий логарифм визначника кореляційної матриці;

Number of factor extraction – Кількість виділених факторів;

Eigenvalues – Власні значення.

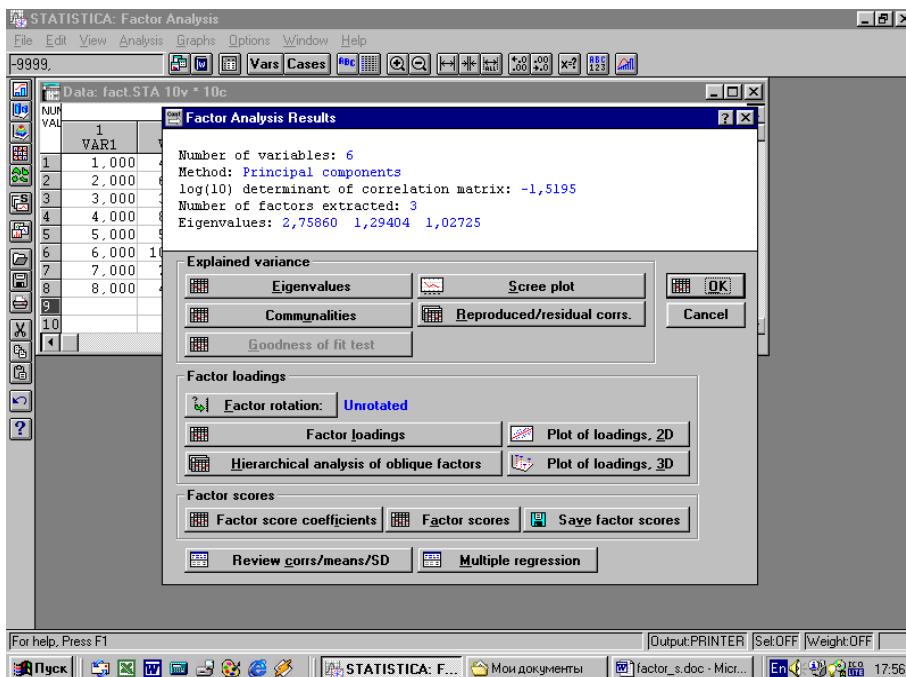


Рис. 16.3. Вікно аналізу результатів

Для компонентного аналізу найважливіше з'ясувати власні значення (характеристичні корені). Для даної матриці вони наведені у таблиці на рис. 16.4.

Тут у першому стовпці знаходяться самі характеристичні корені λ_L , в другому – відсоток загальної дисперсії показників, що пояснюють головні компоненти, в третьому – накопичені значення характеристичних коренів, у четвертому – накопичений відсоток загальної дисперсії змінних, що пояснюють головні компоненти.

Після обчислення власних значень простір загальних факторів вважають знайденим, проте самі ці загальні фактори вважають визначеними однозначно по відношенню до обертання осей знайденого простору. За допомогою процедури *Factor rotation* – Обертання факторів підбирають таку комбінацію загальних факторів, яка дає змістовне тлумачення

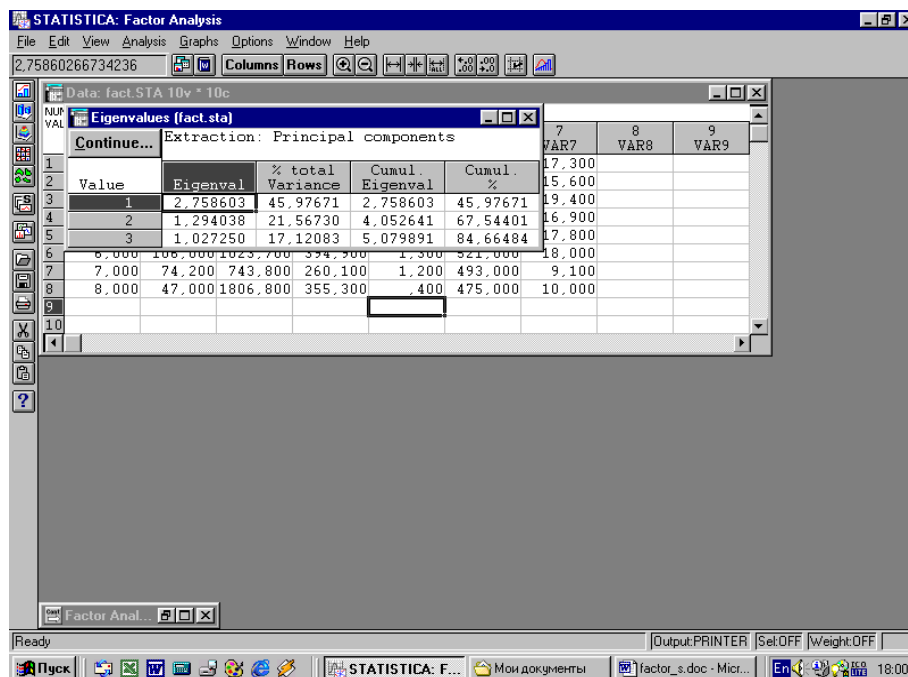


Рис. 16.4. Власні значення

моделі. Серед запропонованих методів ротації: *Varimax*; *Biquartimax*; *Quartimax*; *Equamax*. Якщо факторні навантаження в процедурі нормалізують діленням на квадрат загальностей, то використовують нормалізований (*normalized*) метод, якщо ж використовують початкові (*raw*) дані, то це означає, що факторні навантаження не піддають нормалізації. Наприклад, *Varimax normalized* означає, що буде використано метод варімакс із нормалізацією факторних навантажень.

Можна також розглянути і графічне зображення результатів факторизації, використавши *Plot of loadings*, 2D – Двовимірний графік факторних навантажень.

Для перегляду таблиці факторних навантажень використовують кнопку *Factor loadings*.

Далі аналізують таблицю факторних навантажень (див. рис. 16.5). Якщо за величиною факторних навантажень важко дати змістовне тлумачення факторів, то доцільно проводити поворот різними доступними методами та побудову графіка факторних навантажень аж до отримання задовільної моделі.

З рис. 16.5 робимо висновок, що для наших даних високі факторні навантаження по першому фактору мають показники X_4 та X_6 , по другому – показники X_2 , X_3 , а по третьому – X_1 .

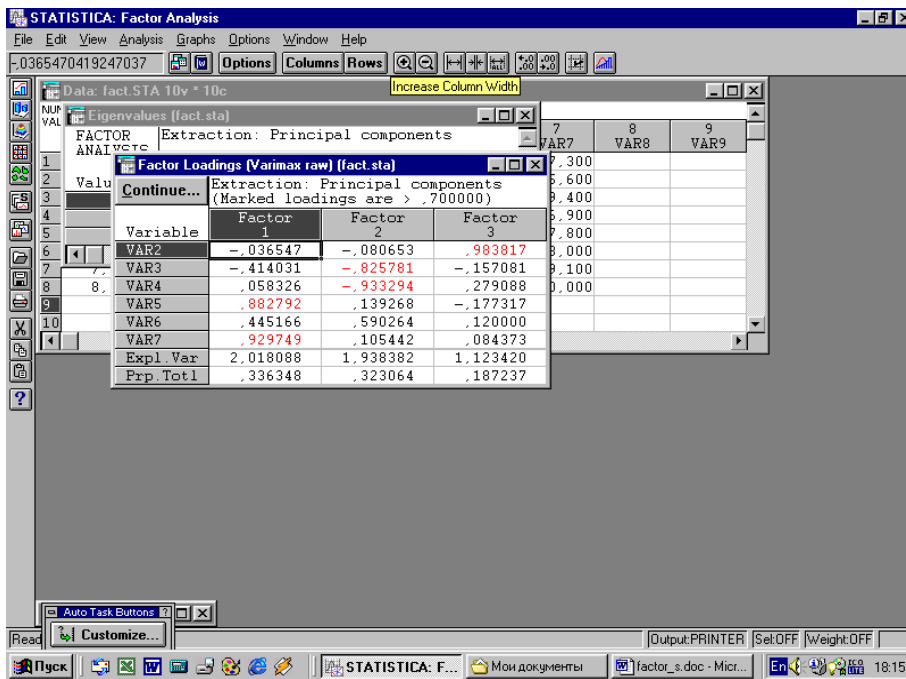


Рис. 16.5. Факторні навантаження

Отже, нами виділено три фактори (знижено вимірність простору ознак з 8 до 3), які описують майже 85 % загальної дисперсії досліджуваних показників.

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] Борисенко О.Д., Майборода Р.Є. *Аналітико-статистичні методи й моделі психології та педагогіки (вибрані лекції)*. – К.: РВЦ “Київський університет”, 2000.
- [9] Боровиков В. П. *Популярное введение в программу Statistica*. – М.: Компьютер Пресс, 1998.

- [10] Гойко О.В. *Практичне використання пакета STATISTICA для аналізу медико-біологічних даних.* – К.: Київська мед. академія ім. П.Л. Шупика, 2004.
- [11] Джессен Р. *Методы социологических обследований.* – М.: Финансы и статистика, 1985.
- [12] Зінченко Н.М., Оленко А.Я. *Аналітичні моделі та методи у соціології.* – К.: РВЦ “Київський університет”, 2000.
- [13] *Електронний підручник по пакету STATISTICA.* –StatSoft, Inc. www.statsoft.ru/home/textbook”
- [14] Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др. *Факторный, дискриминантный и кластерный анализ.* – М.: Финансы и статистика, 1989.
- [15] Кокрен У. *Методы выборочного исследования.* – М.: Финансы и статистика, 1976.
- [16] Крамер Г. *Математические методы статистики.* – М.: Мир, 1975.
- [17] Салин В.Н., Чурилова Э.Ю. *Практикум по курсу “Статистика” в системе STATISTICA.* – М.: Соц. отношения, 2002.
- [18] Турчин В.М. *Математична статистика в прикладах і задачах.* – К.: НМК В, 1993.
- [19] Тюрин Ю.Н., Макаров А.А. *Статистический анализ данных на компьютере.* – М.: Инфра, 1998.
- [20] Феллер В. *Введение в теорию вероятностей и ее приложения.* В 2-х томах. – М.: Мир, 1984.
- [21] Шварц Г. *Выборочный метод.* –М.: Статистика, 1978.
- [22] Янковой А.Г. *Многомерный анализ в системе STATISTICA.* –О.: OPTIMUM, 2001.
- [23] Clifford H.T., Stephenson W. *An introduction to numerical classification.* – N.Y.: Academic Press, 1975.
- [24] Gower J.C. *A general coefficient of similarity and some of its properties// Biometrics*, **27** (1971). –P.857–872.

- [25] Greene W.H. *Econometrics analysis*. – Prentice Hall; 2002.
- [26] Leonard T., Hsu J.S.J. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. – Cambridge Univ. Press, 2001.
- [27] Monahan J. *Numerical Methods of Statistics*. – Cambridge Univ. Press, 2001.
- [28] Robert C.P. *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. – Springer-Verlag, 2001.
- [29] Venables W.N., Ripley B.D. *Modern Applied Statistics with S-Plus*. – Springer-Verlag, 2002.
- [30] Wilcox R.R. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. – Springer-Verlag, 2001.

Предметний покажчик

- H_0 , 53
- H_1 , 53
- $L_{(\theta_1, \theta_2, \dots, \theta_s)}(\zeta)$, 20
- S^2 , 39
- \bar{x} , 39
- $\chi^2_{\alpha, n}$, 41
- \hat{a} , 17
- \hat{m}_i , 19
- $\ln L_{(\theta_1, \theta_2, \dots, \theta_s)}(\zeta)$, 20
- σ^2 , 38
- d_P , 22
- f , 9
- r_s , 85
- $t_{\alpha, n}$, 40
- x_i^* , 8
- t-тест, 61
 - залежних вибірок, 66
 - незалежних вибірок, 61
- апроксимація
 - нормальним розподілом, 33
- варіаційний ряд, 8
 - побудова, 13
- вбірка, 7
 - безповторна, 8
 - генерація, 11
 - з поверненням, 8
 - знаходження характеристик, 42
 - навчаюча, 171
 - проста випадкова, 7
- вбіркоче значення параметра, 19
- власні значення, 154
- гістограма частот
 - побудова, 15
- гіпотеза
 - альтернативна, 53
 - нульова, 53
 - основна, 53
 - про однорідність, 93
 - проста, 53
 - складна, 54
 - статистична, 53
- генеральна сукупність, 6
- гомоскедастичність, 100
- дендрограма, 114
- дискримінантний аналіз
 - нелінійний, 177
 - непараметричний, 172
 - параметричний, 172, 173
- дисперсійний аналіз, 134
 - двофакторний, 140
 - однофакторний, 134
 - основна рівність, 135
- ефективність, 18
- загальна сукупність, 6
- змінні
 - екзогенні, 97
 - ендогенні, 97
 - категоризовані, 145
- канонічні ваги, 155
- квантиль, 22
 - нормального розподілу, 51

202 ПРЕДМЕТНИЙ ПОКАЖЧИК

- кластери, 107
- коефіцієнт
 - відмінності, 111
 - варіації, 45
 - кореляції Спірмена
 - вибірковий, 85
 - подібності
 - бінарних змінних, 108
 - неперервних змінних, 109
- конзистентність, 18
 - строга, 18
- кореляція
 - канонічна, 154
 - рангова, 85
- критерій, 54
 - F^2 , 152
 - χ^2 , 90, 152
 - χ^2 з поправкою Йетса, 152
 - χ^2 скоректований, 152
 - t , 61
 - Вілкоксона, 75
 - знаків, 73
 - Колмогорова – Смірнова, 89
 - Макнемара, 153
 - Манна і Уїтні, 82
 - найбільш потужний, 55
 - непараметричний, 73
 - рівень значущості, 54
 - Фішера, 153
 - функція потужності, 55
- критична
 - множина, 54
 - область, 54
- максимальної вірогідності
 - метод, 20
 - функція, 20
 - логарифмічна, 20
- матриця
 - коваріаційна, 173
 - кореляційна
 - редукована, 186
- метод
 - найменших квадратів, 21, 98
 - кластеризації
 - k-means clustering, 127
 - two way joining, 131
 - агломеративний, 114
 - за центрами, 120
 - найближчих сусідів, 115
 - найвіддаленіших сусідів, 118
 - одиначного зв'язку, 126
 - повних зв'язків, 126
 - подрібнення, 123
 - середніх групових відстаней, 119
 - середнього зв'язку, 126
 - Уорда, 122, 127
 - максимальної вірогідності, 20
 - моментів, 19
- моменти
 - емпіричні, 19
- мультиколінеарність, 101
- надійність, 22
- надійний проміжок, 22
 - величина, 45
 - математичного сподівання, 40
- надлишковість, 156
- незміщенність, 18
- оцінка
 - дисперсії, 38
 - емпірична, 19
 - ефективність, 18
 - інтервальна, 22
 - конзистентність, 18
 - строга, 18
 - незміщенність, 18
 - параметра, 17
 - середнього, 38
 - точкова, 22

- помилка
 - другого роду, 54
 - першого роду, 54
- похибка
 - гранична, 45
 - середньоквадратична, 44
 - стандартна, 44
- ранг, 75
- регресія
 - крос-секційна, 97
 - часових рядів, 97
- середнє
 - вибіркове, 39
- статистика, 55
 - Вілкоксона, 75
 - порядкова, 8
- фактори, 183
 - критерій
 - варімакс, 190
 - визначення кількості, 188
 - інваріантності, 189
 - інтерпретовності, 189
 - Кайзера, 188
 - Каттелла, 188
 - квартимакс, 190
 - Хармана, 188
 - методи обертання, 189
 - квартимін, 191
 - косокутного, 191
 - облімін, 191
 - ортогонального, 190
- факторна структура, 155
- факторний аналіз, 183
- функція
 - втрат, 170
 - дискримінантна, 169
 - класифікуюча, 176
 - розподілу
 - емпірична, 14
- числове значення, 35
- центроїд, 173

Зміст

Передмова	3
1 Методи вибірових обстежень	5
1.1 Планування статистичних обстежень	5
1.2 Основні поняття математичної статистики	6
1.3 Типи вибірок зі скінченної загальної сукупності	8
1.4 Простий та стратифікований випадковий вибір	9
1.5 Багатоступеневий гніздовий відбір	10
1.6 Виконання в пакеті STATISTICA	11
2 Оцінювання параметрів	17
2.1 Властивості оцінок	17
2.1.1 Незміщеність	18
2.1.2 Конзистентність	18
2.1.3 Ефективність	18
2.2 Методи одержання оцінок	19
2.2.1 Емпіричні оцінки	19
2.2.2 Метод моментів	19
2.2.3 Метод максимальної вірогідності	20
2.2.4 Метод найменших квадратів	21
2.3 Точкові та інтервальні оцінки	22
2.4 Квантилі	22
2.5 Виконання в пакеті STATISTICA	23
3 Оцінювання частки та кількості елементів	29
3.1 Точкові оцінки для P і N_1	29
3.2 Довірчі інтервали для P і N_1	32
3.2.1 Точні методи	32
3.2.2 Повторна вибірка. Біноміальний розподіл	33
3.2.3 Нормальна апроксимація. Симетричні інтервали	33
3.2.4 Несиметричні інтервали	34
3.3 Виконання в пакеті STATISTICA	35

4	Оцінювання середніх і сумарних значень	38
4.1	Оцінювання середнього та мінливості	38
4.2	Надійні проміжки для параметрів нормальних випадкових величин	40
4.3	Розподіл вибіркового середнього	41
4.4	Виконання в пакеті STATISTICA	42
5	Визначення обсягу вибірки	44
5.1	Показники точності оцінювання	44
5.2	Визначення обсягу вибірки n при оцінюванні часток P	45
5.2.1	Визначення n при заданій граничній похибці e_p	46
5.2.2	Визначення обсягу вибірки при заданій відносній точності	47
5.3	Обсяг вибірки при дослідженні декількох ознак	48
5.4	Визначення обсягу вибірки при оцінюванні середніх і сумарних значень	48
5.4.1	Визначення обсягу вибірки n при оцінюванні середнього при заданій середній квадратичній похибці	48
5.4.2	Визначення обсягу вибірки при оцінюванні середнього при заданій граничній похибці $e_{\bar{x}}$	49
5.4.3	Визначення обсягу вибірки при оцінюванні сумарного значення при заданій середньо- квадратичній або граничній похибці	49
5.4.4	Визначення обсягу вибірки при заданій відносній точності	50
5.5	Обсяг вибірки за необхідності отримати оцінки для підрозділів сукупності	50
5.6	Виконання в пакеті STATISTICA	51
6	Перевірка статистичних гіпотез	53
6.1	Постановка проблеми, основні поняття	53
6.2	Перевірка гіпотез для нормальних розподілів	55
6.2.1	Перевірка гіпотези про значення середнього	55
6.2.2	Перевірка гіпотези про рівність середніх	57
6.2.3	Перевірка гіпотези про значення дисперсії	58
6.2.4	Перевірка гіпотези про рівність дисперсій	59
6.3	Виконання в пакеті STATISTICA	61

7	Непараметричні критерії	73
7.1	Критерій знаків	73
7.2	Критерій Вілкоксона	75
7.3	Виконання в пакеті STATISTICA	76
7.4	Критерій Манна і Уїтні	82
7.5	Виконання в пакеті STATISTICA	82
7.6	Рангова кореляція	85
7.7	Виконання в пакеті STATISTICA	86
8	Тести про вигляд розподілу	89
8.1	Двовибірковий критерій Колмогорова – Смірнова	89
8.2	Виконання в пакеті STATISTICA	89
8.3	Критерій χ^2 та його застосування	90
8.3.1	Перевірка гіпотези про вид розподілу	90
8.3.2	Перевірка гіпотези про вид розподілу, який залежить від невідомих параметрів	91
8.3.3	Перевірка гіпотези про однорідність	93
8.4	Виконання в пакеті STATISTICA	93
8.5	Перевірка гіпотези про незалежність випадкових величин	94
9	Лінійні регресійні моделі	97
9.1	Парна лінійна регресія	97
9.1.1	Метод найменших квадратів	98
9.1.2	Перевірка моделі на адекватність	99
9.1.3	Перевірка моделі на значущість	100
9.2	Множинна лінійна регресія	100
9.3	Виконання в пакеті STATISTICA	101
10	Кластерний аналіз	107
10.1	Коефіцієнти подібності бінарних змінних	108
10.2	Подібність змінних із неперервними значеннями	109
10.3	Коефіцієнти відмінності	111
10.4	Міжгрупові відстані	112
11	Ієрархічна кластерна техніка	114
11.1	Агломеративні методи	114
11.1.1	Метод найближчих сусідів	115
11.1.2	Метод найвіддаленіших сусідів	118
11.1.3	Метод середніх групових відстаней	119
11.1.4	Кластеризація за центрами	120

11.1.5	Метод Уорда	122
11.2	Вибір кількості кластерів	122
11.3	Методи подрібнення	123
11.4	Виконання в пакеті STATISTICA	126
12	Дисперсійний аналіз	134
12.1	Однофакторний дисперсійний аналіз	134
12.2	Виконання в пакеті STATISTICA	136
12.3	Двофакторний дисперсійний аналіз	140
12.4	Виконання в пакеті STATISTICA	142
13	Категоризовані дані	145
13.1	Гіпотези про розподіл частот	145
13.2	Виконання в пакеті STATISTICA	146
13.3	Гіпотези про незалежність ознак	148
13.4	Виконання в пакеті STATISTICA	148
13.5	Оцінка залежності дворівневих даних	150
13.6	Виконання в пакеті STATISTICA	151
14	Канонічний аналіз	154
14.1	Загальні положення	154
14.2	Виконання в пакеті STATISTICA	157
15	Дискримінантний аналіз	169
15.1	Загальні положення	169
15.2	Параметричний дискримінантний аналіз (випадок нормального розподілу класів)	173
15.3	Нелінійний дискримінантний аналіз.	177
15.4	Виконання в пакеті STATISTICA	178
16	Факторний аналіз	183
16.1	Лінійна модель	184
16.2	Припущення	184
16.3	Алгоритм методу	186
16.4	Критерії визначення кількості факторів	188
16.5	Виконання в пакеті STATISTICA	192
	Рекомендована література	198
	Предметний покажчик	201
	Зміст	204