European Commission
TEMPUS

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

ЛУПАН І. В., АВРАМЕНКО О. В.

Комп'ютерні статистичні пакети

Навчально-методичний посібник

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Комп'ютерні статистичні пакети

Навчально-методичний посібник

Лупан І.В., Авраменко О.В.

Кіровоград, 2010

УДК 31
Л 85
ББК 60.6

Рецензенти:
Жалдак М.І. – *академік Академії педагогічних наук України, директор інституту інформатики Національного педагогічного університету імені М.П.Драгоманова, доктор педагогічних наук, професор*
Плічко А.М. – *доктор фізико-математичних наук, професор кафедри прикладної математики, статистики та економіки Кіровоградського державного педагогічного університету імені Володимира Винниченка*

Лупан І.В., Авраменко О.В.
Комп'ютерні статистичні пакети: навчально-методичний посібник. – Кіровоград, 2010. – 218 с.

ISBN

У посібнику наведено завдання, приклади та методичні рекомендації до виконання лабораторних робіт з дисципліни "Комп'ютерні статистичні пакети" для спеціальності "Статистика" та спеціалізації "Освітні вимірювання". Він також буде корисним при вивченні дисципліни "Аналіз даних" для спеціальності "Інформатика". Метою курсу лабораторних робіт є ознайомлення студентів з інструментарієм декількох статистичних пакетів

У посібнику показано у порівнянні інструментарій пакетів MS Excel, SPSS та Statistica для обчислення параметрів розподілів випадкової величини, виконання параметричного та непараметричного порівняння двох та більше зв'язаних та незв'язаних вибірок, у тому числі одно- та двофакторного дисперсійного аналізу, процедури обчислення коефіцієнтів кореляції та регресії, критерії порівняння емпіричних розподілів та перевірки нормальності емпіричних розподілів випадкової величини, виконання дискримінантного та кластерного аналізу; інструментарій для графічного подання результатів та засоби формування звітів.

Рекомендації розроблялися з огляду на MS Excel 2003, SPSS 10.0 та Statistica 6.0, однак, сподіваємося, що вони будуть на користь і при використанні інших статистичних пакетів та інших версій перелічених пакетів.

Наведено також відповідні статистичні функції пакета OpenOffice.org Calc.

Зміст:

Тема 1: “Описові статистики” 5
Теоретичні відомості....6
Завдання 1 (MS Excel)....9
Завдання 2 (SPSS)....12
Завдання 3 (Statistica):....13
Приклад виконання....15
Контрольні запитання....16
Тема 2: “Порівняння параметрів двох вибірок” 18
Теоретичні відомості....18
Завдання 1: порівняння параметрів двох незалежних вибірок....20
Завдання 2: порівняння параметрів двох залежних вибірок....21
Приклад виконання....22
Контрольні запитання....28
Тема 3: “Дисперсійний аналіз” 30
Теоретичні відомості....30
Завдання 1: однофакторний дисперсійний аналіз....34
Завдання 2: однофакторний дисперсійний аналіз в пакетах SPSS та Statistica....34
Завдання 3: двофакторний дисперсійний аналіз....35
Завдання 4: двофакторний дисперсійний аналіз в пакетах SPSS та Statistica....35
Приклад виконання....36
Контрольні запитання....42
Тема 4: “Кореляційний та регресійний аналіз” 43
Теоретичні відомості....43
Завдання 1: Кореляційний аналіз....47
Завдання 2: Регресійний аналіз....49
Приклади виконання....50
Контрольні за питання....67
Тема 5: “Порівняння розподілів” 69
Теоретичні відомості....70
Завдання1: Оцінка нормальності емпіричного розподілу....77
Завдання 2: Порівняння емпіричного розподілу з деяким теоретичним....79
Завдання 3: Порівняння двох та більше емпіричних розподілів....84

Завдання 4: Визначення зв'язку (кореляції) між якісними ознаками. 86
Приклади виконання 89
Контрольні запитання 111
Тема 6: "Непараметричні методи" 112
Теоретичні відомості. 112
Завдання 1: дві вибірки 122
Завдання 2: декілька вибірок 122
Приклади виконання 122
Контрольні запитання 143
Тема 7: "Дискримінантний аналіз" 145
Теоретичні відомості. 145
Завдання 1: виконання дискримінантного аналізу засобами пакета SPSS 151
Завдання 2: виконання дискримінантного аналізу засобами пакета Statistica 152
Приклад виконання 153
Контрольні запитання 167
Тема 8: "Кластерний аналіз" 169
Теоретичні відомомсті. 169
Завдання 1: Ієрархічний кластерний аналіз 175
Завдання 2: Ітеративний кластерний аналіз 176
Приклади виконання 177
Контрольні запитання 182
Індивідуальні завдання 183
Індивідуальне завдання 1 183
Індивідуальне завдання 2 183
Індивідуальне завдання 3 184
Список використаних джерел 185
Глосарій позначень статистичних функцій: 188
Додаток А. Вибірки 194
Додаток Б. Варіанти завдань до лабораторних робіт 202
Завдання до теми 2 202
Завдання до теми 3 203
Завдання до теми 5 211

Тема 1: “Описові статистики”

Мета:

1. Повторити основні параметри розподілу значень статистичної змінної: міри центральної тенденції (середні, мода, медіана), міри мінливості (дисперсія, стандартне квадратичне відхилення), характеристики діапазону розподілу (розмах, мінімальне, максимальне), характеристики форми розподілу[1] (асиметрія, ексцес), стандартні похибки.
2. Ознайомитися з інструментарієм пакета *MS Excel* для табулювання та ранжування змінних, графічного представлення рядів розподілу, статистичними функціями обчислення параметрів описової статистики.
3. Ознайомитися з інтерфейсом статистичного пакету *SPSS*: структурою робочого аркуша, редактором даних, можливостями зчитування та зберігання даних у файлах різних форматів (текстових, електронних таблиць, баз даних), особливостями формування та збереження звітів, інструментарієм для виконання процедур дескриптивної статистики.
4. Ознайомитися з інтерфейсом статистичного пакету *Statistica*, його інструментарієм для виконання процедур дескриптивної статистики.

Після виконання лабораторної роботи студенти ***повинні знати:***

– параметри та основні формули обчислення параметрів розподілу, умови їх застосування;
– статистичні функції *MS Excel* та умови їх застосування; прийоми роботи з формулами-скалярами та формулами-масивами, засоби побудови гістограми та кумуляти статистичного розподілу;
– засоби зчитування та збереження даних різних форматів в пакеті *SPSS*, засоби виконання процедур дескриптивної статистики;
– засоби виконання процедур дескриптивної статистики пакету *Statistica*.

Студенти ***повинні уміти***:

– подавати експериментальні дані у вигляді, зручному для обчислень;

---

[1] Для скорочення називатимемо далі розподіл імовірності випадкової величини просто розподілом.

– визначати шкали, за якими виконано вимірювання;
– форматувати та редагувати експериментальні дані засобами пакетів *MS Excel, SPSS, Statistica*;
– виконувати в *MS Excel* обчислення за алгоритмом статистичного критерію;
– обчислювати середнє арифметичне, дисперсію, стандартне квадратичне відхилення, асиметрію та ексцес для вибірки засобами пакетів *MS Excel, SPSS, Statistica*;
– будувати ряд розподілу частот та відповідні графіки для нього;
– ранжувати експериментальні дані.

Теоретичні відомості

Побудова інтервального статистичного ряду

Варіаційні ряди будують для виявлення закономірностей варіювання ознаки (змінної), тому при побудові слід розбивати значення ознаки на таку кількість класів, щоб при збереженні точності, досягнутої у вимірюваннях, представити досліджувану ознаку найбільш наочно. Також слід пам'ятати, що при застосуванні критеріїв згоди (с. 71) не повинно бути не лише інтервалів з нульовою частотою, а і з частотою менше 5.

Оптимальну кількість класів розбиття визначають [28]:
а) за формулою Стерджеса (для n<=100): K=1+3,32 *lg n*;
б) за формулою К.Брукса та Н.Каррузерса: K=5 *lg n*;
в) за таблицею [5, с.32]:

Обсяг вибірки (від – до)	Кількість класів	За Стерджесом: 1+3,32 lg N	За Бруксом та Карузерсом: 5 lg N
25 – 40	5 – 6	6,3	8
40 – 60	6 – 8	6,9	8,9
60 – 100	7 – 10	7,6	10
100 – 200	8 – 12	8,6	11,5
>200	10 – 15	11 (N=1000)	15 (N=1000)

Порядок побудови інтервального статистичного ряду такий:
1. Відшукують серед числових значень ознаки найменше (X_{min}) та найбільше (X_{max}).
2. Визначають розмах вибірки $R= X_{max} – X_{min}$.

3. За таблицею або формулою визначають оптимальну кількість класів (K).
4. Обчислюють величину класового інтервалу: i=R/K.
5. Округлюють отримане число так, щоб величина класового інтервалу відповідала точності, з якою досліджувану ознаку вимірювали.
6. Визначають нижню границю першого класового інтервалу так, щоб мінімальна варіанта потрапила в його середину. Тобто $L= X_{min} – i/2$.
7. Визначають нижні границі наступних класових інтервалів, додаючи до L відповідно i, 2i, 3i і так далі.
8. Верхні границі класових інтервалів визначають так, щоб вони відрізнялися від нижніх границь наступних класів на величину, що дорівнює точності вимірювання.

Інколи доводиться використовувати безінтервальні ряди. Тоді кожний класовий інтервал замінюють середнім (центральним) значенням варіанти, тобто значенням Xc=L+i/2.

Порядок ранжування

Ранжуванням називають приписування значенням вибірки порядкових номерів (рангів), таким чином, щоб однакові елементи вибірки отримали однакові ранги, але при цьому сума рангів однакових за обсягом вибірок була однакова (див. п. 3-4 нижче).

1. Найменшому значенню приписати ранг 1 (одиниця).
2. Найбільшому значенню приписати ранг рівний загальній кількості елементів (N).
3. Якщо 2 або більше значень рівні між собою, то їм приписують однаковий ранг – *середнє*[2] рангів, які б мали ці значення, якби були різними (функція РАНГ у такому випадку приписує однаковий мінімальний ранг).
4. Сума отриманих рангів повинна дорівнювати половині добутку N на N+1 (тут N – загальна кількість ранжованих елементів): $\sum R_i = \frac{N \cdot (N+1)}{2}$.

[2] Тільки у цьому випадку буде забезпечено виконання п. 4. Однак у пакеті Excel, наприклад функцією РАНГ однаковим елементам вибірки приписують найменший для даної підвибірки порядковий номер.

Шкали вимірювання

Назвемо вимірюванням процес приписування досліджуваним ознакам об'єктів числових характеристик за певним правилом.

Правило, відповідно до якого об'єктам приписують числові значення, називатимемо шкалою вимірювання.

Характеристики та приклади вимірювальних шкал [17 с. 16].

Характеристики	***Приклади***
Номінативна шкала (шкала назв)	
Об'єкти класифіковано, а класи позначено номерами. Відношення між номерами ніяк не пов'язані з властивостями об'єктів.	Колір очей, номери одягу та взуття, стать, клінічні діагнози, автомобільні номери, номери телефонів, типи темперамента.
Порядкова шкала (рангова)	
Номери об'єктів відображують кількісні характеристики властивостей, а саме відношення "більше-менше". Але, однакові різниці між числами не означають однакових різниць у кількостях властивостей.	Нагороди за заслуги, військові ранги, твердість мінералів, шкільні оцінки, академічні ранги.
Інтервальна шкала	
Існує одиниця вимірювання, за допомогою якої предмети можна не лише впорядкувати, але й приписати їм числа так, щоб однакові різниці чисел, присвоєних предметам, відображували однакові відмінності у кількостях вимірюваної ознаки. Нульова точка інтервальної шкали жовільна і не вказує на відсутність властивості.	Календарний час, шкали температур за Фаренгейтом та Цельсієм.
Шкала відношень (абсолютна)	
Числа, присвоєні об'єктам мають усі властивості об'єктів інтервальної шкали, та й ще крім того, на шкалі визначено абсолютний нуль. Значення нуля свідчить про	Зріст, вага, час, температура за Кельвіном.

відсутність оцінюваної властивості. Відношення чисел відображують кількісні відношення вимірюваної властивості.	

Квантилі розподілу випадкової величини

Квантилями (в Excel ПЕРСЕНТИЛЬ) називають значення випадкової величини, що є розв'язками рівняння $F_n(x)=p$, де $F_n(x)$ – вибіркова функція розподілу випадкової величини, а p – відносне положення аргумента x у вибірці. При p=0,5 маємо медіану, при p=0,25, 0,5, 0,75, 1 – квартилі, при p=0,01, 0,02, … 0,99, 1 – процентилі.

У геометричній інтерпретації квантиль – це абсциса точки, в якій площа під графіком щільності розподілу імовірностей над віссю абсцис дорівнює p.

Завдання 1 (MS Excel)

1. Створити робочу книгу "**Розподіл**". На аркуші "Вибірки" ввести у стовпчик експериментальні дані з вибірок A та B (див. Вибірки, с.194).
2. Скопіювати введені дані на аркуш "**Параметри**" (для зручності роботи можна приховати середні рядки, залишаючи видимими початок та кінець вибірки).
3. Обчислити кількість спостережень, мінімальне та максимальне, використовуючи стандартні функції MS Excel. Обчислити розмах вибірки за формулою: Розмах=максимальне – мінімальне.
4. Обчислити середні значення, моду, медіану, дисперсію, середнє квадратичне відхилення, асиметрію та ексцес для кожної вибірки.
5. Розрахувати похибки обчислення відповідно середнього, асиметрії та ексцеса за формулами:

$$похибка = \frac{S}{\sqrt{n}}\ (^3),\ m_A = \sqrt{\frac{6}{n}}\ \text{ та }\ m_E = 2\cdot\sqrt{\frac{6}{n}}\ .$$

[3] Тут S – стандартне квадратичне відхилення, n – кількість спостережень у вибірці. Якщо обсяг генеральної сукупності (N) відомий, то похибку обчислюють за формулою: $похибка = \frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$.

6. Скопіювати введені дані на аркуш “Обчислення”, використавши команду Правка->Специальная вставка –> Связь (таким чином між даними аркушів “Вибірки” та “Обчислення” збережеться односторонній зв’язок: зміни даних на аркуші “Вибірки” відбиватимуться на аркуші “Обчислення”).
7. На аркуші **“Обчислення”** з даних вибірки В сформувати систематичну вибірку B_c з періодом 5 (скориставшись послугою Выборка з пакета Анализ данных (див. примітку на с. 11)) та обчислити для неї за допомогою функцій MS Excel та за формулами відповідно
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \quad S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$
середнє, дисперсію і стандартне квадратичне відхилення.
8. Відсортувати дані вибірки Вс за зростанням.
9. Проранжувати вибірку Вс за допомогою функції РАНГ (порядок ранжування = 1).
10. *У сусідньому стовпці внести поправки у ранжування згідно до порядку ранжування. Перевірити правильність ранжування за формулою: $\sum R_i = \frac{N \cdot (N+1)}{2}$.
11. Визначити ПРОЦЕНТРАНГ деякого значення з вибірки А, тобто відносне положення даного значення у вибірці з припущення, що максимальному значенню відповідає 1, а мінімальному – 0. Процентний ранг=$\frac{Ранг - 1}{n-1}$, де n – обсяг вибірки. У функції ПРОЦЕНТРАНГ(Массив, X, Значимость) – Массив – це вибірка, X – значення, для якого визначають процент, Значимость – кількість значущих цифр у відповіді [15, с.45].
12. Виконати обернену операцію: визначити значення, якому відповідає обчислений процентний ранг, тобто обчислити X=ПЕРСЕНТИЛЬ(Массив, К), де Массив – задана вибірка, К – значення персентиля у межах від 0 до 1. За формулою $X = x_н + i\left(\frac{n \cdot K - p_s}{p}\right)$, де $x_н$ – нижня границя класу, що

містить X; p – частота класу, що містить X; i – величина класового інтервалу; p_s – накопичена частота попереднього класу; n – обсяг вибірки; K – персентиль [15, с.47; 5, с.62].

Примітка: *знайомство з функціями ПРОЦЕНТРАНГ та ПЕРСЕНТИЛЬ можна відкласти до теми 5 (с. 69)*.

13. Визначити значення, якому відповідає персентиль K=0,5 (тобто медіану). Порівняти з результатом застосування функції МЕДИАНА.
14. Визначити значення першого, другого та третього квартилів вибірки A, використовуючи функцію Квартиль(Массив, Значение), де Значение – номер квартиля (0, 1, 2, 3, 4).
15. На аркуші "**Розподіл**" побудувати ряд розподілу частот за допомогою функції ЧАСТОТА[4]. Поекспериментуйте з різною кількістю класових інтервалів: побудуйте принаймні 3 ряди розподілу – з малим, середнім та великим значенням класового інтервалу.
16. Для кожної частотної таблиці розрахувати і записати у сусідні стовпці відносну частоту, накопичену частоту та відносну накопичену частоту. Застосувати до відносних частот ПРОЦЕНТНИЙ формат.
17. До кожної частотної таблиці побудувати гістограму розподілу та кумуляту (до кожної таблиці окремий аркуш з графіками).
18. Зробити висновки про вид розподілів вибірок A та B. Проінтерпретувати отримані статистичні показники. Який спосіб розбиття виявився найбільш наочним? Запишіть свої висновки у примітку.
19. До даних вибірки B застосувати процедуру Описательная статистика з пакета Анализ Данных[5].

---

[4] Пам'ятайте, що ЧАСТОТА – функція-масив. Для отримання правильного результату її введення слід завершувати натисканням комбінації клавіш **CTRL+Shift+Enter.**

[5] Анализ данных – це набір додаткових статистичних процедур, що викликаються з меню СЕРВИС. Якщо пункта Анализ данных немає, то слід підключити надбудову ПАКЕТ АНАЛИЗА за допомогою послуги меню СЕРВИС –> НАДСТРОЙКИ. (Анализ данных – це не обов'язковий інструмент, який підключається лише за вимогою користувача).

Завдання 2 (SPSS)

1. Завантажити пакет SPSS через меню ПРОГРАМИ, або запустивши файл C:\Program Files\SPSS\spsswin.exe.
2. Завантажити довільний файл *.sav, наприклад, файл Employee data.sav, з папки SPSS (File –> Open –> Data). Ознайомитися зі структурою файлу, параметрами змінних (тип, вид шкали, мітка, значення номінативної шкали), параметрами форматованого виведення даних (ширина стовпця, вирівнювання).
3. Застосувати до змінних educ (рівень освіти у роках) та jobcat (категорія служби) процедури описової статистики пакету SPSS: Analyze –> Descriptives та Analyze –> Frequencies.
4. Зберегти звіт у форматі звіту SPSS та HTML.
5. Завантажити в SPSS дані з книги Розподіл.xls. Для цього слід, по-перше, виконати команду File –> Open –> Data, по-друге, вказати для завантажуваного файлу тип Excel, по-третє, вказати назву файла та аркуша з даними. Застосувати до них процедури описової статистики (Descriptives) пакету SPSS. Порівняти результати, отримані в SPSS та MS Excel. Побудувати за досліджуваними даними гістограму з нормальною кривою (Histogram) та стовпцеві діаграми за частотою та відносною частотою. Зберегти звіт у форматі пакету SPSS та HTML.
6. Відсортувати дані за зростанням, застосувавши процедуру Data –> Sort Cases. Вибрати змінну, за якою сортувати (Sort by:) та вказати порядок сортування (Ascending – за зростанням, або Descending – за спаданням).
7. Проранжувати значення змінної, застосувавши процедуру Transform –> Rank cases. У діалоговому вікні вибрати змінну для ранжування, вказати порядок ранжування, надавши найменшому значенню ранг 1 (Assign rank 1 to smallest value) та порядок нарахування рангів однаковим значенням змінної (Ties –> Rank Assigned to Ties вибрати Mean – обчислювати середнє арифметичне можливих рангів[6]). Результатом буде нова змінна.

[6] Див. сторінку 7.

Завдання 3 (Statistica):

1. Завантажити пакет Statistica запуском файла C:\Program Files\StatSoft\STATISTICA6\statist.exe або через меню ПРОГРАМИ.
2. Завантажити довільний приклад з папки C:\…\STATISTICA6\Examples\Datasets, наприклад файл job_prof.sta, скориставшись командою File –> Open головного меню програми.
3. Ознайомитися із структурою робочого вікна програми Statistica. Переглянути для кожної змінної її властивості, скориставшись командою Variable Specs з контекстного меню до заголовку змінної.
4. Ознайомитися з набором інструментів дескриптивної статистики (Statistica –> Basic Statistics –> Descriptive Statistics): для однієї із змінних прикладу застосувати послідовно усі процедури закладки Quick (Summary: Descriptive statistics, Frequency tables, Histograms, Bow&whisker plot). Ознайомитися із структурою, створеною в результаті застосування обчислювальних процедур робочої книги (Workbook). Зберегти робочу книгу.
5. На закладці Advanced вибрати необхідні параметри описової статистики та запустити процедуру обчислення (Summary: Descriptive statistics). Транспонувати отриману таблицю командою Data –> Transpose –> File головного меню. Зберегти таблицю у файл звіту. Для цього виконати команду контекстного меню виділеного об'єкту Extract as stand-alone window (виділити в окреме вікно) та Add to report (додати до звіту) головного меню.
6. Застосувати процедури Frequency tables та Histograms закладки Normality, встановивши необхідну кількість класових інтервалів (Number of intervals). Додавити результати їх застосування до створеного у попередньому пункті звіту. Зберегти звіт у форматах пакету Statistica, RTF та HTML.
7. Зберегти файл з даними у форматі SPSS.
8. Завантажити дані з робочої книги Розподіл.xls. Застосувати необхідні процедури описової статистики, побудувати необхідні графіки до однієї з вибірок. Створити та зберегти звіт.

9. Впорядкувати значення змінної за зростанням: виділити заголовок змінної та застосувати процедуру Data –> Sort.
10. Проранжувати змінну. Оскільки Statistica не створює нової змінної, як SPSS, а замінює значення рангами, бажано спершу скопіювати ранжовану змінну (команда Copy Variables контекстного меню), та застосувати до копії команду Data –> Rank.
11. Порівняти інструментарій декількох статистичних пакетів. Зробити висновки.

	A	B	C	D	E	F
1		Вибірка A	Вибірка B	Bc	Ранг Excel	Ранг
2		98	96	109	1	1,5
3		110	109	109	1	1,5
4		110	101	123	3	3
5		111	109	125	4	4
6		120	112	148	5	5
7		125	123			
8		142	140			
9		142	125			
10		145	140		Правильна сума рангів	
11		150	148			
12	Кількість	10	10	5		
13	Сума	1253	1203	614	14	15
14	Мінімальне	98	96			
15	Максимальне	150	148			
16	Розмах	52	52			
17	середнє	125,300	120,300	122,800		
18	сер. Геом.	124,093	119,124			
19	сер. Гарм.	122,886	117,970			
20	Урізане	125,300	120,300			
21	Мода	110,000	109,000			
22	Медіана	122,500	117,500			
23	Дисперсія	333,567	317,789	255,200		
24	Відхилення	18,264	17,827	15,975		
25	Похибка сер.	5,776	5,637	7,144		
26	Асиметрія	0,033	0,286			
27	Похибка Ас.	0,775	0,775			
28	Ексцес	-1,621	-1,281			
29	Похибка Екс.	1,549	1,549			

Рис. 1

Приклад виконання

На Рис. 1 у стовпчиках B та C наведено вибірки A та B з результатами обчислення параметрів розподілів за допомогою функцій MS Excel. Повний перелік використаних у лабораторних роботах функцій для *MS Excel* та *OpenOffice.org Calc* наведено у глосарії (див. с. 188).

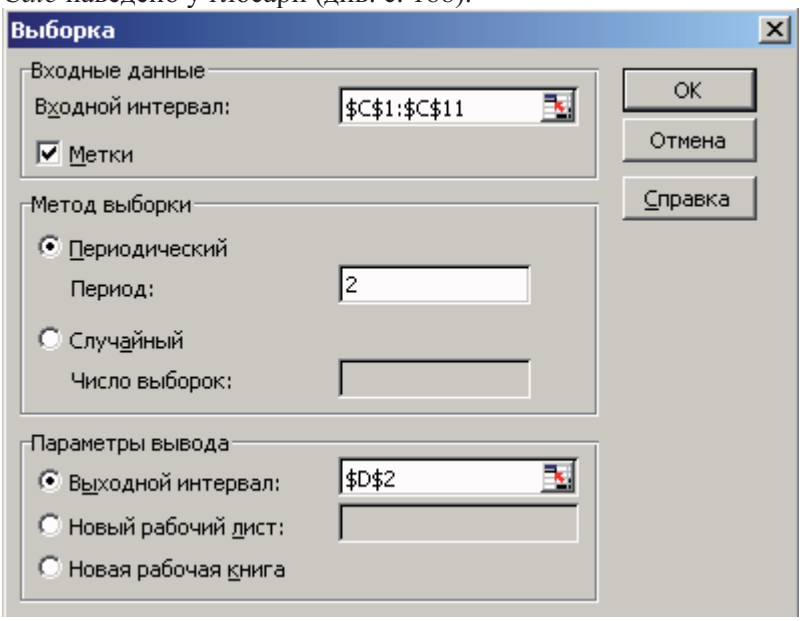
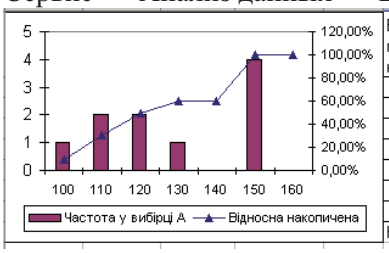


Рис. 2

У стовпці D наведено сформовану за даними вибірки B систематичну вибірку з періодом 2 (тобто до вибірки Bc відібрано кожен 2-й елемент вибірки B)[7]. Вікно процедури Сервис –> Анализ данных –> Выборка наведено на Рис. 2.



Верхня границя класу	Частота у вибірці A	Відносна	Накопичена	Відносна накопичена
100	1	10,00%	1	10,00%
110	2	20,00%	3	30,00%
120	2	20,00%	5	50,00%
130	1	10,00%	6	60,00%
140	0	0,00%	6	60,00%
150	4	40,00%	10	100,00%
160	0	0,00%	10	100,00%
Кількість	10			

Рис. 3

[7] Про використання процедур ПАКЕТА АНАЛІЗА див. примітку на сторінці 11.

На Рис. 3 подано результати застосування функції ЧАСТОТА (стовбець Частота у вибірці А), а також графічне представлення інтервального статистичного ряду розподілу у вигляді гістограми частот та кумуляти (графіка відносних накопичених частот).

Результати застосування функцій ПРОЦЕНТРАНГ та ПЕРСЕНТИЛЬ до вибірки А показано на Рис. 4, а на Рис. 5 – формати введення зазначених функцій у пакеті MS Excel.

	B	G	H	I	J	K
1	Вибірка А	Ранг А	Процентранг	(R-1)/(n-1)	Персентиль	формула
2	98	1	0,00	0,00	98,00	98,00
3	110	2	0,10	0,11	108,80	108,07
4	110	2	0,10	0,11	108,80	108,07
5	111	4	0,30	0,33	110,70	111,56
6	120	5	0,40	0,44	116,40	118,89
7	125	6	0,50	0,56	122,50	125,11
8	142	7	0,60	0,67	131,80	140,67
9	142	7	0,60	0,67	131,80	140,67
10	145	9	0,80	0,89	142,60	145,78
11	150	10	1,00	1,00	150,00	150,00

Рис. 4

Ранг А	Процентранг	Персентиль
=РАНГ(B2;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B2;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H2)
=РАНГ(B3;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B3;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H3)
=РАНГ(B4;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B4;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H4)
=РАНГ(B5;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B5;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H5)
=РАНГ(B6;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B6;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H6)
=РАНГ(B7;$B$2:$B$11;1)	=ПРОЦЕНТРАНГ($B$2:$B$11;B7;1)	=ПЕРСЕНТИЛЬ($B$2:$B$11;H7)

Рис. 5

Контрольні запитання

1. Дайте означення моди, дисперсії, стандартного квадратичного відхилення, асиметрії, ексцеса.
2. Назвіть та запишіть формули обчислення відомих вам середніх.
3. Як обчислити похибку середнього?
4. Які властивості має середнє? Як їх можна застосувати при обчисленнях?
5. В яких випадках доцільно застосовувати середнє арифметичне, середнє гармонійне, середнє геометричне?

6. Які властивості дисперсії? Як їх можна застосувати при обчисленнях?
7. Як визначити величину класового інтервалу? На скільки класів рекомендується розбивати експериментальну вибірку?
8. Який порядок побудови інтервального статистичного ряду розподілу частот?
9. Як обчислити відносну та відносну накопичену частоту?
10. Як графічно можна представити вибірку?
11. Що таке персентиль, квартиль? Що таке медіана? Показниками чого вони є?
12. Який порядок ранжування даних?
13. Як виконати ранжування даних у пакетах MS Excel, OpenOffice.org Calc, SPSS, Statistica?
14. Які статистичні методи відносять до описової статистики?
15. Що таке вибірка?
16. Які шкали вимірювання використовують у статистиці?
17. Наведіть приклади величин, виміряних за різними шкалами.
18. Які арифметичні операції дозволено виконувати з даними, виміряними за номінативною шкалою?
19. Зібрано відомості про розмір взуття 100 студентів першого курсу. В яких випадках шкалу вимірювання даної ознаки слід вважати інтервальною, порядковою, номінативною?
20. З якими форматами даних можна працювати у пакетах MS Excel, OpenOffice.org Calc, SPSS, Statistica?
21. У яких форматах можна зберігати дані у пакетах MS Excel, OpenOffice.org Calc, SPSS, Statistica?
22. Як створювати звіти у пакетах SPSS, Statistica? У яких форматах їх можна зберігати?

Тема 2:
“Порівняння параметрів двох вибірок”

Мета:

Студенти повинні знати:

– критерії та умови порівняння параметрів двох незалежних вибірок;
– критерії та умови порівняння двох залежних вибірок;
– формули для обчислення статистичних параметричних критеріїв порівняння середніх (t-критерій Стьюдента) та дисперсій (F-критерій Фішера);
– правила визначення кількості степенів вільності;
– умови застосовності та обмеження відповідних статистичних критеріїв;
– правила прийняття рішення за критичними значеннями та рівнем похибки;
– особливості та порядок застосування відповідних процедур Пакета Аналізу;
– порядок підключення Пакету Аналізу.

Студенти повинні уміти:

– подавати експериментальні дані у вигляді, зручному для обчислень;
– застосовувати процедури Пакету Аналізу до експериментальних даних;
– будувати необхідні графіки;
– формувати статистичні гіпотези та робити статистичні висновки;
– інтерпретувати результати обчислювальних процедур Пакету Аналізу;
– застосовувати стандартні функції MS Excel для визначення критичних значень статистичних критеріїв;
– застосовувати відповідні процедури пакетів SPSS та Statistica.

Теоретичні відомості

Перш ніж порівнювати середні двох вибірок, слід з’ясувати, чи однорідні (рівні) в них дисперсії. При однорідності дисперсій (гомоскедатичності) кількість степенів вільності для порівняння середніх визначають як $df=n_1+n_2-2$. Якщо дисперсії порівнюваних вибірок достовірно різні, то кількість степенів вільності визначають інакше та використовують менш потужний критерій порівняння.

Порівняння дисперсій найчастіше здійснюють за критерієм Фішера-Снедекора: $F = \frac{S_1^2}{S_2^2}$, де $S^2_1 > S^2_2$ – дисперсії порівнюваних вибірок (F завжди має бути більшим за одиницю). Критичне значення залежить від степенів вільності для порівнюваних вибірок, відповідно $df_1 = n_1 - 1$, $df_2 = n_2 - 1$ (n_1 та n_2 – обсяги порівнюваних вибірок).

Статистичні гіпотези при перевірці дисперсій напрямлені, тому для прийняття рішення застосовують односторонній критерій. Дисперсії вважають різними (гіпотеза H_0 відхиляється), коли $F_{емп} > F_{кр}$, або коли отримане *p*-значення менше за α (прийнятий рівень значущості).

Критерій F може бути коректно застосований лише до нормально розподілених вибірок, тому у деяких статистичних пакетах використовують менш вибагливий до виду розподілу критерій Левена (Leven's Test). Рішення приймається, як і при застосуванні критерія F.

Порівняння середніх виконують за допомогою t-критерія Стьюдента: $t = \frac{\bar{x}_1 - \bar{x}_2}{S_d}$, де $\bar{x}_1$ та $\bar{x}_2$ відповідно середні двох вибірок, а S_d – похибка обчислення різниці середніх. Вона обчислюється для ***зв'язаних*** вибірок за формулою $S_d = \sqrt{\frac{\sum d_i^{\,2} - \frac{\left(\sum d_i\right)^2}{n}}{n \cdot (n-1)}}$ (тут d_i – різниці між парами даних). Та для ***незв'язаних*** вибірок за формулою $S_d = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$, – для вибірок приблизно однакових за обсягом, або $S_d = \sqrt{S^2 \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}$, де $S^2 = \frac{\sum (X - \bar{x}_1)^2 + \sum (X - \bar{x}_2)^2}{n_1 + n_2 - 2}$ – об'єднана дисперсія, для різних за обсягом вибірок.

Нульову гіпотезу відхиляють, коли емпіричне значення більше критичного для кількості степенів вільності $df = n_1 + n_2 - 2$.

Завдання 1: порівняння параметрів двох незалежних вибірок.

1. В MS Excel у робочій книзі "**Стьюдент**" на аркуші "**t-тест**" сформувати з вибірки В дві вибірки значень обсягом 30-40 чисел (бажано щоб сформовані вибірки мали неоднаковий обсяг). Придумати статистичний сюжет до отриманих даних. Записати його у примітку до комірки А1.
2. Переконатися у підключенні Пакета Аналізу (пункт Сервис).
3. Обчислити дисперсії двох вибірок, з'ясувати, яка з них більша при простому порівнянні. Сформувати статистичні гіпотези стосовно порівнюваних дисперсій.
4. Виконати перевірку на рівність дисперсій двох вибірок, застосувавши Двухвыборочный F-тест для дисперсий (критерій Фішера-Снедекора) з Пакету Аналізу (Сервіс–>Аналіз даних). Зробити висновки про дисперсії двох вибірок.
5. Виконати порівняння дисперсій за допомогою стандартних функцій MS Excel:
   – визначити за функцією ФТЕСТ(масив1; масив2) двосторонню імовірність схожості двох сукупностей (2P);
   – визначити степені вільності: $k_1=n_1-1$ та $k_2=n_2-1$ (тут n_1 – кількість спостережень у вибірці з більшою дисперсією, n_2 – кількість спостережень у вибірці з меншою дисперсією);
   – отримати значення односторонньої імовірності P=2P/2;
   – визначити експериментальне значення критерію F, що відповідає отриманій імовірності за функцією FРАСПОБР(P; степені_вільності1; степені_вільності2);
   – визначити критичні значення F для рівнів значущості p=0,05 та p=0,01.
6. Порівняти результати з результатами застосування процедури з Пакета Аналізу.
7. Сформувати статистичні гіпотези стосовно середніх двох порівнюваних вибірок. Виконати для вибірок порівняння середніх: вибрати процедуру Двухвыборочный t-тест с одинаковыми дисперсиями або Двухвыборочный t-тест с различными дисперсиями, залежно від результатів F-тесту (див. п. 4). Зробити статистичні висновки.
8. Отримати критичні значення t, застосувавши функцію СТЬЮДРАСПОБР(p, k), де ***p*** – це обраний рівень значущості, а ***k*** – степені вільності ($k=n_1+n_2-2$).
9. Отримати імовірність похибки першого роду для обчисленого значення t (за пунктом 6), застосувавши

функцію СТЬЮДРАСП(X, k, m), де X – обчислене у t-тесті значення критерію Стьюдента, k – степені вільності, m – кількість хвостів розподілу (m=1 для одностороннього критерію, m=2 для двостороннього критерію).

10. Скопіювати дані з книги MS Excel таким чином: створити дві змінні, до першої змінної скопіювати досліджувані значення, а до другої записати номер вибірки (1 або 2). Поділ на вибірки у пакеті SPSS буде здійснено автоматично за значеннями другої змінної.
11. Застосувати до введених даних процедуру SPSS Analyze –> Compare Means –> Independent Samples t-test. Вказати першу змінну як досліджувану (Test Var), а другу як групувальну (Grouped Var) та визначити її значення для різних груп (1 – для першої, 2 – для другої).
12. Проаналізувати отримані результати, зробити висновки.
13. Виконати порівняння середніх засобами пакету Statistica. Для даних, поданих у двох стовпцях (як в MS Excel), слід застосувати процедуру Basic statistics–>t-test, independent, by variables. Для даних, поданих як в SPSS (в один стовпчик + групувальна змінна), застосовують процедуру Basic statistics –> t-test, independent, by groups.
14. Порівняти результати, отримані у трьох пакетах.

Завдання 2: порівняння параметрів двох залежних вибірок.

1. На аркуші "**Парний t-тест**" ввести значення для парного тесту Стьюдента (спряжені дані). Варіанти наведено у додатку (с. 202).
2. Придумати статистичний сюжет та вставити його як примітку в комірку A1.
3. Застосувати процедуру Парный двухвыборочний t-тест для средних з Пакету Аналізу. Зробити статистичні висновки.
4. До тих самих даних застосувати Двухвыборочный t-тест. Порівняти результати з результатами парного тесту. Зробити висновки.
5. Для тих самих даних застосувати процедуру Analyze –> Compare Means –>Pared-Samples T Test пакета SPSS.
6. Ознайомитися з порядком застосування та настройками процедури Statistics –> Basic statistics –> t-test, dependent samples пакету Statistica.

7. Сформувати та зберегти звіт про виконання тесту, у якому зберегти вхідні дані та результати t-теста. (Для перенесення результатів застосування процедури до звіту, по-перше, слід виділити об'єкт робочої книги в окреме вікно застосуванням команди Extract as stand-alone window з контекстного меню, по-друге, додати до звіту, натиснувши кнопку Add to report на панелі інструментів.

Приклад виконання

Приклад 1. За методикою Д.Векслера у студентів двох факультетів вимірювали рівень невербального інтелекту. Результати представлено на таблиці. Чи можна стверджувати, що в одній з вибірок рівень інтелекту вищий, ніж у другій? [19, с.45]

Група 1	Група 2
111	113
104	107
107	123
90	122
115	117
107	112
106	105
107	108
95	111
116	114
127	102
115	104
102	
99	

У пакеті MS Excel спочатку слід виконати перевірку на гомоскедатичність (однорідність дисперсій) за критерієм Фішера-Снедекора.

Статистичні гіпотези будуть такі:

H_0: дисперсія у групі 1 не більша за дисперсію у групі 2.

H_1: дисперсія у групі 1 більша за дисперсію у групі 2.

Примітка: *Спочатку слід визначити, в якій групі дисперсія більша при простому порівнянні, оскільки значення критерія F завжди має бути більшим за 1 (одиницю). У даному випадку більшою виявилася дисперсія у групі 1 (88,95>45,73).*

Застосування процедури Сервис –> Анализ Данных –> Двухвыборочный F-тест для дисперсий дасть такий результат:

Двухвыборочный F-тест для дисперсии			
(5%)	Група 1	Група 2	
Среднее	107,21	111,5	– середні для вибірок
Дисперсия	88,95	45,72	– дисперсії для вибірок
Наблюдения	14	12	– кількості спостережень

df	13	11	– кількості степенів вільності
F	1,94		– емпіричне (обчислене) значення F
P(F<=f) одностороннее	0,13		– p-значення
Fкритическое одностороннее	2,761		– критичне значення F

Гіпотези у критерії Фішера напрямлені, тому і критерій односторонній. Гіпотезу H_0 відхиляють, коли $F_{емп}>F_{кр}$. У даному випадку цього зробити не можна, оскільки 1,94<2,76, тобто слід прийняти нульову гіпотезу і визнати, що дисперсії двох вибірок відрізняються випадково (тобто статистично однакові).

Про це ж свідчить і p-значення – імовірність похибки відхилити нульову гіпотезу, коли вона правильна. У різних експериментах приймають H_0, коли p-значення > α (встановлений рівень значущості), і відхиляють H_0, коли p-значення < α. У даному випадку P=0,13, тобто більше за 0,05.

Далі можна переходити до, власне, порівняння середніх. Як з'ясувалося, з двох можливих процедур у даному випадку слід застосувати Сервис –> Анализ Данных –> Двухвыборочный t-тест с одинаковыми дисперсиями.

Результат застосування буде таким:

Двухвыборочный t-тест с одинаковыми дисперсиями		
5%		
	Група 1	Група 2
Среднее	107,2143	111,5
Дисперсия	88,95055	45,727273
Наблюдения	14	12
Объединенная дисперсия	69,13988	
Гипотетическая разность средних	0	
df	24	
t-статистика	-1,31017	
P(T<=t) одностороннее	0,10127	
t критическое одностороннее	1,710882	
P(T<=t) двухстороннее	0,202541	
t критическое двухстороннее	2,063899	

Тут крім групових середніх та дисперсій визначено об'єднану дисперсію та гіпотетичну різницю середніх.

Об'єднана дисперсія обчислюється за відповідною формулою (див. с. 19). А гіпотетична різниця середніх – це припущення про її величину. Найчастіше перевіряють рівність середніх, тобто припущення про те, що різниця середніх дорівнює нулю.

Критичне значення t та p-значення подано у двох варіантах: одностороннє та двостороннє. Перший варіант застосовують до напрямлених гіпотез. У даному прикладі напрямлені гіпотези будуть такі:

H_0: середнє групи 1 **не** менше за середнє групи 2 (або середнє групи 2 **не** більше за середнє групи 1).

H_1: середнє групи 1 менше за середнє групи 2 (або середнє групи 2 більше за середнє групи 1).

Для прийняття рішення абсолютне значення обчисленого t порівнюють з одностороннім критичним. Тут $|t_{емп}| < t_{кр}$, тобто нульову гіпотезу відхилити не можна.

При дослідженні ненапрямлених гіпотез абсолютне значення емпіричного t порівнюють з двостороннім критичним.

У даному випадку ненапрямлені гіпотези будуть такі:

H_0: середнє групи 1 не відрізняється від середнього групи 2 (або $\bar{x}_1 - \bar{x}_2 = 0$).

H_1: середнє групи 1 відрізняється від середнього групи 2 (або $\bar{x}_1 - \bar{x}_2 \neq 0$).

За правилом прийняття рішення слід визнати різницю середніх випадковою, тобто прийняти нульову гіпотезу.

У пакеті SPSS перевірка на гомоскедатичність здійснюється у процедурі порівняння середніх. Результати її застосування будуть представлені у такому вигляді:

У таблиці Таблиця 1 наведено основні групові статистики: кількість спостережень, середнє, стандартне квадратичне відхилення та похибка обчислення середнього.

Таблиця 1

Group Statistics

	VAR00002	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	1,00	14	107,2143	9,4314	2,5206
	2,00	12	111,5000	6,7622	1,9521

У таблиці Таблиця 2 наведено, по-перше, результати теста Левена на гомоскедатичність (Leven's Test for Equality of Variance). Р-значення, що відповідає обчисленому значенню F=0,573, вказує на рівність дисперсій (Sig. = 0,471), бо воно більше за α. Отже емпіричному значенню t-критерія відповідатиме значення з рядка Equal variances assumed (припускається, що дисперсії рівні). У протилежному випадку слід буде брати до уваги значення t-критерія з другого рядка (Equal variances not assumed) та додатково використовувати непараметричні критерії порівняння двох вибірок.

Таблиця 2

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
VAR01	Equal variances assumed	,537	,471	-1,310	24	,203	-4,2857	3,2711	-11,0370	2,4655
	Equal variances not assumed			-1,344	23,345	,192	-4,2857	3,1881	-10,8755	2,3041

Отже у даному випадку дисперсії можна вважати однаковими, та скористатися більш потужним критерієм, значення якого у даному випадку t = 1,31 (за абсолютною величиною). Йому відповідає двостороннє p-значення = 0,203, тобто слід прийняти нульву гіпотезу (ненапрямлену). Одностороннє p-значення буде вдвічі менше (p=0,102). Тобто і для випадку напрямленої гіпотези правильною буде H_0.

У пакеті Statistica результати буде подано аналогічно:

	Mean	Mean	t-value	df	p	Valid N	Valid N	Std.Dev.	Std.Dev.	F-ratio	p
Var1	107,21	111,50	-1,31	24	0,20	14	12	9,43	6,76	1,94	0,28

Приклад 2 (парний t-тест): Група з 10 школярів протягом літніх канікул перебувала у спортивному таборі. До і після канікул у них визначали місткість легенів (у міліметрах). За результатами вимірювань необхідно визначити, чи істотно змінився цей показник під впливом інтенсивних фізичних вправ.

Вхідні дані будуть представлені у таблиці MS Excel у двох стовпцях, причому дані кожного рядка представляють собою результати двох замірів для одного учня, тобто “кортеж”.

Далі слід з пункту головного меню MS Excel Сервис → Анализ данных вибрати пункт “Парный двухвыборочный t-тест для средних”.

	A	B	C
1		X	Y
2	учень1	3400	3800
3	учень2	3600	3700
4	учень3	3000	3300
5	учень4	3660	3600
6	учень5	2900	3100
7	учень6	3100	3200
8	учень7	3200	3200
9	учень8	3400	3300
10	учень9	3200	3500
11	учень10	3400	3600

Для наведеного прикладу у вікні майстра парного тесту слід встановити наступні параметри (Рис. 6).

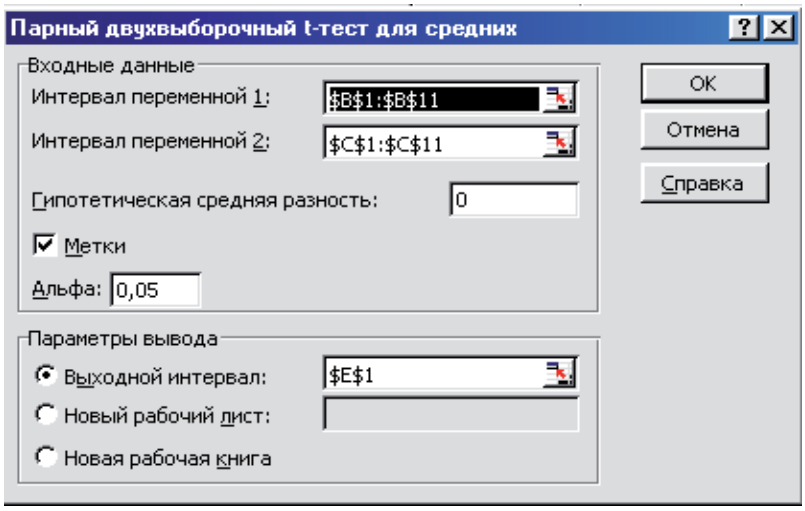


Рис. 6

У таблиці відповідно отримаємо:

E	F	G
Парный двухвыборочный t-тест для средних		
	X	Y
Среднее	3286	3430
Дисперсия	61960	57888,88889
Наблюдения	10	10

Корреляция Пирсона	0,772158142	
Гипотетическая разность средних	0	
df	9	
t-статистика	-2,752988806	
P(T<=t) одностороннее	0,011183595	
t критическое одностороннее	1,833113856	
P(T<=t) двухстороннее	0,022367189	
t критическое двухстороннее	2,262158887	

Тут маємо:

t-статистика – обчислене значення критерію (враховувати абсолютне значення!);

t критическое двухстороннее – критичне значення t для визначеного рівня значущості (у даному випадку 0,05 – див. Рис. 6)

Критичне значення визначається за двостороннім критерієм у тому разі, коли статистична гіпотеза ненапрямлена, тобто з'ясовується достовірністьт або недостовірність різниці середніх як такої.

Оскільки 2,75>2,26 ($t_{емп}>t_{кр}$), то гіпотезу H_0 слід відкинути і визнати, що місткість легенів під впливом фізичних вправ змінилася не випадково.

Про це ж свідчить і отримане значення імовірності похибки:

P(T<=t) двухстороннее=0,022367189, тобто значно менше за 0,05 – прийнятий рівень значущості.

Якщо статистична гіпотеза напрямлена, тобто у гіпотезі припускається, що середнє однієї з вибірок не більше (не менше) – нульова гіпотеза, – або більше (менше) – альтернативна гіпотеза, – ніж в іншій вибірці, то критичне значення визначається за одностороннім критерієм. У даному випадку, $t_{емп}>t_{кр}$ (2,75>1,83), тобто слід прийняти гіпотезу H_1 про те, що середнє вибірки X достовірно *менше* за середнє вибірки Y.

У пакеті SPSS результати парного t-тесту (Analyze –> Compare Means –>Pared-Samples T Test) для даних прикладу будуть такі:

Таблиця 3

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	X	3286,0000	10	248,9177	78,7147
	Y	3430,0000	10	240,6011	76,0847

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	X & Y	10	,772	,009

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	X - Y	144,0000	165,4086	52,3068	-262,3262	-25,6738	-2,753	9	,022

Крім значення t-критерію додатково обчислюється коефіцієнт кореляції Пірсона. У даному випадку кореляція достовірна на рівні значущості α=0,01 (p-значення = 0,009), а різниця середніх достовірні на рівні значущості α=0,05 (p-значення = 0,022).

У пакеті Statistica результат застосування процедури Statistics –> Basic statistics –> t-test, dependent samples матимуть такий вигляд:

	Mean	Std.Dv.	N	Diff.	Std.Dv.	t	df	p
X	3286,000	248,9177						
Y	3430,000	240,6011	10	-144,000	165,4086	-2,75299	9	0,022367

Контрольні запитання

1. Що таке статистична гіпотеза?
2. Які бувають статистичні гіпотези?
3. Що таке помилка першого роду? Яка її імовірність?
4. Що таке помилка другого роду? Яка її імовірність?
5. Який критерій називають одностороннім?
6. Призначення критерію Фішера-Снедекора.
7. Що таке гомоскедатичність?

8. Як визначити кількість степенів вільності для критерія Фішера?
9. Як формулюються статистичні гіпотези для критерія Фішера?
10. За якою критичною областю (одно- чи двосторонньою) робиться висновок?
11. Який порядок порівняння середніх?
12. Як приймається рішення для критерія Стьюдента?
13. В яких випадках застосовують парний тест Стьюдента?
14. Які обмеження до застосування критеріїв Фішера та Стьюдента?

Тема 3: “Дисперсійний аналіз”

Мета:

Студенти ***повинні знати***:
– критерії порівняння параметрів двох незалежних вибірок;
– критерії порівняння трьох та більше незалежних вибірок;
– призначення однофакторного дисперсійного аналізу;
– критерії порівняння двох та більше залежних вибірок;
– призначення двофакторного дисперсійного аналізу;
– особливості графічного подання даних для дисперсійного аналізу;
– склад Пакета Аналізу;
– порядок застосування процедур Пакета Аналізу;
– порядок підключення Пакету Аналізу.

Студенти ***повинні уміти:***
– подавати експериментальні дані у вигляді, зручному для обчислень;
– застосовувати процедури Пакету Аналізу до експериментальних даних;
– будувати необхідні графіки
– інтерпретувати результати обчислювальних процедур пакету аналізу;
– застосовувати стандартні функції MS Excel для визначення критичних значень статистичних критеріїв
– порядок та особливості виконання дисперсійного аналізу засобами пакетів SPSS та Stitistica.

Теоретичні відомості

Дисперсійний аналіз (ДА) (в англомовній літературі – ANOVA – *ANalysis Of VAriance*) дозволяє оцінити відмінності між вибірковими середніми для довільної кількості вибірок. Суть метода полягає у розкладі дисперсій однієї або кількох змінних на складові, та оцінювання за допомогою F-критерія внеску цих складових до загальної варіації даних. Змінна, вплив якої вивчається у ДА називається *фактором* або *контрольованим фактором.* Якщо такий фактор один, то застосовують однофакторний ДА, якщо факторів багато, то – багатофакторний ДА.

В основу ДА покладено лінійну модель, відповідно до якої значення кожного елемента вибірки складається з вибіркового середнього (математичного сподівання), відхилення від вибіркового середнього для даної градації фактора та “похибки” лінійної моделі, тобто внеску унікальності конкретного

елемента: $x_{ij} = M_{виб} + (M_j - M_{виб}) + e_{ij}$. Тут $M_{виб}$ – математичне сподівання генеральної сукупності (його оцінкою є вибіркове середнє), M_j – середнє у групі, що відповідає j-тій градації фактора, e_{ij} – "похибка" лінійної моделі, тобто внесок окремого елемента вибірки.

Основна гіпотеза ДА стверджує, що всі вибірки сформовано з однієї генеральної сукупності, тобто їхні середні однакові.

ДА засновано на припущеннях про те, що 1) значення ознак, що відповідають кожному рівню контрольованого фактора, нормально розподілені навколо свого середнього; 2) дисперсії вибіркових розподілів, що відповідають кожному рівню контрольованого фактора, однорідні; 3) отримані спостереження незалежні. Однак дослідження показали, що ДА насправді мало чутливий до порушення нормальності та при великих обсягах вибірок майже однакового розміру є також стійким до неоднорідності дисперсій.

Таблиця однофакторного аналізу (у випадку незв'язаних вибірок) матиме вигляд як на таблиці Таблиця 4.

Тут k – кількість градацій фактора, n – кількість спостережень у кожній комірці дисперсійного комплекса. Висновок робиться на основі значення F: якщо $F_{emp} < F_{kr}$, то нульову гіпотезу не можна відхилити, тобто вплив контрольованого фактора слід вважати випадковим (неістотним).

Таблиця 4

Компоненти дисперсії	Суми квадратів Q	Степені вільності *df*	Дисперсії	F
Міжгрупова (Between groups)	$Q_{bg} = n \cdot \sum_{i=1}^{k} (\overline{X}_i - \overline{X})^2$	k–1	$S^2_{bg} = \frac{Q_{bg}}{k-1}$	$\frac{S^2_{bg}}{S^2_{wg}}$
В середині груп (Within groups)	$Q_{wg} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)^2$	k·(n–1)	$S^2_{wg} = \frac{Q_{wg}}{k \cdot (n-1)}$	
Загальна (Total)	$Q_{total} = \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \overline{X})^2$	k·n–1	$S^2_{total} = \frac{Q_{total}}{N-1}$	

Нульова гіпотеза стверджує, що відмінності між групами випадкові, а альтернативна – що відмінності між групами обумовлені впливом контрольованого фактора.

У разі, коли змінні вимірюються за порядковою шкалою, для одного фактора використовують непараметричні критерії: Н-критерій Крускала-Уолліса (незв'язані вибірки), χ^2_r-критерій Фрідмана (зв'язані вибірки) (див. лабораторну роботу №6, с. 115).

За однофакторним дисперсійним комплексом можна також дослідити достовірність нелінійної кореляції, показником якої є величина $\eta^2 = \frac{Q_{bg}}{Q_{total}}$ – кореляційне відношення (ета-квадрат). Достовірність коефіцієнта η^2 визначають так само, як і достовірність лінійної кореляції (див. с. 43).

За коефіцієнтом η^2 також визначають величину (силу) факторного впливу, тобто внеску незалежної змінної або взаємодії змінних у розсіювання залежної змінної: при відсутності факторного ефекту η^2 дорівнює 0, а при значному впливі коефіцієнт η^2 близький до 1 [23, 5].

Для двофакторної моделі (у разі незв'язаних вибірок) формується три пари гіпотез: про вплив першого фактора, дру гого фактора та їхню взаємодію. Розрахункові формули компонентів дисперсії для двофакторної моделі подано у таблиці Таблиця 5.

Лінійна модель двофакторного аналізу є розширенням однофакторної моделі (див. с. 30):

$$x_{ijg} = M_{виб} + (M_j - M_{виб}) + (M_g - M_{виб}) + a_{jg} + e_{ijg}.$$

Тут $M_{виб}$ – математичне сподівання генеральної сукупності (або його оцінка – вибіркове середнє), M_j – середнє у групі, що відповідає j-тій градації першого фактора, M_g – середнє у групі, що відповідає g-тій градації другого фактора, a_{jg} – вплив міжфакторної взаємодії, e_{ijg} – "похибка" лінійної моделі, тобто внесок окремого елемента вибірки.

При неврівноважених комплексах процедури обчислення дисперсій дещо ускладнюються [23, 18]. А у разі зв'язаних вибірок додатково перевіряється вплив фактора індивідуальних відмінностей [19, 23]. Опис моделі дисперсійного аналізу та відповідних процедур статистичних пакетів для зв'язаних вибірок буде наведено у темі 6 (див. с. 120). -

Таблиця 5

Компоненти дисперсії	Суми квадратів	Степені вільності	Дисперсії	F
По рядках (фактор А)	$Q_a = b \cdot n \cdot \sum_{i=1}^{a} (\overline{X}_{i**} - \overline{X})^2$	a–1	$S_a^2 = \frac{Q_a}{a-1}$	$\frac{S_a^2}{S_z^2}$
По стовпцях (фактор В)	$Q_b = a \cdot n \cdot \sum_{j=1}^{b} (\overline{X}_{*j*} - \overline{X})^2$	b–1	$S_b^2 = \frac{Q_b}{b-1}$	$\frac{S_b^2}{S_z^2}$
Взаємодія	$Q_{ab} = n \cdot \sum_{i=1}^{a}\sum_{j=1}^{b} (\overline{X}_{ij*} - \overline{X}_{i**} - \overline{X}_{*j*} + \overline{X})^2$	(a–1)·(b–1)	$S_{ab}^2 = \frac{Q_{ab}}{(a-1)(b-1)}$	$\frac{S_{ab}^2}{S_z^2}$
Залишкова	$Q_z = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (X_{ijk} - \overline{X}_{ij*})^2$	a·b·(n–1)	$S_z^2 = \frac{Q_z}{a \cdot b \cdot (n-1)}$	
Загальна (Total)	$Q_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} (X_{ijk} - \overline{X})^2$	a·b·n–1	$S_{total}^2 = \frac{Q_{total}}{a \cdot b \cdot n - 1}$	

Завдання 1: однофакторний дисперсійний аналіз

1. На аркуші **“Однофакторний”** ввести значення для однофакторного дисперсійного аналізу (вибірки D та E, див. Вибірки, с.194, та варіанти завдань с.203).
2. Сформувати статитстичні гіпотези.
3. Виконати процедуру Однофакторный дисперсионный анализ з Пакета аналізу.
4. Побудувати графік групових середніх.
5. Проінтерпретувати отримані результати, зробити статистичні висновки.

Завдання 2: однофакторний дисперсійний аналіз в пакетах SPSS та Statistica

1. Перенести дані лабораторної роботи в пакет SPSS. Добавити змінну, що містить значення градацій контрольованого фактора.
2. В пакеті SPSS виконати однофакторний дисперсійний аналіз: Analyze –> Compare Means –> One-Way ANOVA, вказавши залежною змінною (Dependent List) основну досліджувану змінну, а змінну, що містить градації досліджуваного фактора як Factor.
3. В настройках (Options) встановити прапорці Descriptive, Homogeneity-of-variance та Means plot. Відповідно у звіті про дисперсійний аналіз отримаємо описову статистику для комірок дисперсійного комплекса, тест на однорідність дисперсій та графік середніх.
4. У пакеті Statistica оформити вхідні дані так само, як і в пакеті SPSS. Виконати процедуру Statistics –> ANOVA, вибрати тип аналізу – One-Way ANOVA. У наступному діалоговому вікні вказати залежну змінну (Dependent variable) та факторну змінну і її рівні (в пакеті Statistica для аналізу можна вибрати не всі градації фактора).
5. На закладці Summary у наступному діалоговому вікні переглянути результати однофакторного аналізу: Cell statistics – результати описової статистики для кожної комірки дисперсійного комплексу, Univariate results – власне результати однофакторного аналізу, All effects/Graphs – графіки середніх.

6. Виконати однофакторний дисперсійний аналіз за допомогою послуги Statistics –> Basic Statistics/Tables –> Breakdown & one-way ANOVA. На закладці Quick у діалоговому вікні Statistics by Groups Results отримати Summary Table of Statistics – результати описової статистики, Analysis of Variance – власне результати дисперсійного аналізу та Interaction plots – графіки середніх.
7. Порівняти результати застосування однофакторного дисперсійного аналізу в різних статитсичних пакетах.

Завдання 3: двофакторний дисперсійний аналіз

1. Ввести значення для двофакторного аналізу (див. с.208).
2. Виконати процедуру Двухфакторный дисперсионный анализ с повторениями (аналіз без повторів здійснюється у разі, коли кожна комірка дисперсійного комплексу містить лише одне значення, наприклад, середнє по групі).
3. Побудувати графіки зміни середнього арифметичного для різних градацій впливаючих факторів. За графіком висунути припущення про наявність чи відсутність взаємодії факторів.
4. Проінтерпретувати отримані результати, зробити висновки.

Завдання 4: двофакторний дисперсійний аналіз в пакетах SPSS та Statistica

1. Перенести дані лабораторної роботи в пакет SPSS (добавити змінні, що містять значення градацій контрольованих факторів).
2. Виконати процедуру Analyze –> General Linear Model –> Univariate.
3. Перенести залежну змінну у поле Dependent Variable, а змінні-фактори у поле Fixed Factor(s).
4. У діалоговому вікні Options встановити прапорці Descriptive statistics та Estimates of effect size (оцінка величини ефекта).
5. Для побудови графіків вибрати кнопку Plots та додати два графіки, вказавши один із факторів як горизонтальну вісь (Horisontal Axis), а інший – як окрему лінію (Separate Lines).
6. Порівняти отримані результати з результатами попереднього завдання, зробити висновки.

Приклад виконання

Приклад 1. Однофакторний дисперсійний аналіз.

Три групи операторів слідкували за рухомим об'єктом. З кожним досліджуваним було проведено по 10 дослідів та обчислено середню кількість помилок. З'ясувати, чи залежить кількість помилок від професійного досвіду досліджуваних? [187, с.41].

Таблиця 6 містить дані для задачі, оформлені в MS Excel.

Таблиця 6

	A	B	C
1	Фактор А: Досвідченість операторів		
2	Досвідчені	Новачки	Студенти
3	3,13	1,39	5,47
4	3,25	5,38	5,6
5	3,64	4,07	6,88
6	3,4	3,87	6,4
7	2,59	4,37	3,02
8	1,97	3,79	6,18
9	3,16	3,33	5,52
10	4,22	5,39	4,15
11	1,36	3,37	2,07
12	3,47	4,74	4,68

У вікні майстра функції для однофакторного дисперсійного аналізу слід вказати наступне:

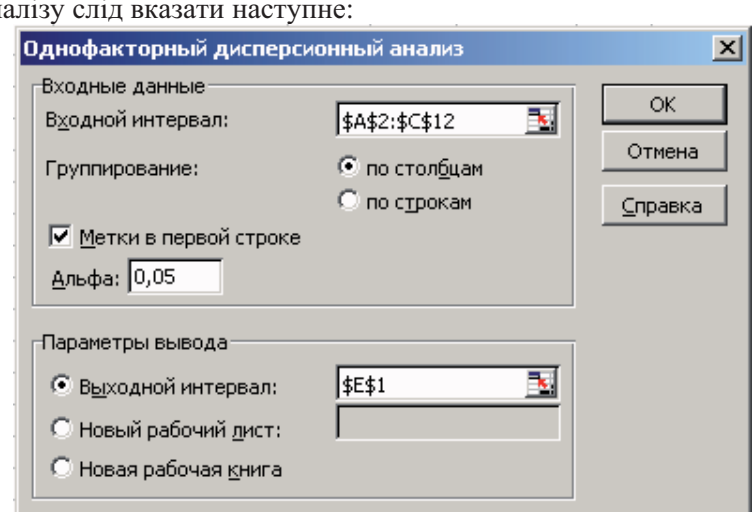


Рис. 7

Гіпотеза H_0 стверджуватиме, що відмінності між групами за кількістю допущених помилок випадкові. Альтернативна гіпотеза полягатиме у тому, що ці відмінності будуть визнані невипадковими.

В результаті виконання процедури однофакторного аналізу з пакета Аналізу даних в MS Excel отримається (див. Таблиця 7): $F_{емпіричне}$=6,72. Воно більше за $F_{критичне}$=3,35. Отже нульову гіпотезу слід відхилити: відмінності у кількості допущених помилок значно більші, ніж відмінності, обумовлені випадковими причинами.

Таблиця 7

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Досвідчені	10	30,19	3,019	0,7
Новачки	10	39,7	3,97	1,36
Студенти	10	49,97	4,997	2,341

Дисперсионный анализ						5%
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F кр*
Между группами	19,572	2	9,7860233	6,672	0,004420313	3,3541
Внутри групп	39,603	27	1,4667741			
Итого	59,175	29				

На графіку (Рис. 8) слід зобразити значення середніх для кожної умови:

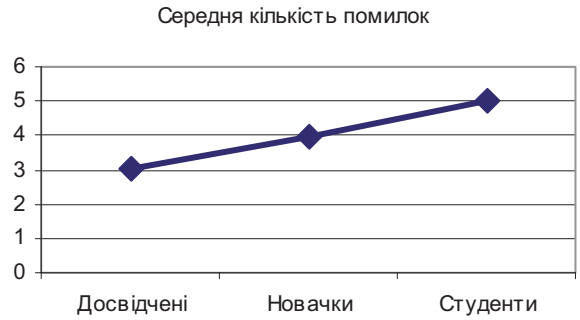


Рис. 8

У пакетах SPSS та Ststistica результати будуть аналогічні.

Приклад 2. Двофакторний дисперсійний аналіз (варіант 11, с. 208): вивчали залежність часу завантаження комп'ютерів від їхнього типу (MAC або Dell) та виду браузера (Netscape Communicator або Internet Explorer).

На таблиці Таблиця 8 представлено дані, у вигляді, в якому їх слід подати для виконання двофакторного дисперсійного аналізу у пакеті MS Excel (комірки, що містять назви градацій факторів, також слід включати до вхідного інтервалу).

Таблиця 8

	MAC	Dell
Netscape Communicator	142	284
	132	304
	125	273
	136	340
	127	326
	138	301
	147	291
	143	285
Internet Explorer	198	285
	210	292
	199	305
	202	325
	196	297
	213	301
	207	285
	201	290

Таблиця 9

Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	MAC	Dell	Итого
Netscape Communicator			5%
Счет	8	8	16
Сумма	1090	2404	3494
Среднее	136,25	300,5	218,375
Дисперсия	61,0714	511,714	7461,45
Internet Explorer			
Счет	8	8	16
Сумма	1626	2380	4006
Среднее	203,25	297,5	250,375
Дисперсия	37,0714	174,857	2467,72
Итого			
Счет	16	16	
Сумма	2716	4784	
Среднее	169,75	299	
Дисперсия	1242,87	322,8	

У таблицях (Таблиця 9 та Таблиця 10) подано результати застосування процедури дисперсійного аналізу.

Таблиця 10

Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F крит.
Выборка	8192	1	8192	41,76	5,3E-07	4,196
Столбцы	133645	1	133645	681,24	3,4E-21	4,196
Взаимодействие	9800	1	9800	49,95	1,1E-07	4,196
Внутри	5493	28	196,18			
Итого	157130	31				

У даному дослідженні фактором А є браузер, фактором В – платформа (тип комп'ютера). Гіпотези дослідження можна сформулювати так:

H_{0A}: час завантаження не залежить від типу браузера.

H_{1A}: час завантаження залежить від типу браузера.

H_{0B}: час завантаження не залежить від типу комп'ютера.

H_{1B}: час завантаження залежить від типу комп'ютера.

H_{0AB}: час завантаження на різних комп'ютерах однаково залежить від виду браузера і навпаки.

H_{1AB}: час завантаження на різних комп'ютерах по різному залежить від типу браузера і навпаки.

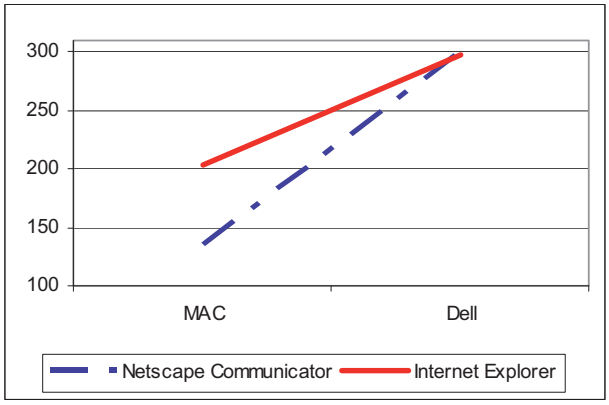


Рис. 9

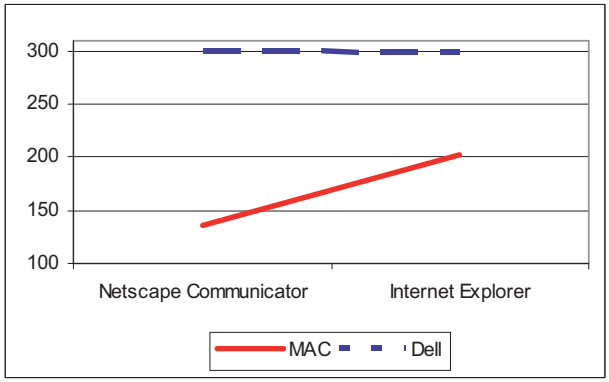


Рис. 10

За результатами дисперсійного аналізу усі коефіцієнти F виявилися більшими за критичні значення, отже нульові гіпотези слід відхилити та визнати достовірними вплив на час

завантаження як типу комп'ютера, так і типу браузера. Цікаво також проінтерпретувати достовірність взаємодії цих двох факторів: на графіку (Рис. 9, Рис. 10) добре видно, що на комп'ютерах Dell обидва браузери завантажуються однаково повільно, а на комп'ютерах Mac значно швидше, але по-різному.

Аналогічні результати буде отримано у пакеті SPSS. Однак вхідні дані слід подати інакше: кожному фактору повинна відповідати змінна. Для даного прикладу слід додати змінну ***browser*** із значеннями 1 – Netscape Communicator та 2 – Internet Explorer, та змінну ***computer*** із значеннями 1 – Mac та 2 – Dell.

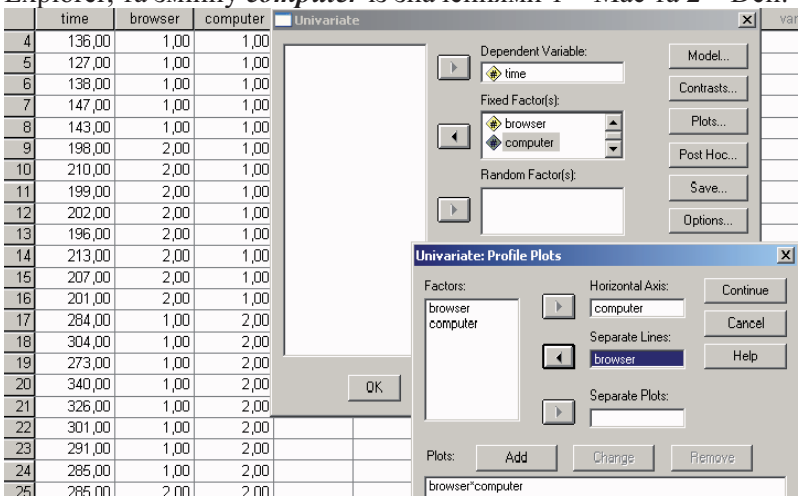


Рис. 11

На рисунку Рис. 11 показано фрагмент списку даних та заповнені поля діалогових вікон процедури дисперсійного аналізу.

Звіт міститиме результати описової статистики (Таблиця 11, Таблиця 12).

Таблиця 11

Between-Subjects Factors

		Value Label	N
BROWSER	1,00	NC	16
	2,00	IE	16
COMPUTER	1,00	mac	16
	2,00	dell	16

Таблиця 12

Descriptive Statistics

Dependent Variable: T_LOAD

BROWSE	COMPUTE	Mean	Std. Deviation	N
NC	mac	136,2500	7,8148	8
	dell	300,5000	22,6211	8
	Total	218,3750	86,3797	16
IE	mac	203,2500	6,0886	8
	dell	297,5000	13,2234	8
	Total	250,3750	49,6761	16
Total	mac	169,7500	35,2543	16
	dell	299,0000	17,9666	16
	Total	234,3750	71,1948	32

Таблиця 13

Tests of Between-Subjects Effects

Dependent Variable: TIME

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	151636,5[a]	3	50545,5	257,7	,000	,965
Intercept	1757813	1	2,E+06	8960	,000	,997
BROWSER	8192,000	1	8192,00	41,758	,000	,599
COMPUTER	133644,5	1	133645	681,2	,000	,961
BROWSER * COMPUTER	9800,000	1	9800,00	49,954	,000	,641
Error	5493,000	28	196,179			
Total	1914942	32				
Corrected Total	157129,5	31				

a. R Squared = ,965 (Adjusted R Squared = ,961)

Основна таблиця дисперсійного аналізу (Таблиця 13) додатково міститиме значення коефіцієнта η^2 (ета квадрат) – внесок незалежної змінної або взаємодії змінних на розсіювання значень залежної змінної.

Corrected model – скоригована сума квадратів моделі – враховує відхилення, обумовлені незалежніми змінними та їхньої взаємодією.

Corrected total – скоригована повна сума квадратів – характеризує усю дисперсію.

Error – залишкова сума квадратів – характеризує відхилення, не обумовлене незалежними змінними або їхньою взаємодією.

Отримані графіки (Рис. 12) також аналогічні побудованим у MS Excel.

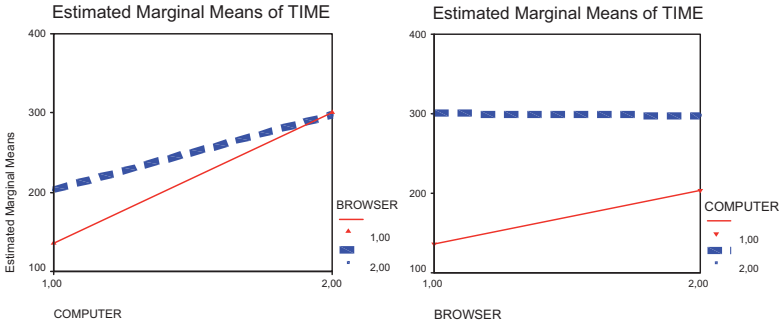


Рис. 12

Аналогічні результати буде отримано і у пакеті Statistica.

Контрольні запитання

1. Призначення дисперсійного аналізу?
2. Умови застосування дисперсійного аналізу.
3. Назвіть та прокоментуйте джерела варіації в однофакторному дисперсійному аналізі.
4. Якими мають бути розміри вибірок (дисперсійних комплексів) у дисперсійному аналізі.
5. Що таке фактор. Які можуть бути фактори?
6. Що таке кореляційне відношення? Як його обчислюють? Як визначити його достовірність?
7. Які статистичні гіпотези формулюються в однофакторному дисперсійному аналізі. Навести приклади.
8. Які статистичні гіпотези формулюються у двофакторному дисперсійному аналізі. Навести приклади.
9. Що таке взаємодія факторів? Коли її визначають? Як її інтерпретувати?

Тема 4: “Кореляційний та регресійний аналіз”

Мета:

Студенти повинні знати:
- статистичні функції Excel для обчислення коефіцієнту кореляції;
- функції для визначення достовірності коефіцієнту кореляції;
- правила прийняття статистичних гіпотез;
- прийоми побудови діаграми розсіювання;
- умови застосування критеріїв статистичного висновку;
- правила обчислення критеріїв для емпіричних та теоретичних розподілів;
- правила обчислення частот теоретичних розподілів.

Студенти повинні уміти:
- подавати експериментальні дані у вигляді, зручному для обчислень;
- виконувати обчислення за алгоритмом статистичного критерію;
- будувати діаграми розсіювання;
- подавати експериментальні дані у вигляді, зручному для обчислень;
- за допомогою статистичних функцій розраховувати частоти нормального розподілу;
- виконувати обчислення за алгоритмом статистичного критерію.

Теоретичні відомості

Кореляційний аналіз виконують для перевірки гіпотези про зв’язок між досліджуваними змінними. Нульовою гіпотезою стверджується, що коефіцієнт кореляції дорівнює нулю, тобто зв’язок між змінними відсутній. Альтернативна гіпотеза – це гіпотеза про те, що кореляція не дорівнює нулю, тобто між змінними існує зв’язок, прямий або обернений, залежно від знаку коефіцієнта кореляції, який може набувати значення – $1 \le \boldsymbol{r} \le 1$.

Для числових нормально розподілених даних лінійну кореляцію обчислюють за формулою Пірсона: $r_{xy} = \dfrac{\sum (X - \bar{x})(Y - \bar{y})}{N \cdot s_x \cdot s_y}$, де s_x, s_y – стандартні квадратичні відхилення змінних X та Y відповідно, $\bar{x}$ та $\bar{y}$ – їхні середні, X та Y – значення вибірок; N – кількість порівнюваних пар чисел.

Для даних, які можна вважати порядковими, кореляцію обчислюють за ранговими критеріями (Спірмена або Кендала) – менш потужними, але і менш вибагливими до виду розподілу змінних та шкал вимірювання. За критеріем Спірмена порівнювані змінні слід проранжувати та обчислити різниці рангів у відповідних парах значень (d)[8]: $r_s = 1 - \frac{6 \cdot \sum d^2}{N \cdot (N^2 - 1)}$.

Достовірність коефіцієнта кореляції оцінюють за таблицями критичних значень, або за допомогою критерія Стьюдента. Значення t-статистики обчислюють як відношення вибіркової оцінки досліджуваного параметра до його стандартної похибки. У разі кореляції вибірковою оцінкою буде *r*.

Обчислене за формулою $t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$ (тут n – обсяг вибірки) значення t-статистики порівнюють з критичним для визначеного рівня значущості і степенів вільності df=n-1. Якщо є підстави відкинути гіпотезу H_0 і прийняти гіпотезу H_1, то кореляцію можна вважати достовірною. У протилежному випадку кореляцію вважають недостовірною, а зв'язок між змінними, незалежно від абсолютної величини коефіцієнта кореляції, випадковим.

Кореляцію якісних ознак, тобто змінних, виміряних за номінативною шкалою, буде розглянуто у темі №5 (с. 86).

На практиці важливо буває з'ясувати кореляцію двох ознак, обумовлену загальним впливом третьої змінної, тобто часткову кореляцію – Partitial Correlation.

Не завжди зв'язок між змінними виявляється лінійним. Показником криволінійного зв'язку є кореляційне відношення η (ета) – відношення дисперсії групових середніх до загальної дисперсії. Розрізняють кореляційне відношення Y по X ($\eta_{yx} = \frac{s_{yx}}{s_y}$) та кореляційне відношення X по Y ($\eta_{xy} = \frac{s_{xy}}{s_x}$), які виявляються однаковими ($\eta_{yx}=\eta_{xy}$) лише за умови лінійності зв'язку. Тут s_x, s_y – загальні стандартні квадратичні відхилення

[8] При наявності однакових рангів коефіцієнт Спірмена обчислюється з поправкою, яка тут не наводиться.

за змінними X та Y відповідно; $s_{yx} = \sqrt{\frac{1}{n}\sum p_x(\bar{y}_x - \bar{y})^2}$ та $s_{xy} = \sqrt{\frac{1}{n}\sum p_y(\bar{x}_y - \bar{x})^2}$ – групові стандартні квадратичні відхилення; p_x, p_y – частоти рядів X та Y; x, y – загальні середні; y_x, x_y – середні у класах рядів розподілу; n – обсяг вибірки.

Достовірність кореляційного відношення також визначається за критерієм Стьюдента [17, 5, 18].

Порядок обчислення кореляційного відношення схожий на порядок виконання однофакторного дисперсійного аналізу, тому у статистичних пакетах коефіцієнт η обчислюється у межах відповідних процедур (див. с. 35, 41). Детальніше про це у прикладі 2.

Називають чотири основні властивості кореляції: напрямленість, силу, форму та напрямок.

Напрямленість характеризує обумовленість зміни значень однієї випадкової величини змінами іншої. Одностороння напрямленість означає, що зміни X обумовлені змінами Y, але не навпаки, або зміни Y обумовлені змінами X, але не навпаки. Двостороння (взаємна) напрямленість означає, що X обумовлює Y, а Y обумовлює X.

Форма кореляції може бути лінійною або нелінійною.

Напрямок кореляції визначається її знаком: якщо більшим значенням змінної X відповідають більші значення змінної Y, то кореляція позитивна, якщо менші, то негативна.

Показником сили кореляції є її квадрат (r^2 для лінійної, η^2 для нелінійної), який називають коефіцієнтом детермінації.

Достовірність будь-якого коефіцієнта детермінації визначається за критерієм Фішера: $F = \frac{\eta^2(mn-m)}{(1-\eta^2)(m-1)}$. Тут η^2 – коефіцієнт детермінації, m – кількість змінних, для яких обчислено коефіцієнт детермінації; n – кількість значень кожної змінної. Обчислене значення $F_{емп}$ порівнюють з критичним для степенів вільності $df_1=m-1$ та $df_2=mn-m$ [18].

Якщо між змінними існує кореляційний зв'язок, то доцільно припустити також наявність функціонального зв'язку між ними, а отже цікавою для прогнозування значень однієї

змінної за відомими значеннями іншої або інших є задача побудови за експериментальними даними апроксимуючої функції $Y=f(X)+\varepsilon$, яку називають регресією.

Ознаку Y можна розглядати і як функцію кількох аргументів $x_1, x_2, x_3, \ldots x_m$. Тоді говорять про множинну регресію: $y = a+bx_1+cx_2+\ldots$

Символом ε позначено випадкову величину – похибку прогнозування.

У найпростішому випадку розглядають лінійну регресію, однак на практиці часто зустрічаються залежності, які краще апроксимуються параболічними (поліноміальними), показниковими, степеневими та іншими нелінійними функціями.

Показником ефективності регресійної моделі є коефіцієнт детермінації R^2 – квадрат коефіцієнта кореляції, – який показує долю загальної варіації змінної Y, поясненої змінною X. Тобто $r^2 = \frac{\sum(Y'-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$, де $\bar{Y}$ – середнє емпіричних значень Y (значення у знаменнику дробу є загальною варіацією змінної Y), Y' – значення, отримані за допомогою регресійної моделі (прогнозовані), тобто $Y'=a+bX$. Отже значення у чисельнику – це варіація, пояснена впливом змінної X. Для лінійної моделі регресії Y по X значення коефіцієнтів a та b обчислюють за формулами: $b = \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2}$ або $b = r \cdot \sqrt{\frac{\sum(y-\bar{y})^2}{\sum(x-\bar{x})^2}}$ та $a = \bar{y} - b\bar{x}$. Аналогічно можна обчислити коефіцієнти регресії X по Y.

Достовірність показників регресії (відмінність від нуля) визначається за критерієм Стьюдента. Коефіцієнт регресії вважають значущим, коли $t_{емп}>t_{кр}$.

Відповідність математичної моделі експериментальним даним, тобто значимість рівняння регресії, визначають за співвідношенням дисперсій врахованих та неврахованих регресійною моделлю факторів. Вважають, що рівняння регресії незначуще, математична модель погано узгоджується з експериментальними даними, коли $F_{емп}<F_{кр,}$ тобто при H_0.

Підтвердити правильність математичної моделі (коли $F_{емп}>F_{кр}$) можна за аналізом залишків, тобто різниць між

експериментальними даними та обчисленими на основі отриманого рівняння регресії. У класичних методах регресійного аналізу залишки розглядаються як незалежні випадкові величини з нормальним законом розподілу [14, 8, 14].

Завдання 1: Кореляційний аналіз

Завдання 1а: кореляційний аналіз у пакеті MS Excel

1. Ввести експериментальні дані – стовпці X, Y, Z за вибіркою C[9] (див. Вибірки, с.194).
2. Для кожного стовпця обчислити кількість (СЧЕТ), середнє значення (СРЗНАЧ), дисперсію (ДИСП), середнє квадратичне відхилення (СТАНДОТКЛОН).
3. Обчислити коефіцієнти кореляції r_{xy}, r_{xz}, r_{yz} за формулою Пірсона (ПИРСОН).
4. Визначити достовірність коефіцієнта кореляції за критерієм Стьюдента:
   а) обчислити значення критерія Стьюдента;
   б) визначити p-значення (імовірність помилки I роду): p = СТЬЮДРАСП(t, n-2, **2**). Тут першим параметром є обчислене значення критерію, другим – кількість степенів вільності, а третім – кількість хвостів розподілу (**2** – для двостороннього критерію та **1** – для одностороннього). Порівняти p-значення з прийнятим рівнем значущості (α=0,05 або α=0,01). При отриманому p<α, гіпотеза H0 відкидається і приймається H1, тобто кореляція достовірна. При p>α – кореляцію вважають недостовірною, тобто зв'язок між змінними, незалежно від абсолютної величини коефіцієнта кореляції, випадковий.
5. Побудувати точкові діаграми (діаграми розсіювання) для кожної пари змінних.
6. Оцінити мінімальний обсяг вибірки, необхідний для забезпечення запланованої точності коефіцієнту кореляції, використовуючи перетворення Фішера (функція ФИШЕР). Для цього обчислити мінімальне n за формулою:

---

[9] Кожен рядок вибірки C (трійка чисел x, y, z) – це результати вимірювання одного об'єкта, тобто "запис", тому деякі операції, зокрема сортування даних, виконуються лише над записами (рядками), а не над змінними.

$n = \frac{t^2}{z^2} + 3$, де n – шуканий обсяг вибірки, t – величина, задана за прийнятим рівнем значущості (для α=0,05 t=1,96; для α=0,01 t=2,58); z – перетворене за Фішером значення коефіцієнту кореляції. Зробити висновки про необхідний обсяг вибірки (див. Приклад 1).

7. Виконати оцінку коефіцієнта кореляції змінних X та Y, обчисливши коефіцієнт кореляції Спірмена для частини основної вибірки (15-20 значень). Сформувати малочисельну вибірку випадковим чином, скориставшись процедурою ВИБОРКА з пакета Аналізу даних (див. с. 10) або згенерувавши 15-20 випадкових чисел за допомогою функції СЛУЧМЕЖДУ(1; N) (тут N – обсяг вибірки С) та скопіювавши на окремий аркуш записи з вибірки С з відповідними номерами. Обчислити для них коефіцієнт кореляції Спірмена за алгоритмом. Зробити висновки про його достовірність. Порівняти результат дослідження випадкової малочисельної вибірки з результатом для генеральної сукупності (тут вибірка С).
8. Обчислити для тієї ж малочисельної вибірки коефіцієнт кореляції Пірсона. Порівняти результати застосування різних алгоритмів. Зробити висновки.

Завдання 1б: кореляційний аналіз у пакеті SPSS

1. Для цього застосувати процедуру Analyze –> Correlate –> Bivariate, вибрати для порівняння усі змінні та вказати коефіцієнти кореляції, які слід обчислити: Пірсона (Pearson), τ-Кендала (Kendall's tau-b) та Спірмена (Spearman). У пункті Options можна вказати спосіб обробки пропущених даних та необхідність обчислення деяких параметрів дескриптивної статистики.
2. Побудувати діаграми розсіювання, застосувавши процедуру Graps –> Scatterplot –> Simple.
3. Побудувати тривимірну діаграму розсіювання Graps –> Scatterplot –> 3D. Вибрати найкращій ракурс тривимірної діаграми (відкрити редактор діаграм командою контекстного меню SPSS Chart Object –> Open та скористатися інструментом 3D-Rotation ().

4. Порівняти отримані результати з обчисленими в MS Excel. Зробити висновки.

Завдання 1в: кореляційний аналіз у пакеті Statistica

1. Ознайомитися з інструментарієм процедури Statistics –> Correlation Matrices. Для даних вибірки С виконати на закладці Advanced/Plot операції Summary: Correlation matrix (обчислення коефіцієнта кореляції Пірсона для кожної пари вибратних змінних), Partitial correlation (часткова кореляція), Scatterplot matrix (матриця точкових діаграм), 3D scatterplots (тривимірна точкова діаграма). На тривимірній діаграмі підібрати найкращій ракурс зображення (кнопка  панелі інструментів графічних об'єктів). Результати застосування операцій внести до звіту та зберегти.
2. Застосувати критерії рангової кореляції (процедура Statistics –> Nonparametrics –> Correlation (Spearman, Kendall tau, gamma).
3. Порівняти результати застосування трьох пакетів. Зробити висновки.

Завдання 2: Регресійний аналіз

1. За даними вибірки С побудувати рівняння регресії Y по X та X по Y, скориставшись функцією ЛИНЕЙН() MS Excel.
2. На відповідних точкових діаграмах за допомогою команди Добавить линию тренда побудувати графіки лінійної та деякої нелінійної регресії (підібрати таку нелінійну функцію, яка найкраще пояснює емпіричні дані, тобто має $R^2 \approx 1$). Висунути гіпотезу про те, яка функція краще апроксимує експериментальні дані.
3. За допомогою процедури Регрессия з Пакета Анализа виконати лінійний регресійний аналіз Y по X. Зробити висновки.
4. За допомогою процедури Регрессия з Пакета Анализа побудувати рівняння множинної лінійної регресії змінної Y від X та Z. Зробити аналіз отриманих результатів.
5. Ознайомитися з інструментами регресійного аналізу пакета SPSS (Analyze –> Regression). У пакті SPSS виконати регресійний аналіз Y по X та множинний регресійний аналіз Y по X та Z.

6. За допомогою процедури Analyze –> Curve Estimation побудувати до даних графіки лінійної та квадратичної регресії. Порівняти результати, з'ясувати, яка функція краще апроксимує дані.
7. Ознайомитися з інструментами регресійного аналізу пакета Statistica (Statistics –> Multiple Regression). Виконати регресійний аналіз Y по X у пакті Statistica.
8. Порівняти результати, отримані у трьох пакетах. Зробити висновки.

Приклади виконання

***Приклад 1*:** Студентам було запропоновано виміряти довжину та ширину (у мм) середнього листка у вибраних навмання рослин лісової полуниці та порахувати кількість зубців на листку. Результати вимірювань наведено у таблиці.

	A	B	C	D
1		Довжина (X)	Ширина (Y)	Кількість зубців (Z)
2	рослина 1	30	20	18
3	рослина 2	30	25	18
4	рослина 3	30	20	18
5	рослина 4	37	23	19
6	рослина 5	49	30	21
7	рослина 6	41	21	20
8	рослина 7	25	16	15
9	рослина 8	45	27	20
10	рослина 9	40	25	17
11	рослина 10	39	24	19

З'ясувати, чи існує зв'язок між змінними. Побудувати діаграми розсіювання.

Виконання:

У пакеті MS Excel Зручно подати результати обчислень під таблицею вхідних даних.

Нижче наведено формули, які слід ввести до комірок у даному випадку B14, C14, D14 та E14 відповідно для обчислення кореляції між змінними X таY (R_{xy}).

B14	Коефіцієнт кореляції за Пірсоном для першого та другого стовпців даних	=ПИРСОН(B2:B11;C2:C11)
C14	Емпіричне значення t-критерія для оцінки достовірності отриманого коефіцієнту кореляції: =B14*КОРЕНЬ(10–2)/КОРЕНЬ(1–B14^2)	
D14	Імовірність помилки I роду (α)	=СТЬЮДРАСП(C14;10–2;2)
E14	Результати порівняння α з обраним рівнем значущості та висновки	=ЕСЛИ(D14<0,05;"H1";"H0")

Для наведеного прикладу отримаємо:

	Кореляція (*r*)	t-критерій	alpha	Висновок
Rxy	0,826997062	4,160574433	0,003162722	H1
Rxz	0,821279263	4,071553121	0,00357572	H1
Ryz	0,718129688	2,918738829	0,019328577	H1

Тобто в усіх випадках кореляція достовірна. Для r_{xy} та r_{xz} навіть на рівні значущості α=0,01. А для r_{yz} лише на рівні значущості α=0,05.

Для оцінки необхідного обсягу вибірки вводимо формули:

Перетворене за Фішером значення коефіцієнту кореляції	=ФИШЕР(B14)
Обсяг вибірки необхідний для рівня значущості 1%	=2,58^2/$B18^2+3
Обсяг вибірки необхідний для рівня значущості 5%	=1,96^2/$B18^2+3

Виконавши обчислення у даному випадку отримуємо таке:

		n1%	n5%
Z(xy)	1,178560185	7,792210445	5,76572256
Z(xz)	1,160734982	7,940526868	5,851320236
Z(yz)	0,903772255	***11,14932058***	7,703207431

З таблиці видно, що необхідне значення n менше 10 (наявного обсягу вибірки), тобто даних цілком достатньо для того, щоб вважати отриманий коефіцієнт кореляції достовірним. Однак для коефіцієнта кореляції між змінними Y та Z на рівні

значущості α=0,01 для прийняття достовірних висновків кількість спостережень слід збільшити хоча б до 12.

Діаграму розсіювання для змінних X та Y зображено на Рис. 13. На ній видно, що більшість точок майже лежить на прямій з додатним кутовим коефіцієнтом. Тобто за діаграмою можна припустити наявність сильного прямого зв'язку між змінними, що і підтверджується обчисленнями.

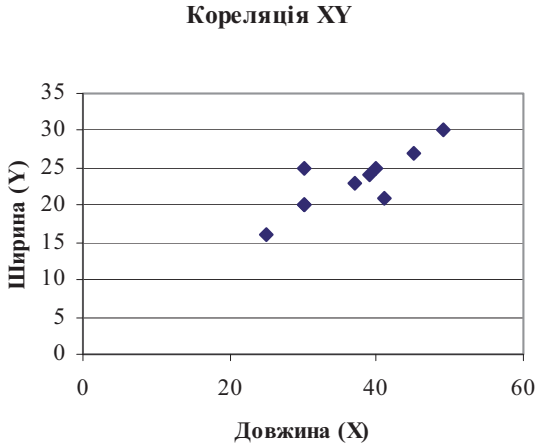


Рис. 13

Порядок обчислення коефіцієнта кореляції за формулою Спірмена для змінних X та Y представлено у наступній таблиці:

	A	B	C	D	E	F	G	H
1		Довжина (X)	Ширина (Y)	Кількість зубців (Z)	Ранг X	Ранг Y	d	d^2
2	рослина 1	30	20	18	3	2,5	0,5	0,25
3	рослина 2	30	25	18	3	7,5	-4,5	20,25
4	рослина 3	30	20	18	3	2,5	0,5	0,25
5	рослина 4	37	23	19	5	5	0	0
6	рослина 5	49	30	21	10	10	0	0
7	рослина 6	41	21	20	8	4	4	16
8	рослина 7	25	16	15	1	1	0	0
9	рослина 8	45	27	20	9	9	0	0
10	рослина 9	40	25	17	7	7,5	-0,5	0,25
11	рослина 10	39	24	19	6	6	0	0

Сума значень останнього стовпця $\Sigma d^2=37$. Відповідно $r_s = 0{,}776$. Це значення менше за коефіцієнт кореляції за Пірсоном, однак воно теж достовірне (за критерієм Стьюдента) і свідчить про значний позитивний зв'язок між змінними, у даному разі – між шириною та довжиною листка.

Таблиця 14

Descriptive Statistics

	Mean	Std. Deviation	N
X	36,60	7,65	10
Y	23,10	4,01	10
Z	18,50	1,72	10

У пакеті SPSS результати кореляційного аналізу можна отримати одразу для усіх вказаних змінних. Якщо вказано виводити параметри описової статистики, то у таблиці (Таблиця 14) для кожної змінної будуть виведені середнє та стандартне квадратичне відхилення.

У таблиці Таблиця 15 буде виведено результати обчислення коефіцієнтів кореляції за Пірсоном. Однією зірочкою (*) позначено результати, достовірні на рівні значущості $\alpha=0{,}05$, а двома зірочками (**) – результати, достовірні на рівні значущості $\alpha=0{,}01$. Крім коефіцієнтів кореляції у кожній клітинці таблиці наведено також p-значення та кількість порівнюваних пар чисел. Як бачимо результат не відрізняється від отриманого в MS Excel.

Таблиця 15

Correlations

		X	Y	Z
X	Pearson Correlation	1,000	,827**	,821**
	Sig. (2-tailed)	,	,003	,004
	N	10	10	10
Y	Pearson Correlation	,827**	1,000	,718*
	Sig. (2-tailed)	,003	,	,019
	N	10	10	10
Z	Pearson Correlation	,821**	,718*	1,000
	Sig. (2-tailed)	,004	,019	,
	N	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

На таблиці (Таблиця 16) наведено результати обчислення за ранговими критеріями Кендала та Спірмена. За цими критеріями кореляцію між змінними Y та Z не визнано достовірною.

Таблиця 16

Nonparametric Correlations

Correlations

			X	Y	Z
Kendall's tau_b	X	Correlation Coefficient	1,000	,682**	,732**
		Sig. (2-tailed)	,	,008	,005
		N	10	10	10
	Y	Correlation Coefficient	,682**	1,000	,458
		Sig. (2-tailed)	,008	,	,080
		N	10	10	10
	Z	Correlation Coefficient	,732**	,458	1,000
		Sig. (2-tailed)	,005	,080	,
		N	10	10	10
Spearman's rho	X	Correlation Coefficient	1,000	,772**	,806**
		Sig. (2-tailed)	,	,009	,005
		N	10	10	10
	Y	Correlation Coefficient	,772**	1,000	,565
		Sig. (2-tailed)	,009	,	,089
		N	10	10	10
	Z	Correlation Coefficient	,806**	,565	1,000
		Sig. (2-tailed)	,005	,089	,
		N	10	10	10

**. Correlation is significant at the .01 level (2-tailed).

Діаграма розсіювання у пакеті SPSS матиме вигляд як на Рис. 14.

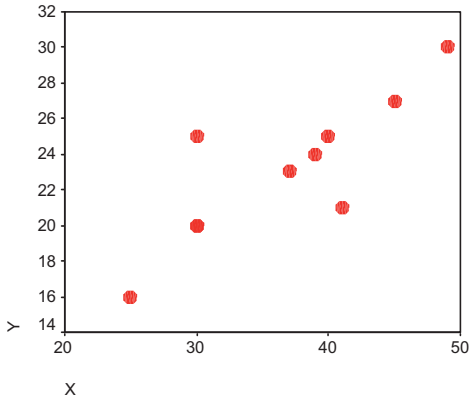


Рис. 14

Застосування процедури обчислення часткової кореляції (Analyze –> Correlate –> Partitial Correlation) при виключеному впливі змінної Y дасть такий результат:

Таблиця 17

Partitial Correlation Coefficients
Controlling for.. Ширина

	Довжина	Кількість
Довжина	1,0000 (0) P= ,	,5812 (7) = ,101
Кількість	,5812 (7) = ,101	1,0000 (0) P= ,

(Coefficient / (D.F.) / 2-tailed Significance)
" , " is printed if a coefficient cannot be computed

У термінах розглядуваного прикладу цей результат означає, що при однаковій ширині листка (Y) між довжиною (X) та кількістю зубців (Z) існує кореляційний зв'язок: $r_{xy,z} = 0{,}5812$. У даному випадку значення коефіцієнту недостатньо велике, щоб визнати кореляцію достовірною. Можливо слід збільшити обсяг вибірки та повторити вимірювання і обчислення.

У пакеті Statistica результати кореляційного аналізу процедурою Statistics –> Correlation Matrices будуть представлені:

1) кореляційною матрицею:

Variable	Correlations (Spreadsheet1 in Workbook1) Marked correlations are significant at p < ,05000 N=10 (Casewise deletion of missing data)		
	Var1	Var2	Var3
Var1	1,00	0,83	0,82
Var2	0,83	1,00	0,72
Var3	0,82	0,72	1,00

2) матрицею точкових діаграм (Рис. 15)

3) тривимірною точковою діаграмою.

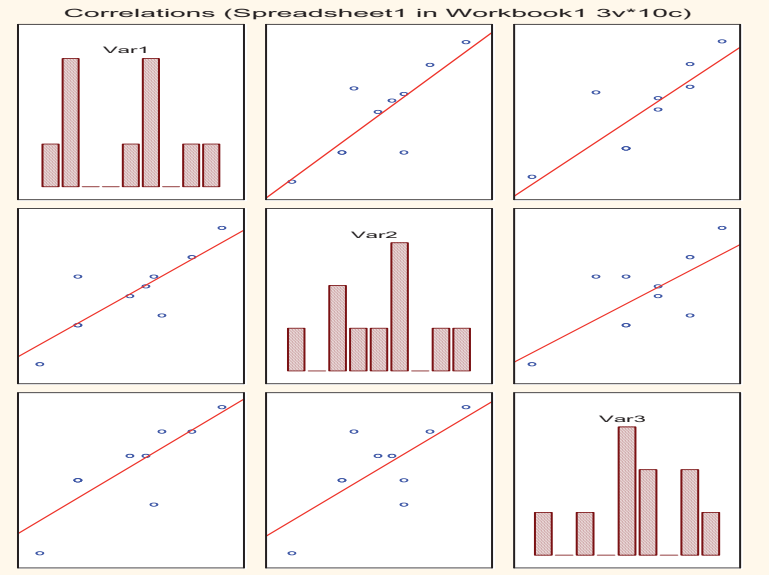


Рис. 15

Приклад 2: У групі студентів перед початком екзамена вимірювали рівень інтелекту (A) та рівень тривожності (B):

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	76	80	80	87	90	97	99	100	100	105	110	112	115	120	126
B	18	23	20	23	25	25	26	28	29	27	25	24	21	18	19

З'ясувати, чи існує зв'язок між виміряними змінними?

Виконання:

Перш ніж формулювати гіпотези, доцільно побудувати точкову діаграму (Рис. 16).

Очевидно, що зв'язок, якщо і достовірний, то нелінійний, тому застосування процедур обчислення коефіцієнта лінійної кореляції буде на користь нульової гіпотези ($r_{ab} \approx 0$).

Для обчислення кореляційного відношення слід застосувати процедуру Analyze –> Compare Means –> Means пакета SPSS. При першому застосуванні вказати залежною змінною A, а незалежною – B. При другому – навпаки. У меню Options вибрати Anova table and eta та Test for linearity.

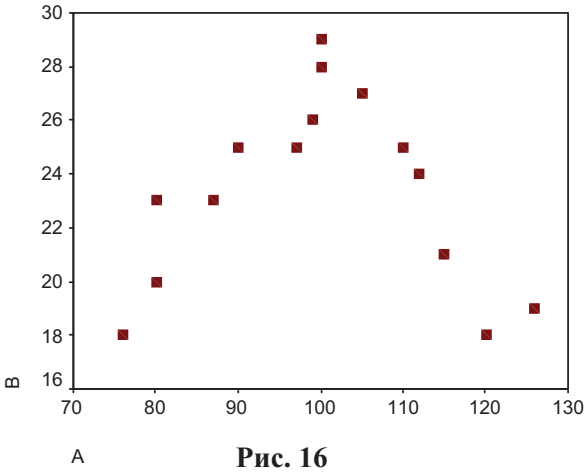


Рис. 16

Результати будуть представлені у чотирьох таблицях:

1. Case Processing Summary – підсумкові відомості про враховані та виключені з обчислень дані.
2. Report – частотна таблиця, побудована за незалежною змінною.
3. Таблиця однофакторного дисперсійного аналізу:

Таблиця 18

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
B * A	Between Groups	(Combined)	170,600	12	14,217	5,687	,159
		Linearity	,275	1	,275	,110	,772
		Deviation from Linearity	170,325	11	15,484	6,194	,147
	Within Groups		5,000	2	2,500		
	Total		175,600	14			

Слід звернути увагу на те, що значна частина міжгрупової дисперсії обумовлена відхиленням від лінійності: Deviation from Linearity = 170,325 при повному значенні міжгрупової дисперсії (Combined) = 170,6.

4. Міри зв'язку:

Measures of Association

	R	R Squared	Eta	Eta Squared
B * A	-,040	,002	,986	,972

Тут R – коефіцієнт кореляції Пірсона; R Squared – квадрат кореляції (коефіцієнт детермінації) – показує, в якій мірі мінливість однієї змінної обумовлена впливом іншої; Eta – кореляційне відношення.

Як бачимо, коли незалежною змінною визначено A, коефіцієнт лінійної кореляції близький до нуля, а кореляційне відношення – майже одиниця.

Аналогічний результат отримається, коли незалежною змінною визначити змінну B:

Measures of Association

	R	R Squared	Eta	Eta Squared
A * B	-,040	,002	,793	,628

Кореляційне відношення обчислюється також у процедурі Analyze –> Descriptives –> Crosstabs –> Statistics (Eta):

Directional Measures

			Value
Nominal by Interval	Eta	A Dependent	,793
		B Dependent	,986

Застосування процедури Analyze –> Compare Means –> Means до даних ***прикладу 1*** дасть такі результати:

Measures of Association

	R	R Squared	Eta	Eta Squared
Y * X	,827	,684	,941	,885

Measures of Association

	R	R Squared	Eta	Eta Squared
X * Y	,827	,684	,951	,905

Як бачимо, тут різниця між двома кореляційними відношеннями незначна, отже зв’язок між змінними майже лінійний, що буде відображено також на таблиці Anova (Таблиця 19).

Таблиця 19

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Y * X	Between Groups	(Combined)	128,233	7	18,319	2,198	,348
		Linearity	99,101	1	99,101	11,892	,075
		Deviation from Linearity	29,133	6	4,855	,583	,743
	Within Groups		16,667	2	8,333		
	Total		144,900	9			

Відхилення від лінійності тут обумовлюють лише 22% міжгрупової дисперсії (29,133 із 128,233).

Приклад 3: Виконати регресійний аналіз за даними прикладу 1.
Виконання:

У пакеті MS Excel рівняння регресії можна побудувати за допомогою функції ЛИНЕЙН() та відповідної процедури Пакета Анализа. Добір відповідної регресійної моделі також можна здійснити на діаграмі за допомогою послуги Добавить линию тренда.

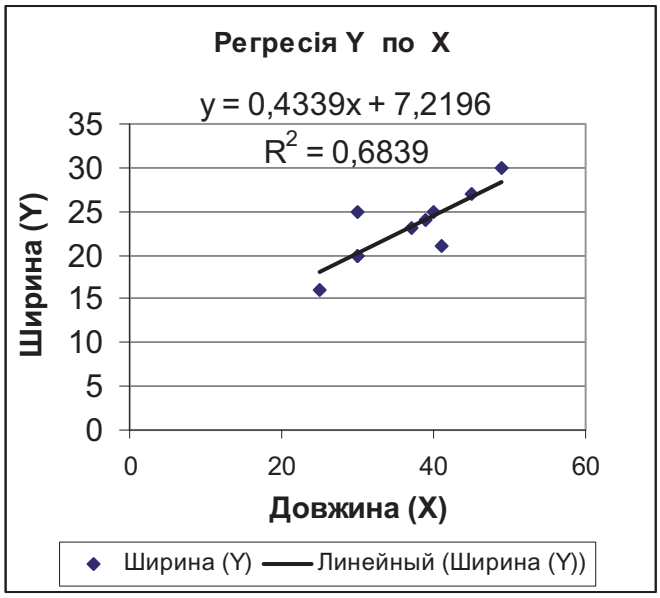


Рис. 17

Якщо на закладці Параметри вибрати настройки “показывать уравнение на диаграмме” та “поместить на диаграмму величину достоверности аппроксимации (R^2)”, то вони будуть виведені на графіку (Рис. 17).

За допомогою лінії тренда можна апроксимувати експериментальні значення, крім лінійної, також степеневою, експоненційною, поліноміальною або логарифмічною функцією.

За допомогою функції ЛИНЕЙН() можна отримати коефіцієнти регресійної прямої або повну статистику регресійного аналізу. У будь-якому разі дана функція є функцією-масивом (див. примітку на с. 11).

У першому випадку результат буде розміщено у масиві з двох комірок. Наприклад, C14:D14. Вставимо туди формулу =ЛИНЕЙН(B2:B11; C2:C11; 1; **0**). Отримаємо: C14=0,4339 – значення b; D14=7,2196 – значення *a*. Тобто формула рівняння регресії буде така: y=0,4339b+7,2196.

У функції ЛИНЕЙН (Y; X; Конст.; Стат.) першим параметром є експериментальні значення залежної змінної (один стовбець); другим – експериментальні значення незалежних змінних (може бути декілька змінних, тобто стовпців); третій параметр вказує, чи потрібно, щоб коефіцієнт *a* дорівнював нулю (для *a*=0 вказують Конст.=0); параметр Стат включає або виключає обчислення повної статистики критерію.

При застосуванні формули =ЛИНЕЙН(B2:B11; C2:C11; 1; **1**) отримаємо результат у два стовпці та 5 рядків:

0,4339	7,2196
0,1043	3,8911
0,6839	2,3927
17,31	8
99,101	45,799

У клітинках відповідно розміщені:

Коефіцієнт *b*	0,4339	Коефіцієнт *a*	7,2196
Стандартна похибка обчислення *b*	0,1043	Стандартна похибка обчислення *a*	3,8911
Коефіцієнт детермінації R^2	0,6839	Стандартна похибка регресійних залишків	2,3927
Обчислене значення F-критерія	17,31	df (степені вільності)	8
Сума квадратів регресії	99,101	Сума квадратів залишків	45,799

При обчисленні множинної регресії результат подається у масив з п’яти рядків та k+1 стовпця (де k – кількість незалежних змінних).

Більш детально регресійний аналіз здійснюється за допомогою процедури Регрессия з Пакета Анализа. Вікно процедури зображено на Рис. 18.

Результатом її застосування будуть такі таблиці:

1. Регресійна статистика:

Регрессионная статистика		
Множественный R	0,826997	– корінь квадратний з коефіцієнта детермінації (R>0)
R-квадрат	0,683924	– коефіцієнт детермінації
Нормированный R-квадрат	0,644415	
Стандартная ошибка	2,392681	
Наблюдения	10	

2. Дисперсійний аналіз:

Дисперсионный анализ					
	df	SS	MS	F	Значимость F
Регрессия	1	99,10061	99,10061	17,31038	0,003163
Остаток	8	45,79939	5,724924		
Итого	9	144,9			

Тут за значенням F та значимістю F можна зробити висновок про те, що регресійна модель досить добре узгоджується з експериментальними даними.

3. Коефіцієнти регресійної прямої та їх значимість:

	Коэффи-циенты	Стандартная ошибка	t-статис-тика	P-Значе-ние	Нижние 95%	Верхние 95%
Y-пересе-чение	7,219605	3,891148	1,85539	0,1006	-1,7534	16,1926
Довжина (X)	0,433891	0,104286	4,16057	0,00316	0,19340	0,67437

Тут у другому стовпці наведено коефіцієнти регресійної прямої, у третьому – похибки їх обчислення, у двох останніх – границі довірчих інтервалів (межі допустимих відхилень від обчислених значень). За величиною t-статистики та її p-значенням визначається достовірність відхилення коефіцієнта регресії від нуля.

4. Аналіз залишків:

ВЫВОД ОСТАТКА				ВЫВОД ВЕРОЯТНОСТИ	
Наблюде-ние	Предсказанное Ширина (Y)	Остатки	Стандартные остатки	Персен-тиль	Ширина (Y)
1	20,23632	-0,23632	-0,10476	5	16
2	20,23632	4,763678	2,111708	15	20
3	20,23632	-0,23632	-0,10476	25	20
4	23,27356	-0,27356	-0,12127	35	21
5	28,48024	1,519757	0,673698	45	23
6	25,00912	-4,00912	-1,77722	55	24
7	18,06687	-2,06687	-0,91623	65	25
8	26,74468	0,255319	0,113181	75	25
9	24,57523	0,424772	0,188299	85	27
10	24,14134	-0,14134	-0,06265	95	30

За даними цієї таблиці будуються відмічені у вікні процедури графіки (Рис. 18).

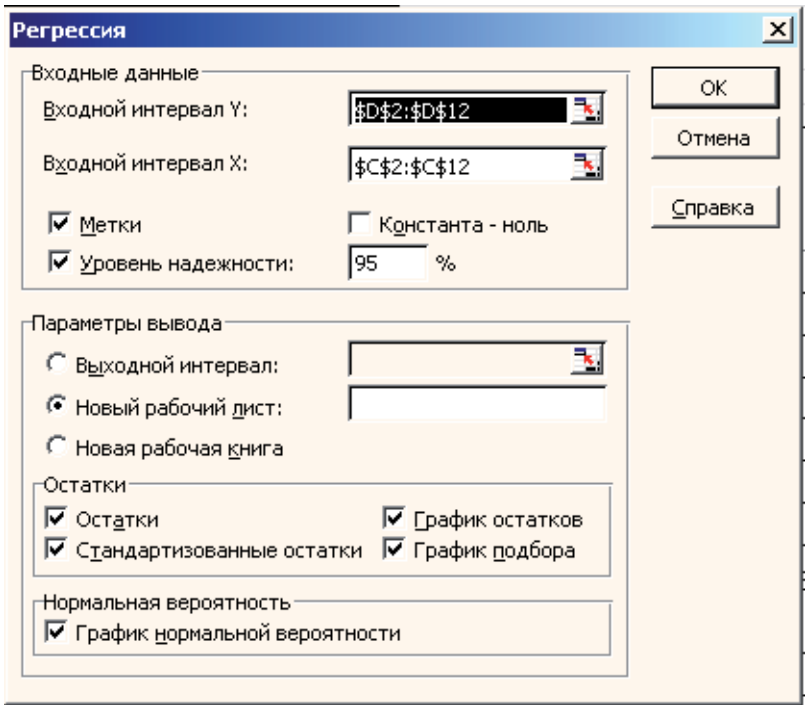


Рис. 18

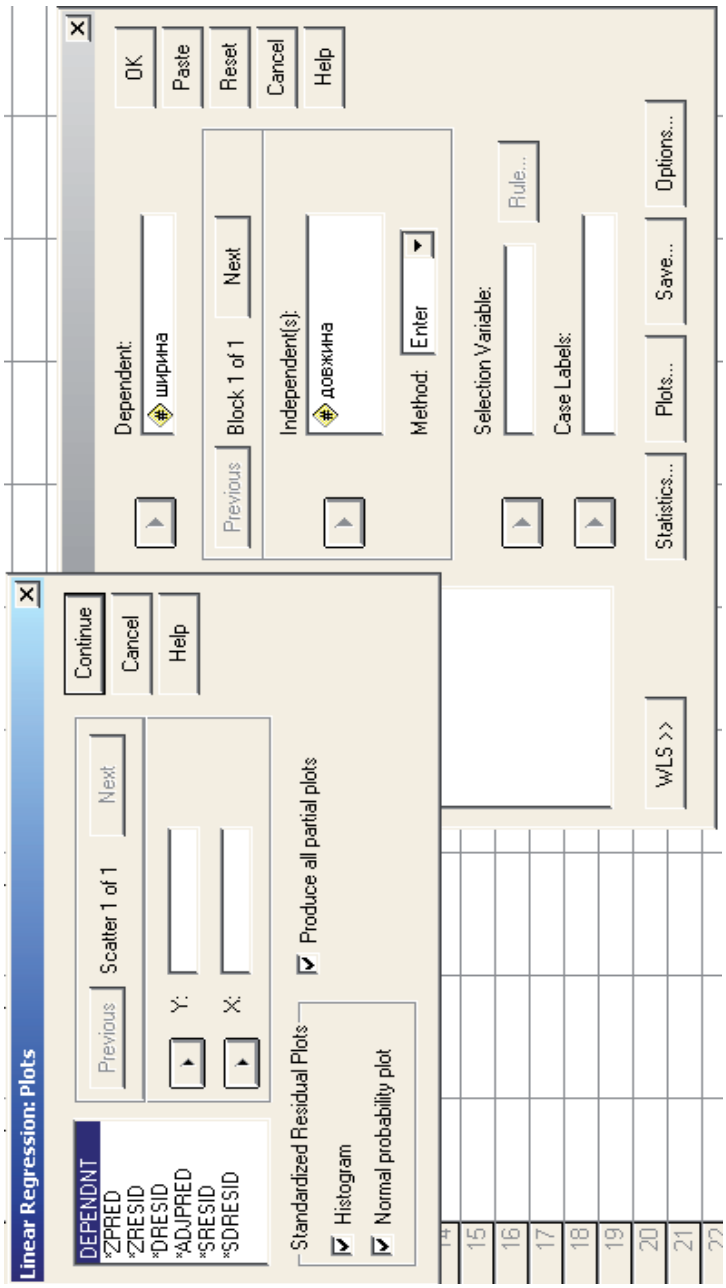

Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Previous
Scatter 1 of 1
Next
Y:
X:
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Continue
Cancel
Help
Dependent:
ширина
Previous
Block 1 of 1
Next
Independent(s):
довжина
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS >>
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help

Рис. 19

"График подбора" показує експериментальні значення залежної змінної та значення, обчислені за регресійною моделлю.

"График нормального распределения" дає можливість окомірним способом перевірити нормальність розподілу залишків. Розподіл залишків тим ближчий до нормального, чим краще відмічені на графіку точки вкладаються на пряму (див. Тема 5, с. 89).

"График остатков" показує стандартизовані (відносно горизонтальної вісі) значення задишків.

У пакеті SPSS для виконання лінійної регресії у вікні процедури Analyze –> Regression –> Linear слід виконати настройки як на Рис. 19.

Отримані результати будуть представлені у вигляді наступних таблиць:

1. Короткий звіт за моделлю:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,827[a]	,684	,644	2,3927

a. Predictors: (Constant), ДОВЖИНА

b. Dependent Variable: ШИРИНА

2. Дисперсійний аналіз відповідності моделі експериментальним даним:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	360,018	1	360,018	17,310	,003[a]
	Residual	166,382	8	20,798		
	Total	526,400	9			

a. Predictors: (Constant), довжина

b. Dependent Variable: ширина

3. Аналіз коефіцієнтів регресійної прямої:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,188	8,870		,021	,984
	довжина	1,576	,379	,827	4,161	,003

a. Dependent Variable: ширина

Отримані результати аналогічні до отриманих в MS Excel.

Порівняти лінійну та нелінійну (наприклад, квадратичну) регресійні моделі за допомогою SPSS можна за допомогою процедури Analyze –> Regression –> Curve Estimation (Рис. 20).

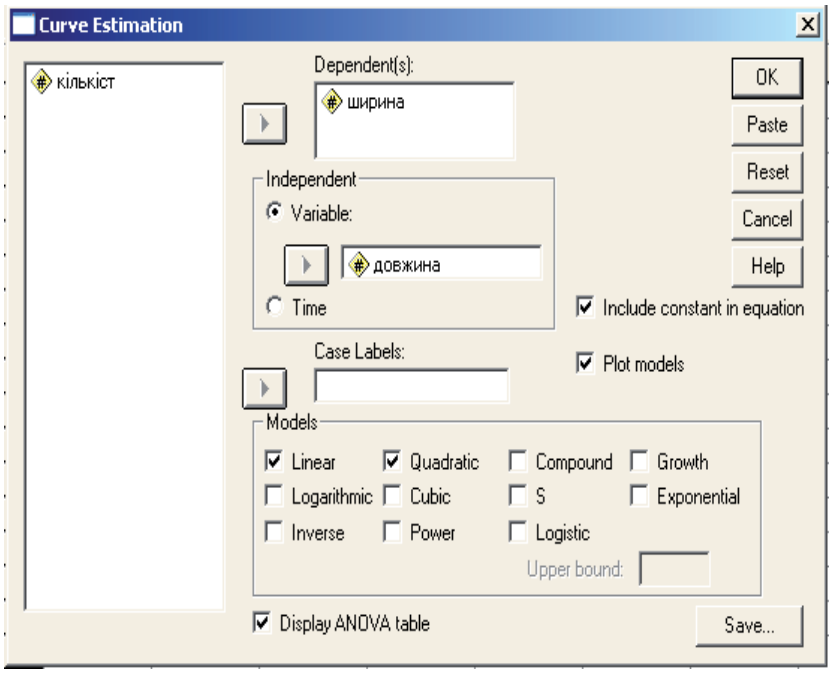


Рис. 20

Для вибраних кривих у звіті будуть утворені таблиці, які нижче наведено у двох стовпцях: зліва для лінійної моделі, справа – для квадратичної:

1. Оцінки моделі (MODEL):

Dependent variable.. ШИРИНА Method.. **LINEAR**	Dependent variable.. ШИРИНА Method.. **QUADRATI**
Listwise Deletion of Missing Data	Listwise Deletion of Missing Data
Multiple R ,82700	Multiple R ,82825
R Square ,68392	R Square ,68600
Adjusted R Square ,64441	Adjusted R Square ,59628
Standard Error 2,39268	Standard Error 2,54948

У даному прикладі обидві моделі відрізняються мало: і лінійна, і квадратична пояснюють приблизно 68% варіації залежної змінної (R Square ≈ 0,68).

2. Дисперсійний аналіз (Analysis of Variance):

Dependent variable.. ШИРИНА Method.. **LINEAR**				Dependent variable.. ШИРИНА Method.. **QUADRATI**			
	DF	Sum of Squares	Mean Square		DF	Sum of Squares	Mean Square
Regression	1	99,1006	99,1006	Regression	2	99,1006	49,7005
Residuals	8	45,7994	5,7249	Residuals	7	45,7994	6,4999
F = 17,31038 Signif F = ,0032				F = 7,64637 Signif F = ,0173			

Дисперсійний аналіз показує, що лінійна модель усеж краще відповідає емпіричним даним: відношення поясненої дисперсії до неврахованої регресійною моделлю F = 99,1/5,72 = 17,31. Це більше, ніж у випадку квадратичної регресії (F = 49,7/6,49 =7,64637), хоча в обох випадках є підстави відкинути нульову гіпотезу (у лінійному випадку на рівні значущості α=0,01, оскільки p-значення = 0,0032, а у квадратичному випадку на рівні значущості α=0,05, оскільки p-значення = 0,0173). Тобто і лінійна, і квадратична регресійні моделі добре узгоджуються з емпіричними даними.

3. Коефіцієнти рівняння регресії (Variables in the Equation):

Dependent variable.. ШИРИНА Method.. **LINEAR**					
Variable	B	SE B	Beta	T	Sig T
ДОВЖИНА	,4339	,1043	,8270	4,161	,0032
(Constant)	7,2196	3,8911		1,855	,1006

Dependent variable.. ШИРИНА Method.. **QUADRATI**					
Variable	B	SE B	Beta	T	Sig T
ДОВЖИНА	,1796	1,188	,342	,151	,88410
ДОВЖИНА**2	,0035	,016	,487	,215	,8359
(Constant)	11,7167	21,329		,549	,5999

Коефіцієнти регресії потрібні для побудуви рівняння регресійної функції. У лінійному випадку отримаємо рівняння y=0,4339 x+7,2196. А у квадратичному випадку y=0,0035 x^2+0,1796 x+11,7167.

Однак t-критерій показує, що у квадратичному випадку коефіцієнти мало відрізняються від нуля, тобто незначимі.

Отже більш прийнятною є лінійна модель.

Графічно у пакеті SPSS це буде зображено так, як показано на Рис. 21.

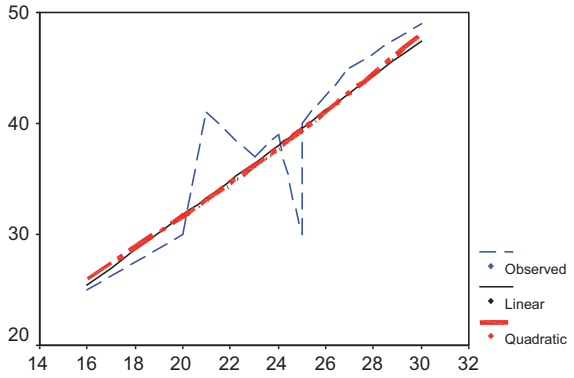


Рис. 21

Ламаною пунктирною лінією тут з'єднано експериментальні дані.

У пакеті Statistica графічно побудувати регресійну криву можна за допомогою процедури Graphs –> 2D Scatterplots (Рис. 22).

Контрольні за питання

1. Яке призначення коефіцієнта кореляції?
2. Які є види кореляційного зв'язку?
3. Як визначити достовірність коефіцієнта кореляції?
4. Схематично зобразити на діаграмі розсіювання достовірний зворотний зв'язок між двома змінними.

5. Які умови застосування коефіцієнта кореляції Пірсона та рангових критеріїв?
6. Що таке часткова кореляція, як її проінтерпретувати?
7. Як визначити кореляцію якісних ознак?
8. Як визначити нелінійну кореляцію?
9. Що показує коефіцієнт детермінації?
10. Що таке регресія? Як пов'язані між собою лінійна кореляція та лінійна регресія?
11. Як оцінити регресійну модель?

12.Як побудувати та оцінити регресійну модель експериментальних даних у різних статистичних пакетах?

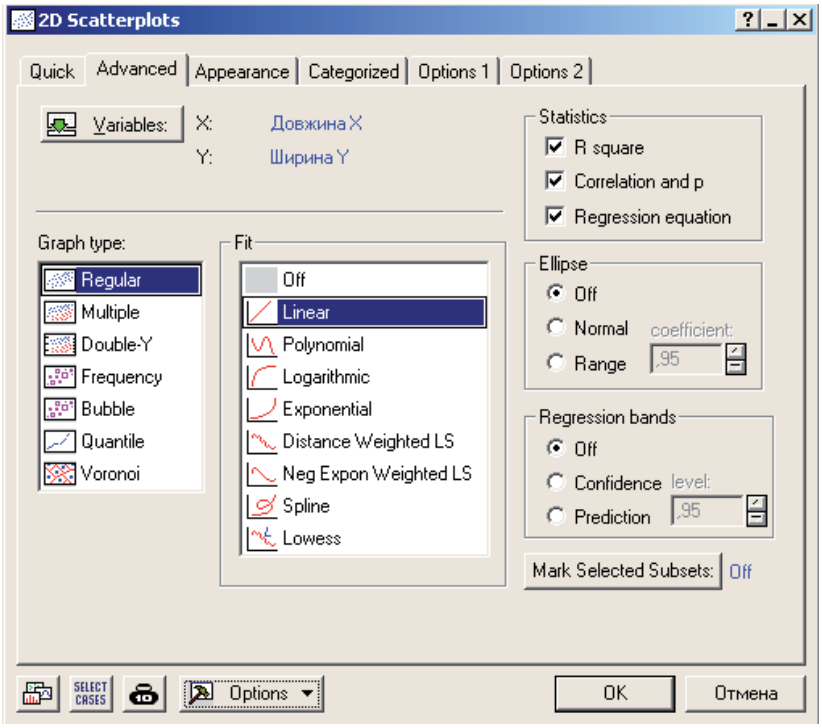


Рис. 22

Тема 5:
“Порівняння розподілів”

Мета:

1. Ознайомитися з критеріями та методами перевірки нормальності емпіричного розподілу даних. Виконати перевірку нормальності засобами пакетів MS Excel, SPSS та Statistica.
2. Ознайомитися та навчитися застосовувати критерії порівняння емпіричного розподілу з деяким теоретичним. Вивчити умови застосування та алгоритм обчислення критеріїв χ^2-Пірсона та λ-Колмогорова-Смирнова.
3. Вивчити порядок порівняння двох та більше емпіричних критеріїв.
4. Ознайомитися із застосуванням критерію χ^2-Пірсона до обчислення кореляції якісних ознак.

Студенти повинні знати:

– особливості нормального та рівномірного розподілу;
– критерії та прийоми порівняння емпіричного розподілу з будь-яким теоретичним, зокрема з нормальним;
– критерії та способи порівняння двох та більше емпіричних розподілів;
– функції MS Excel для роботи з нормальним розподілом;
– призначення, обмеження, особливості застосування критеріїв χ2-Пірсона та λ-Колмогорова-Смирнова, правила прийняття рішення про статистичну гіпотезу.

Студенти повинні уміти:

– подавати експериментальні дані у вигляді, спряжених таблиць та формувати спряжені таблиці даних засобами статистичних пакетів;
– виконувати обчислення за алгоритмами розглянутих статистичних критеріїв;
– будувати відповідні графіки;
– обчислювати процентилі, нормалізовані частоти тощо;
– формулювати статистичні гіпотези та робити статистичні висновки;
– інтерпретувати отримані результати у термінах експериментального дослідження.

Теоретичні відомості

Перевірка нормальності

Статистичні методи, в основу яких покладено оперування з параметрами статистичного розподілу, такими, як, наприклад, середнє, дисперсія, та іншими, називають параметричними. Більшість параметричних методів розраховані на нормально розподілені дані. У випадках відхилення від нормальності не всі параметричні методи будуть коректно працювати, тому в таких випадках перевагу слід надавати менш вибагливим до виду розподілу даних непараметричним методам (див. тему 6, с. 112).

Для перевірки нормальності використовують як специфічні методи (окомірний, критерії асиметрії та ексцеса), так і більш універсальні методи, які застосовують і для вирішення інших статистичних задач (критерії згоди, наприклад, χ2-Пірсона та λ-Колмогорова-Смирнова).

Критерії асиметрії та ексцеса полягають у порівнянні відповідних параметрів розподілу з похибками їх обчислення.

За критерієм Н.О.Плохінського можна вважати, що розподіл не відрізняється від нормального, коли асиметрія та ексцес відрізняються від своїх похибок репрезентативності не більше як у три рази. Тобто, якщо $t_A = \frac{|A|}{m_A} \le 3$ та $t_E = \frac{|E|}{m_E} \le 3$, де $m_A = \sqrt{\frac{6}{n}}$ та $m_E = 2 \cdot \sqrt{\frac{6}{n}}$ відповідно похибки репрезентативності асиметрії та ексцеса, А – асиметрія, Е – ексцес, то розподіл досліджуваної ознаки можна вважати нормальним,.

За критерієм Є.І.Пустильника [25, с. 155] розподіл не відрізняється від нормального, якщо обчислені (емпіричні) значення асиметрії та ексцесу не перевищують критичних. Тобто $A_{емп} < A_{кр}$, $E_{емп} < E_{кр}$, де $A_{емп}$ та $E_{емп}$ – обчислені значення асиметрії та ексцеса, а $A_{кр} = 3 \cdot \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+3)}}$ і $E_{кр} = 5 \cdot \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+2)^2 \cdot (n+3) \cdot (n+5)}}$ відповідно їхні критичні

значення, обчисленні з урахуванням дисперсій: $A_{кр} = 3 \cdot \sqrt{D(A)}$ та $E_{кр} = 5 \cdot \sqrt{D(E)}$.

Критерії асиметрії та ексцеса застосовують лише для малих вибірок. Для великих (n≥20) рекомендується використовувати більш загальні критерії згоди.

Ще одним специфічним методом є окомірний метод, який не дає імовірнісної оцінки висновків про нормальність, але дозволяє її оцінити візуально.

Емпіричний розподіл тим ближчий до нормального, чим краще вкладаються на пряму відповідні квантилі емпіричного інтервального статистичного ряду та стандартного нормального розподілу [8]. Порядок побудови відповідних графіків та їхнього аналізу наведено у завданні і прикладах.

Критерії згоди

Критерії згоди застосовують не лише для перевірки нормальності. До цієї групи відносять статистичні критерії, призначені для виявлення відхилень між будь-якою гіпотетичною статистичною моделлю та реальними даними, для опису яких її використано [8, 3]. Такою моделлю може бути деякий теоретичний розподіл, тобто розподіл, у якому імовірність появи кожного значення випадкової величини можна розрахувати за певною формулою. Такими є, наприклад, нормальний розподіл, рівномірний, біноміальний, Пуассона, Стьюдента, Фішера та інші.

Перевірка відповідності реального розподілу теоретичному потребує великих обсягів даних. При цьому не існує методів, які б дозволили стовідсотково визначити характер розподілу даних. Насправді можна лише перевірити, наскільки добре експериментальні дані відповідають обраній моделі.

Статистичні гіпотези, що формуються для даного класу задач, можуть бути *простими* (у разі, коли перевіряється відповідність деякому закону з відомими визначеними параметрами) та *складними* (коли перевіряється відповідність деякому закону з невідомими параметрами).

Проста гіпотеза має вид H_0: $F(x)=F(x, \theta)$, где $F(x, \theta)$ – функція розподілу імовірностей, узгодженість з якою досліджуваної вибірки перевіряють, а θ – відоме значення параметра (скалярного чи векторного).

Складна гіпотеза має вид H_0: $F(x) \in \{F(x,\theta), \theta \in \Theta\}$, де Θ – область визначення параметра θ. В такому випадку оцінку параметра розподілу $\hat{\theta}$ обчислюють за тою ж вибіркою, за якою перевіряють узгодженість. Якщо оцінку $\hat{\theta}$ обчислюють за іншою вибіркою, то гіпотеза проста.

При порівнянні емпіричного розподілу з теоретичним, як правило, перш за все розраховують частоти теоретичного розподілу для даного обсягу вибірки (порівнювати зручно вибірки однакового обсягу).

На другому етапі порівняння застосовують обраний критерій згоди. У даній роботі розглянуто застосування критеріїв χ^2-Пірсона та λ-Колмогорова-Смирнова для порівняння їх підходів до виявлення різниці та потужності, тобто чутливості до відмінностей.

Побудова теоретичного розподілу

При побудові теоретичного розподілу спочатку будують варіаційний ряд емпіричного розподілу, обчислюють його середнє та стандартне квадратичне відхилення. Потім за допомогою формули відповідного теоретичного розподілу розраховують теоретичні частоти.

Для побудови нормального розподілу [3] нормалізують кінці класових інтервалів: від x_i та x_{i+1} переходять до $z_i = \frac{x_i - \bar{x}}{S}$ та $z_{i+1} = \frac{x_{i+1} - \bar{x}}{S}$ (тут $\bar{x}$ – середнє емпіричної вибірки, а S – її стандартне квадратичне відхилення). За допомогою інтегральної функції Лапласа обчислюють теоретичні імовірності потрапляння випадкової величини X до інтервалу (x_i, x_{i+1}): $P_i = \Phi(z_{i+1}) - \Phi(z_i)$. І, нарешті, знаходять теоретичні частоти: $n_i' = N \cdot P_i$. Тут N – обсяг емпіричної вибірки.

Інколи замість функції Лапласа використовують функцію щільності нормального розподілу. Тоді $n_i' = \frac{N * i}{S} \varphi(z_i)$, де i – величина класового інтервалу [5].

Критерій Колмогорова-Смирнова

Критерій Колмогорова-Смирнова застосовують до даних, які можна впорядкувати (цій вимозі не задовольняють номінативні дані). Розбіжності між теоретичним та емпіричним розподілами визначають за абсолютною величиною максимальної різниці між відповідними відносними накопиченими частотами.

При простій гіпотезі для малих вибірок (n<20) нульову гіпотезу (гіпотезу про те, що емпіричний розподіл не відрізняється від теоретичного) відхиляють, коли $|d_{emp}| \geq d_{kr}$ на обраному рівні значущості. Для великих вибірок обчислюють статистику $\lambda = | d_{max} | \cdot \sqrt{n}$ (або $d_{кр} = \frac{\lambda_{кр}}{\sqrt{n}}$) і нульову гіпотезу відхиляють, коли $\lambda_{емп} \geq \lambda_{кр}$. Критичні значення для d та λ подано у таблицях (Таблиця 20, Таблиця 21).

Таблиця 20

Критичні значення d_{max} [19]

Відмінності між емпіричним та теоретичним розподілами можна вважати достовірними, коли $|d_{emp}| \geq d_{kr}$.

n	p=0,05	p=0,01	n	p=0,05	p=0,01
5	0,6074	0,7279	50	0,1921	0,2302
10	0,4295	0,5147	60	0,1753	0,2101
15	0,3507	0,4202	70	0,1623	0,1945
20	0,3037	0,3639	80	0,1518	0,182
25	0,2716	0,3255	90	0,1432	
30	0,248	0,2972	100	0,1358	
40	0,2147	0,2574	>100	1,36/√n	1,63/√n

Таблиця 21

Квантилі розподілу Колмогорова λ_{1-p} [25]

p	0,30	0,25	0,20	0,15	0,10	0,05	0,02	0,01
$\lambda_{кр}$	0,97	1,02	1,07	1,14	1,22	1,36	1,52	1,63

Враховуючи, що критерій коректно працює не для згрупованих даних, а лише для неперервних розподілів, рівні значущості беруть “жорсткі”: p=0,2 або навіть p=0,3 [8, 25]. (У деяких статистичних критеріях обмежуються вказанням нижньої границі p-значення, достатньої для прийняття рішення. Наприклад, p<0,2 означатиме, що є підстави відхилити нульову гіпотезу принаймні на рівні значущості α=0,2. Навпаки,

значення p>0,2 (тобто 0,2 є нижньою границею p-значення) не дає підстав відхилити нульову гіпотезу.)

При складній гіпотезі (тобто у випадках, коли параметри теоретичного розподілу невідомі або за них приймають їхні оцінки, обчислені за досліджуваною вибіркою, слід використовувати модифіковану статистику[10]. У випадку нормального розподілу для обчислення критерію використовують формулу $D^* = d\left(\sqrt{n} - 0{,}01 + \frac{0{,}85}{\sqrt{n}}\right)$. Критичні значення статистики D^* наведено у наступній таблиці:

Таблиця 22

α	0,15	0,10	0,05	0,03	0,01
D^*	0,775	0,819	0,895	0,955	1,035

Для порівняння двох емпіричних розподілів статистику Колмогорова-Смирнова визначають за формулою $\lambda = d_{\max} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$, де d_{max}=максимальна за абсолютною величиною різниця між відповідними відносними накопиченими частотами розрядів порівнюваних розподілів; n_1 та n_2 – обсяги порівнюваних вибірок. Критичні значення статистики визначають за таблицею Таблиця 21.

Нульову гіпотезу відхиляють, коли $\lambda_{емп} \geq \lambda_{кр}$.

Критерії χ^2-Пірсона

Порівняння двох та більше емпіричних розподілів – майже єдиний спосіб дослідження номінативних та порядкових даних, хоча його застосовують і до числових даних також.

---

[10] Детальніше про особливості застосування статистики Колмогорова-Смирнова та помилки, які виникають при застосування критеріїв згоди див.:

1. Р 50.1.037–2002. РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II Непараметрические критерии Издание официальное.ГОССТАНДАРТ РОССИИ. Москва – 2002. http://ami.nstu.ru/~headrd/seminar/nonparametric/index.htm
2. Критерий Колмогорова-Смирнова <http://www.machinelearning.ru/wiki/index.php>

Критерій λ-Колмогорова-Смирнова тут можна застосувати лише до порядкових даних (даних, які можна впорядкувати за зростанням або спаданням). Порядок застосування не відрізняється від порівняння емпіричного розподілу з теоретичним, однак порівнювати можна лише два розподіли.

Критерії χ^2-Пірсона є найпотужнішим непараметричним критерієм. Його можна застосовувати як до числових, так і до номінативних даних. До того ж кількість порівнюваних розподілів не обмежується.

При порівнянні емпіричного розподілу з теоретичним необхідно розрахувати частоти відповідного теоретичного розподілу (див. стор.72).

Значенням критерію є наступна сума: $\chi^2 = \sum_{i=1}^{k} \frac{(f_{emp} - f_{teor})^2}{f_{teor}}$,

де k – кількість класових інтервалів (градацій) досліджуваної ознаки, f_{emp} та f_{teor} – відповідно емпірична та теоретична частоти, що відповідають визначеним градаціям змінної. Критичне значення визначають для кількості степенів вільності $df=k–r–1$, де k – кількість класових інтервалів, r – кількість оцінюваних параметрів розподілу. Наприклад, при порівнянні емпіричного розподілу з нормальним *проста* статистична гіпотеза полягатиме у тому, що досліджуваний (емпіричний) розподіл не відрізняється від нормального розподілу з середнім $\bar{x}$ та стандартним квадратичним відхиленням S (визначеними за емпіричною вибіркою). У такому разі $r=2$, тобто $df=k–3$.

При порівнянні декількох емпіричних розподілів виходять з міркування про те, що усі порівнювані розподіли сформовано з однієї генеральної сукупності. Наслідком такого міркування є те, що фактично емпіричні розподіли порівнюють не один з одним, а з теоретичними розподілами, частоти яких для кожного розподілу та розряду (класового інтервалу) визначають як долю відповідного значення досліджуваної ознаки у загальній сукупності емпіричних даних (див. **Приклад 4**, с. 108).

Обчислене емпіричне значення критерію порівнюють з критичним для кількості степенів вільності $df=(k–1)(m–1)$, де k – кількість розрядів (класових інтервалів або градацій ознаки), m – кількість порівнюваних розподілів.

Критерій χ^2 є надзвичайно чутливим до обсягу вибірок. По-перше значення $\chi^2_{\text{експ}}$ зростає пропорційно обсягу вибірки, відповідно зростає імовірність відхилити нульову гіпотезу незалежно від усіх інших факторів. По-друге, для малих вибірок критерій коректно працює, якщо в усіх класових інтервалах очікувані частоти не менші 5. Щоб задовольнити цю вимогу доводиться об'єднувати деякі класові інтервали. По-третє, Критерій χ^2 не застосовують, якщо дані подано відносними частотами, відсотками і т.п. [5].

Ще одним застосуванням критерія χ^2-Пірсона є перевірка гіпотези про відсутність зв'язку між ознаками (змінними). У випадку, коли змінні номінативні, параметричні або рангові критерії, згадані у лабораторній роботі №4, застосувати неможливо.

Для встановленні кореляційного зв'язку між якісними (номінативними) даними використовують такі критерії [22, 5, 18]:

для таблиці 2x2

– коефіцієнт спряженості: $Q = \dfrac{ad - cb}{ad + cb}$;

– коефіцієнт асоціації: $\Phi = \dfrac{ad - cb}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$;

для таблиці mxn

– φ-критерій Пірсона: $\varphi = \sqrt{\dfrac{\chi^2}{\chi^2 + N}}$;

– коефіцієнт Крамера $V = \sqrt{\dfrac{\chi^2}{N \cdot (k-1)}}$, де k – найменше з числа градацій двох змінних;

– коефіцієнт Чупрова $T = \sqrt{\dfrac{\chi^2}{N \cdot (c-1) \cdot (k-1)}}$, де c та k кількості градацій двох ознак.

Тут N – загальний обсяг вибірки, а χ^2 – результат порівняння емпіричних розподілів, що відповідають різним градаціям однієї з ознак.

Для критеріїв, що використовують χ^2, кореляція достовірна, якщо $\chi^2_{емп} \geq \chi^2_{крит}$. Кількість степенів вільності визначається як при порівнянні емпіричних розподілів.

Завдання1: Оцінка нормальності емпіричного розподілу.

1. За емпіричними вибірками А та В з теми 1 обчислити емпіричні квантилі (0.05, 0.10, ..., 0.95) –> P_k=ПЕРСЕНТИЛЬ (Емпіричний масив, К), де К=0.05, 0.10 ... 0.95.
2. Для кожного квантиля обчислити нормалізоване значення –> Z_k=НОРМСТОБР(К).
3. Побудувати точкову діаграму (P_k, Z_k).
4. Провести для отриманої сукупності точок лінійну лінію тренда. Візуально оцінити, наскільки графік (P_k, Z_k) відхиляється від прямої.
5. Зробити висновки про нормальність експериментальних вибірок та достовірність окомірного методу оцінки нормальності.
6. Обчислити асиметрію та ексцес для кожної вибірки. Порівняти емпіричний розподіл з нормальним за критичними значеннями асиметрії та ексцесу за критеріями Н.О.Плохінського та Є.І.Пустильника.
7. Порівняти отримані висновки з висновками, зробленими на основі застосування окомірного методу.
8. У пакеті SPSS побудувати квантильні графіки (Graphs –> Q-Q Plots) та графіки накопичених частот (Graphs –> P-P Plots). Порівняти значення асиметрії та ексцесу, обчислені процедурами дескриптивної статистики пакету SPSS, з їхніми похибками обчислення. Порівняти отримані результати з результатами обчислень та побудов у пакеті MS Excel.
9. Виконайти перевірку нормальності емпіричного розподілу за допомогою процедури пакета SPSS Analyze –> Descriptive Statistics –> Explore:

У вікні процедури перенесіть досліджувану змінну до списку залежних змінних (Dependent List). Якщо є можливість та потреба згрупувати значення залежної змінної за значеннями деякого фактора, наприклад “стать”, то перенесіть назву групуючої змінної до списку факторів (Factor List). У полі Display вкажіть Both, щоб отримати і графіки, і статистику.

Діаграма віток та листків (Stem-and-Leaf Plot) показує частоти емпіричного розподілу для інтервального статистичного ряду. Розбиття показано у вигляді віток (старші розряди числових границь інтервалів, які не змінюються на даному інтервалі) та листків (листок відповідає окремому значенню змінної).

На блочній діаграмі (Box Plot) представлено проміжок від першого до третього квартиля вибірки, відмічені мінімальне та максимальне значення і медіана. Екстремальні значення (викиди) будуть відмічені зірочками та кружечками.

При включенні прапорця Normality plots with tests буде проведено порівняння емпіричного розподілу з нормальним за тестом Колмогорова-Смирнова. Якщо значення похибки (Sig.) більше 0,05, то розподіл можна вважати нормальним.

10. Перевірити, як впливають на результати "теста нормальності" викиди, тобто екстремальні значення. Для цього "профільтрувати" вибірку за допомогою процедури Data –> Select Cases та виконати для відфільтрованих даних попередній пункт завдання.

Діалогове вікно процедури показано на Рис. 23. Для фільтації за умовою слід обрати перемикач "If condition …" та ввести умову відбору[11]. В результаті буде створено службову змінну filter_$. Далі при виконанні аналітичних процедур будуть враховуватися лише ті значення досліджуваної змінної, для яких filter_$=1.

11. Побудувати такі ж графіки у пакеті Statistica:
    Graphs –> 2d Graphs –> Normal Probability Plot
    –> Quantile-Quantile Plot
    –> Probability-Probability Plot.
12. Виконати у пакеті Statistica тест на нормальність емпіричного розподілу, за допомогою пунктів меню Statistica –> Basic Statistics/Tables –> Frequency Tables. На закладці Normality діалогового вікна Frequency Tables слід встановити

[11] За допомогою процедури Data –> Select Cases можна формувати випадкові вибірки на основі заданої. Для цього слід вказати порядок відбору "Random sample of cases".

прапорець Kolmogorov-Smirnov test та натиснути кнопку Test for normality.

13. Зробити висновки про нормальність досліджуваних даних. Порівняти результати, отримані засобами різних статистичних пакетів.

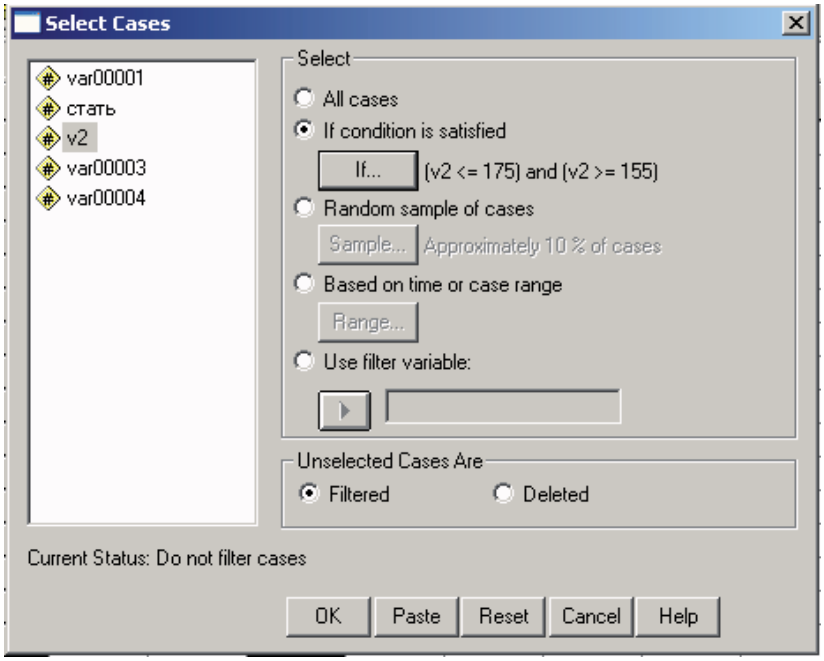


Рис. 23

Завдання 2: Порівняння емпіричного розподілу з деяким теоретичним.

Завдання 2а: порівняння емпіричного розподілу з теоретичним (нормальним):

1. За даними вибірок А та В з теми 1 розрахувати частоти нормального розподілу.

1.а) використовуючи функцію НОРМРАСП:

	A	B	C	D
1	X	Частота	Нормалізація	Частота нормального розподілу
2	111	3	=НОРМРАСП(A2,середнє, S, 0)	=C2*N*k
3	112	9	=НОРМРАСП(A3,середнє, S, 0)	=C3*N*k

Тут у стовпці А вказано значення X – верхню границю відповідного класового інтервалу.

У стовпці В – обчислені за формулою ЧАСТОТА частоти емпіричного розподілу.

У стовпці С – виконується нормалізація емпіричних значень за параметрами емпіричного розподілу – середнім та стандартним квадратичним відхиленням S (їх слід попередньо обчислити). При значенні четвертого параметра 0 (ХИБНЕ) формується значення вагової функції розподілу (функції щільності розподілу), а при значенні 1 (ІСТИНА) – інтегральної (накопичені відносні частоти). Для обрахунку частот слід обирати вагову функцію, тобто 0. У формулі можна посилатися на адреси комірок, які містять раніше обчислені параметри.

У стовпці D нормалізовані значення частот узгоджують з обсягом емпіричної вибірки, множачи на N – кількість значень в емпіричній вибірці, та k – розмір класового інтервалу. У наведеному прикладі k=1 (різниця між сусідніми значеннями у стовпці змінної X).

1.б) використовуючи функції НОРМАЛИЗАЦИЯ та НОРМСТРАСП:

	A	B	C	D	E	F	G
1	X_i	X_{i+1}	Z_i	Z_{i+1}	$\Phi(Z_i)$	$\Phi(Z_{i+1})$	$(\Phi(Z_{i+1})-\Phi(Z_i))*N$
2	110	111	$-\infty$				
3	111	112					

X_i – нижня границя класового інтервалу;

X_{i+1} – верхня границя класового інтервалу;

Z_i та Z_{i+1} – нормалізовані значення границь класового інтервалу.

У таблиці Z_i=НОРМАЛИЗАЦИЯ(X_i; середнє; відхилення). Найменше Z_i та найбільше Z_{i+1} заміняють відповідно на $-\infty$ та $+\infty$ (тут на деяке велике за абсолютною величиною значення).

$\Phi(Z_i)$ та $\Phi(Z_{i+1})$ – значення функції Лапласа.

У таблиці можна замість функції Лапласа застосувати інтегральну функцію стандартного нормального розподілу $F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{z^2}{2}}dz$ (F(z)=НОРМСТРАСП(z)), яка пов'язана

з функцією Лапласа рівнянням $F(z) = \frac{1}{2} + \Phi(z)$, оскільки константа не впливатиме на кінцевий результат.

Якщо ж використати інтегральну форму функції – НОРМРАСП (X_i,середнє, S, 1), то можна буде обійтись без нормалізації (не обчислювати Z_i та Z_{i+1}).

Отримані на останньому кроці (множення на N) частоти не рекомендується округлювати, щоб не знижувати достовірність подальших обчислень.

2. На одному графіку побудувати гістограми емпіричного та теоретичного розподілів.
3. Порівняти ряд розподілу частот за емпіричною вибіркою з отриманим теоретичним (нормальним) рядом:

а) за допомогою критерія χ^2 Пірсона (використовуйте функцію ХИ2ТЕСТ для порівняння та функцію ХИ2ОБР для отримання критичних значень критерія для визначеного рівня значущості та степенів вільності);

б) за алгоритмом обчислення λ-критерія Колмогорова-Смирнова (визначити вид статистичної гіпотези, зробити висновки про результати помилкового визначення виду статистичної гіпотези).

4. Зробити висновки про те, чи можна вважати емпіричний ряд розподілу нормальним.

Примітка: схожим чином емпіричний розподіл можна порівняти з теоретичними розподілами інших видів: лонг-нормальним, Стьюдента, Фішера, Вейбулла, експоненційним, гамма-, бета-, біноміальним, гіпергеометричним та Пуассона. Для цього відповідно слід застосувати одну з функцій: БИНОМРАСП(), СТЬЮДРАСП(), FРАСП(), ВЕЙБУЛЛ(), ЭКСПРАСП(), ПУАССОН(), ГАММАРАСП(), БЕТАРАСП(), ГИПЕРГЕОМЕТ(), ЛОГНОРМРАСП(),

Завдання 2б: порівняння емпіричного розподілу з теоретичним (рівномірним)

1. Для експериментальних вибірок з теми 1 обчислити теоретичні частоти ***рівномірного*** розподілу.

2. На одній діаграмі побудувати гістограми емпіричного та теоретичного розподілів частот.
3. За допомогою функцій ХИ2ТЕСТ та ХИ2ОБР перевірити, чи можна емпіричний ряд вважати рівномірним.
4. Порівняти емпіричний ряд з рівномірним, застосовуючи алгоритм обчислення критерію λ-Колмогорова-Смирнова.
5. Зробити висновки про рівномірність емпіричного розподілу. Порівняти результати застосування різних критеріїв.

Завдання 2в: порівняння емпіричного розподілу з теоретичним (рівномірним) в пакеті SPSS

1. Скопіювати значення вибірки В з лабораторної роботи 1.
2. Створити нову змінну.

2.а)За допомогою процедури Transform –>Count, значеннями якої будуть номери інтервалів інтервального статистичного ряду частот вибірки В. Для цього у вікні процедури Count вказати назву нової змінної (Target Variable), перенести до списку Numeric Variables досліджувану змінну та визначити границі інтервалів (Define Value), вказавши їх у пункті Range ***k*** through ***L*** (тут ***k*** – розмір інтервалу, а ***L*** – найбільше включене значення, тобто, наприклад, *"10, включаючи 140*" означатиме діапазон від 131 до140) або у пункті Range Lowest through ***L*** (діапазон включатиме значення, менші за L). Всі умови слід додати (Add) до списку Values to Count.

2.б) Або за допомогою процедури Transform –> Recode –> Into Different Variables. Для цього у вікні процедури Recode вказати вхідну змінну (Input Variables), назву нової змінної (Output Variable), відкрити вікно Old and New Values, в якому вказати нові значення або назви діапазонів аналогічно до процедури Count. Всі умови слід додати (Add) до списку Old –> New.

3. Застосувати до отриманої змінної непараметричний тест χ^2: Analyze – Nonparametric Tests – Chi-Square (All categories equal).
4. Зробити висновки про рівномірність емпіричного розподілу.
5. За розрахованою змінною побудувати гістограму розподілу: Graphs –> Bar Charts –> Simple.

6. Порівняти емпіричний розподіл з теоретичними (нормальним та рівномірним) за допомогою процедури Analyze –> Nonparametric –> 1-Sample K-S (критерій Колмогорова-Смирнова). Порівняти отриманий у даній процедурі результат з тестом нормальності процедури Analyze –> Descriptive Statistics –> Explore.

Застереження: як зазначено у [8], процедура Nonparametric –> 1-Sample K-S працює некоректно: p-значення визначається з припущення про те, що перевіряється проста гіпотеза, хоча насправді, вона складна (за параметри теоретичного розподілу прийнято їхні вибіркові оцінки).

7. Зробити висновки про рівномірність та нормальність експериментального розподілу. Порівняти результати з отриманими в MS Excel.

Завдання 2г: порівняння емпіричного розподілу з теоретичним (рівномірним) засобами пакету Statistica:

1. У програмі Statistica вибрати пункт меню Statistica –> Distribution Fitting. У діалоговому вікні Distribution Fitting обрати характер розподілу – неперервний.

Вибір неперервного (Continuous Distribution) розподілу дозволить виконати порівняння емпіричного з нормальним (Normal), рівномірним на проміжку (Rectangular), а також іншими (Exponential, Gamma, Log-normal, Chi-square, Weibull, Gompertz). Вибір дискретного розподілу (Descrete Distribution) дозволить виконати порівняння з біноміальним, геометричним, та розподілами Бернулі і Пуассона.

2. У діалоговому вікні Fitting Continuous Distributions вказати досліджувану змінну (Variable) та тип теоретичного розподілу (Distribution). На закладці Parameters (параметри) вказати кількість класових інтервалів для емпіричного розподілу (Numbers of categories), нижню границю (Lower limit) першого та верхню границю (Upper limit) останнього класового інтервалу.

Примітка: ці величини потрібно визначити самостійно відповідно до правил, наведених у темі1 (с. 6).

3. На закладці Options необхідно виконати установки, як показано на Рис. 24: слід встановити прапорець Combine Categiries (об’єднувати категорії) у розділі Chi-Square test,

встановити перемикач Yes (continuous) у розділі Kolmogorov-Smirnov test, обрати частотний графік, тобто графік-гістограму (Frequency distribution) та "сирі" частоти (Raw frequencyes) у розділі Graph.

4. Після налаштування слід натиснути кнопки Summary та Plot of observed and expected distribution на закладці Quick.
5. Проаналізувати отримані результати. Порівняти з отриманими в інших пакетах.

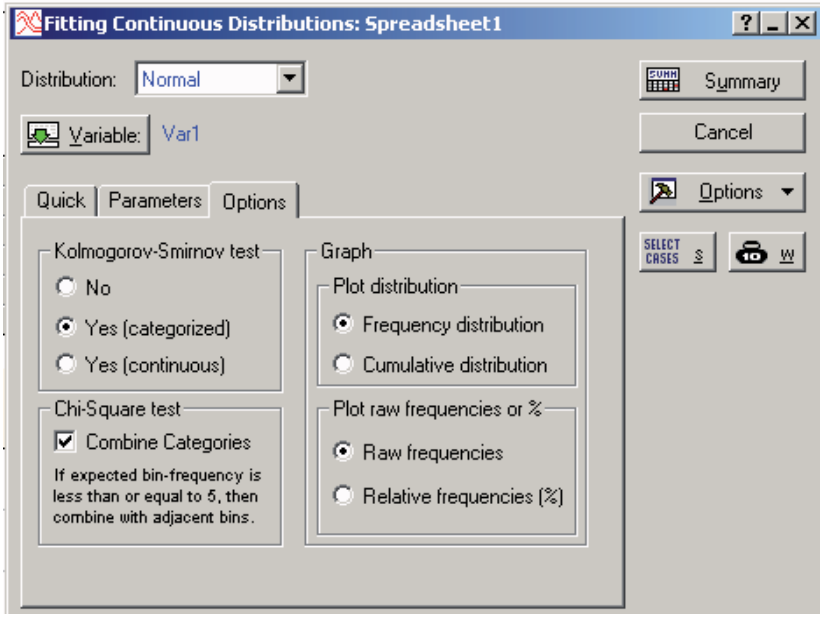


Рис. 24

Завдання 3: Порівняння двох та більше емпіричних розподілів.

1. За допомогою генератора випадкових чисел (функції СЛЧИС() або СЛУЧМЕЖДУ(A; B)) сформувати три ряди чисел однакового обсягу (50-80 значень): перший ряд – значення змінної A1 (у діапазоні 50-200 за інтервальною шкалою), другий ряд – значення змінної A2 (0-5 за порядковою шкалою), третій – значення змінної A3 (1-3 за номінативною шкалою). Скласти сюжет експериментального дослідження: придумати назви змінним та категоріям).

2. Скопіювати згенеровані **значення** на Аркуш2: за допомогою послуги ПРАВКА –> СПЕЦИАЛЬНАЯ ВСТАВКА вставити лише значення[12]. Сформувати статистичні гіпотези для порівняння емпіричних рядів розподілу.
3. Побудувати розподіли частот за ознакою А2 для рядів, визначених значеннями змінної А3 за допомогою послуги Данные –> Сводная таблица.

У зведеній таблиці значення змінної А2 використати для назв стовпців (розподіли), а значення змінної А3 – для назв рядків (градації ознаки). До даних використати підрахунок кількості.

Примітка: в таблиці не повинно бути порожніх комірок! Якщо таке сталося, відкоректуйте вхідні значення та оновіть зведену таблицю.

4. Порівняти між собою отримані емпіричні розподіли. Зробити висновки про відмінності між ними. Проінтерпретувати отримані результати в термінах сюжету.

У пакеті SPSS

5. Скопіювати значення змінних А1, А2, А3 з Аркуша2.
6. Побудувати таблицю спряженості для змінних А2 (Row) та А3 (Column). Застосувати для цього процедуру Analyze –> Descriptive Statistics –> Crosstabs. Обчислити критерій χ^2-Пірсона та коефіцієнт V Крамера (див. Завдання 4), зробивши відповідні настройки у вікні Crosstabs–>Statistics.
7. Побудувати діаграми порівнюваних розподілів. Graphs –> Bar Charts –> Clustered. Визначити як вісь категорій (Category Axis) – змінну А2, а групувати за змінною А3 (Define Clusters by)[13]. Зробити висновки про відмінності між розподілами. Проінтерпретувати результати у термінах експериментального сюжету.
8. Порівняти з результатом обчислення критерія χ^2-Пірсона в MS Excel.

[12] Генеровані випадкові числа мають властивість автоматично змінюватися.

[13] Або навпаки.

Завдання 4: Визначення зв'язку (кореляції) між якісними ознаками.

1. Засобами MS Excel порівняти експериментальні ряди даних (див. с. 211), обчислити кореляцію якісних ознак за формулою Крамера та зробити висновки про наявність або відсутність зв'язку між досліджуваними змінними. Проінтерпретувати отримані результати у термінах дослідження.
2. Побудувати діаграми порівнюваних розподілів.
3. У пакеті SPSS провести дослідження за даними файлу GSS93 subset.sav, який інсталюється разом з пакетом. Файл містить результати опитування громадської думки з низки питань. Зокрема "Чи потрібен закон про необхідність отримання в поліції дозволу на носіння вогнепальної зброї (Так/Ні)" (відповіді представлено у стовпці gunlaw). Дослідити зв'язок між відповідями на це запитання та статтю (sex), сімейним станом (marital) або віросповіданням (relig) опитуваних. Для цього вибрати процедуру Analyze –> Descriptive… –> Crosstabs та вказати в рядку (Row) змінну gunlaw (тобто залежну), а впливаючі змінні (тобто sex, relig, marital тощо) – у стовпці (Column). У вікні Statistics вибрати Chi-Square та Phi and Kramer's V. У вікні Cells вибрати Counts Observed та Expected (рахувати спостережувані та очікувані значення).
4. Зробити висновки про наявність або відсутність зв'язку між досліджуваними змінними.

У		К	Персентиль(Y;k)	Нормстобр(К)
122		0,05	139,6	-1,64
144		0,1	148,8	-1,28
152		0,15	152,4	-1,04
153		0,2	153,4	-0,84
155		0,25	155	-0,67
155		0,3	155	-0,52
156		0,35	155,6	-0,39
160		0,4	157,6	-0,25
163		0,45	160,6	-0,13
164		0,5	163	0,00
168		0,55	163,8	0,13
169		0,6	166,4	0,25
173		0,65	168,4	0,39
176		0,7	169,8	0,52
178		0,75	173	0,67
181		0,8	175,4	0,84
194		0,85	177,2	1,04
		0,9	179,2	1,28
		0,95	183,6	1,64

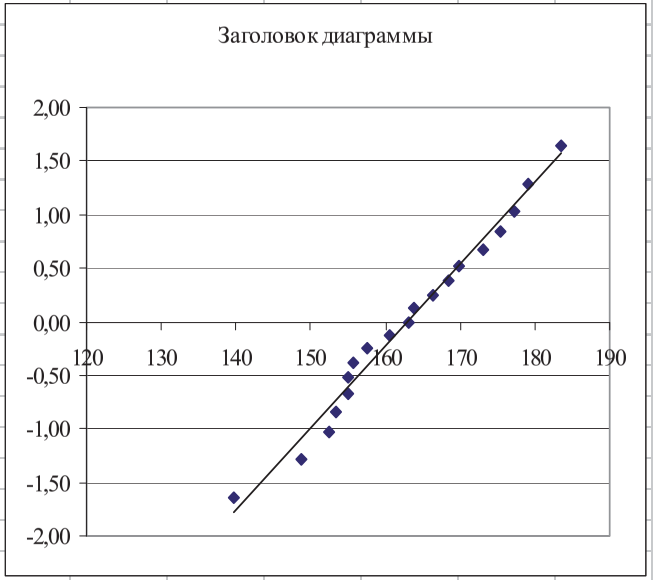


Рис. 25

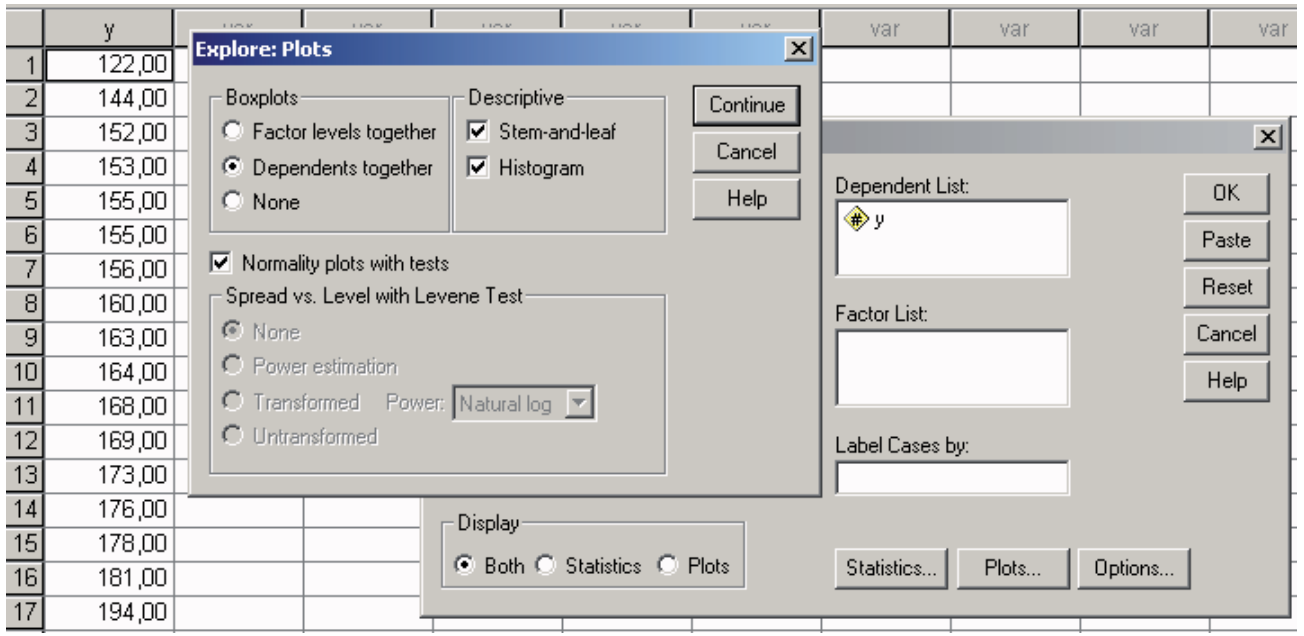

y
1 122,00
2 144,00
3 152,00
4 153,00
5 155,00
6 155,00
7 156,00
8 160,00
9 163,00
10 164,00
11 168,00
12 169,00
13 173,00
14 176,00
15 178,00
16 181,00
17 194,00
var
var
var
var
Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Continue
Cancel
Help
Normality plots with tests
Spread vs. Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
Display
Both
Statistics
Plots
Dependent List:
y
Factor List:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Options...

Рис. 26

Приклади виконання

Приклад 1: Нехай Y – час реакції на звук, виміряний у 17-ти досліджуваних [8, с. 104]. Необхідно з'ясувати, чи можна вважати, що змінна Y має нормальний розподіл?

Розв'язок. На рисунку (Рис. 25) у стовпці Y записано виміряні значення. У другому стовпці (К) записано проміжки розбиття одиничного інтервала на 20 рівних частин. (Одиничний інтервал можна розбити на довільну кількість частин, але зручно, щоб кількість проміжків була кратною 5, тобто 10, 20 тощо). У кожній комірці наступного (третього) стовпця записано результат обчислення формули ПЕРСЕНТИЛЬ(Y, K), де Y – масив значень із стовпця Y, а К – значення відповідного проміжку одиничного інтервала. Формула ПЕРСЕНТИЛЬ визначає, яке значення досліджуваної змінної поділяє усю вибірку у співвідношенні К/(1–К). У четвертому стовпці персентилі, які відповідають стандартизованому нормальному розподілу, обчислено за допомогою функції НОРМСТОБР(К). Тут К – відповідні значення з другого стовпця.

За даними третього та четвертого стовпців будують точкову діаграму. Чим краще побудовані точки вкладаються на пряму, тим ближчий емпіричний розподіл до нормального. Пряму для порівняння можна взяти регресійну (Диаграмма –> Добавить линию тренда).

За графіком видно, що емпіричні дані досить мало відхиляються від прямої, тобто розподіл можна вважати нормальним. Це підтверджується також результатами порівняння обчислених значень асиметрії та ексцеса з їхніми похибками та критичними значеннями (*перевірити самостійно*).

Отже до змінної Y можна буде застосувати без обмежень будь-які параметричні критерії.

Виконання завдання у пакеті SPSS полягає у введенні даних та виборі параметрів процедури. На рисунку (Рис. 26) показано, які слід вибрати настройки у вікні процедури Analyze –> Descriptive Statistics –> Explore та у вікні Plots цієїж процедури.

Результати будуть представлені таким чином:

1. Резюме по кількості та відносній чатоті валідних та пропущеих значень:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Y	17	100,0%	0	,0%	17	100,0%

2. Результати порівняння емпіричної вибірки з нормальною за тестом Лілльєфора (модифікацією теста Колмогорова-Смирнова) та тестом Шапіро-Уілкса (проводиться для вибірок з обсягом менше 50 значень).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Y	,142	17	,200*	,966	17	,713

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

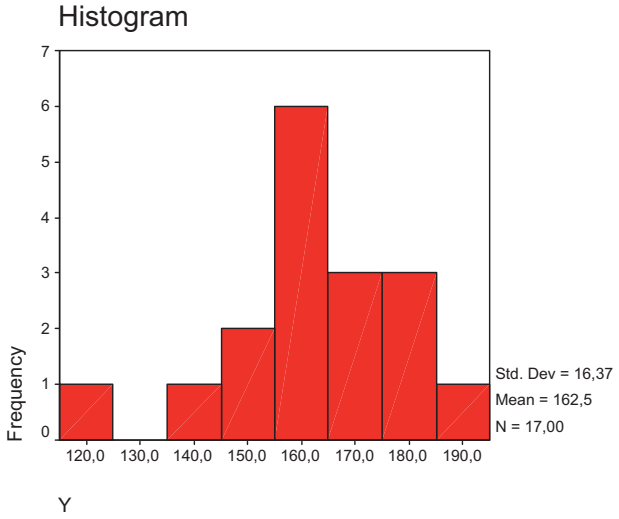
Наведено нижню границю p-значення (Sig.). Вона відповідає рівню значущості α=0,2. Це свідчить на користь того, що емпіричний розподіл можна вважати нормальним.

3. Статистичні характеристики вибірки (Глосарій, с. 188):

Descriptives

			Statistic	Std. Error
Y	Mean		162,5294	3,9697
	95% Confidence Interval for Mean	Lower Bound	154,1141	
		Upper Bound	170,9447	
	5% Trimmed Mean		163,0327	
	Median		163,0000	
	Variance		267,890	
	Std. Deviation		16,3673	
	Minimum		122,00	
	Maximum		194,00	
	Range		72,00	
	Interquartile Range		20,5000	
	Skewness		-,499	,550
	Kurtosis		1,479	1,063

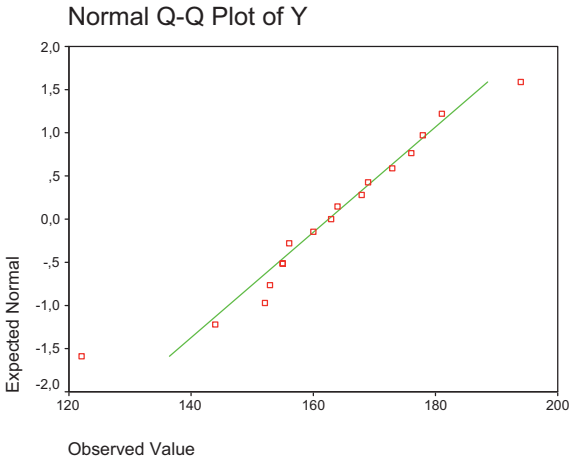
4. Гістограма інтервального статистичного ряду розподілу частот (границі та розміри інтервалів визначено автоматично).



5. Діаграма віток та листків.

Stem-and-Leaf Plots Y Stem-and-Leaf Plot Frequency Stem & Leaf 1,00 Extremes (=<122) 1,00 14 . 4 5,00 15 . 23556 5,00 16 . 03489 3,00 17 . 368 1,00 18 . 1 1,00 19 . 4 Stem width: 10,00 Each leaf: 1 case(s)	Оскільки досліджувана ознака має значення у діапазоні від 122 до 194, то автоматично визначено інтервали 140-149, 150-159 і т. д. Значення 122 позначено як екстремальне. Для решти ж визначено вітки: 14, 15 і т.д. відповідно, – та листки. Так на вітці 17 маємо листки 3, 6 та 8, тобто до інтервалу 170-179 потрапляють значення 173, 176 та 178. Відповідно частота даного інтервалу – 3. Отже ширина кожної вітки – 10. Кожен листок відповідає одному значенню.

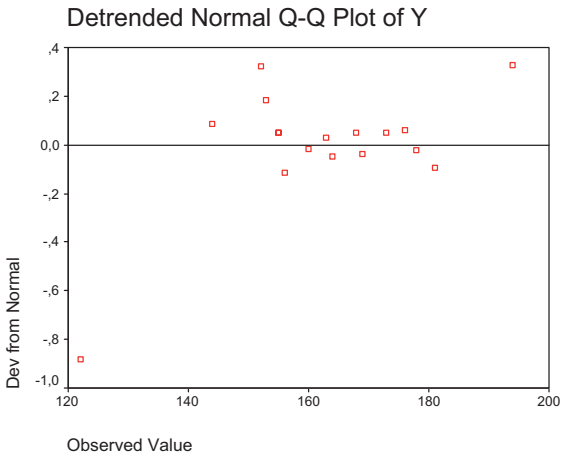
6. Діаграма нормального розподілу.



Для кожної точки на графіку координата X відповідає спостережуваному значенню, а координата Y – очікуваному нормальному. Якщо емпіричний розподіл нормальний, то усі точки будуть лежати на зображеній прямій.

7. Діаграма з виключеним трендом.

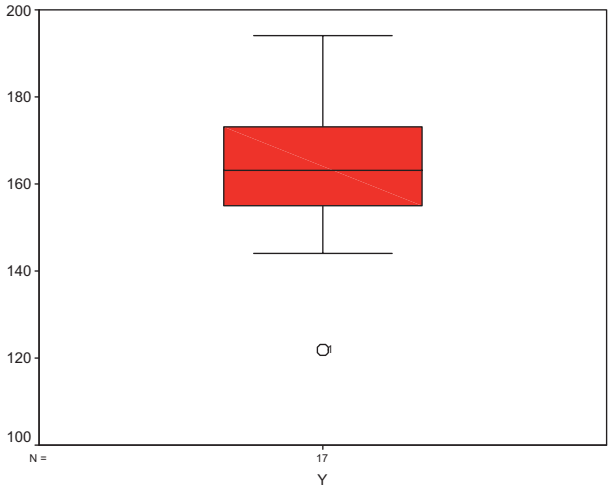
Detrended Normal Q-Q Plots



На ній зображено відхилення кожного емпіричного значення від відповідного очікуваного.

Для нормального розподілу усі точки мають вкладатися на горизонтальну вісь (Y=0).

8. Блочна діаграма (Box Plot):



На блочній діаграмі висота прямокутника відповідає ширині міжквартильного інтервалу, тобто інтервалу від 25 до 75 процентиля.

Горизонтальна лінія в середині прямокутника відповідає медіані.

“Вусиками” відмічено мінімальне та максимальне значення, а кружечком – значення 122 (викид, екстремальне значення).

У пакеті Statistica для тесту нормальності отримається аналогічний результат і графіки:

	Tests of Normality (Spreadsheet1)		
Variable	N	max D	K-S p
Var2	17	0,142362	p > .20

Приклад 2: В одному дослідженні порівнювали частоти народження хлопчиків в індіанських родинах англійського міста, де переважають мешканці-вихідці з Америки, із середньою частотою народження хлопчиків в Англії. Середня частота становила тоді 52%, а у досліджуваних родинах за період досліджень з 20 дітей, що народилися, виявилося 5 хлопчиків. Чи можна вважати, що в індіанських родинах

хлопчики народжуються достовірно рідше ніж в цілому по Англії?

Розв'язок. Спочатку слід сформувати статистичні гіпотези:

H_0: P=0,52 (дані вибірки узгоджуються з імовірністю народження хлопчиків p=0,52, тобто емпіричний розподіл не відрізняється від теоретичного).

H_1: P≠0,52 (дані вибірки не узгоджуються з імовірністю народження хлопчиків p=0,52, тобто емпіричний розподіл відрізняється від теоретичного не випадково).

Далі слід розрахувати частоти теоретичного розподілу, однакового за обсягом з емпіричним:

– теоретична частота народження хлопчиків у досліджуваній вибірці була б рівною 20*0,52=10,4;
– теоретична частота народження дівчаток у досліджуваній вибірці була б рівною 20–10,4=9,6.

Заносимо дані до таблиці та виконуємо обчислення за алгоритмом критерія χ^2-Пірсона. У даному випадку кількість степенів вільності *df*=2–1=1, оскільки розподіл має лише дві категорії. Тому слід буде застосувати поправку на неперервність (поправку Йетса). Вона полягає у тому, щоб модулі різниць чатот для кожної категорії зменшити на 0,5.

	Розподіли		R	R	Q	
Категорії	Емпір.	Теор.	$\lvert f_{емп}-f_{теор}\rvert$	R–0,5	R^2	$Q/f_{теор}$
Хлопчики	5	10,4	5,4	4,9	24	2,31
Дівчатка	15	9,6	5,4	4,9	24	2,5
Разом	20	20			Сума=	4,81

Поправку застосовують лише при **df=1** ↑

Отже отримано значення χ^2 =4,81. За допомогою функції Хи2Обр(вероятность, Степени_свободы) отримаємо табличні значення χ^2. $\chi^2_{0,05}$= Хи2Обр(0,05; 1)=3,84. $\chi^2_{0,01}$= Хи2Обр(0,01; 1)=6,63. Як бачимо, емпіричне значення знаходиться між критичними. Отже нульову гіпотезу можна відхилити тільки на рівні значущості α=0,05.

За допомогою функції Хи2Расп(X, Степени_свободы) отримаємо p-значення обчисленого критерію. Якщо воно менше, за визначений рівень значущості, то нульову гіпотезу можна буде відхилити. У даному випадку отримано p=0,03. Тобто на рівні значущості α=0,05 порівнювані розподіли дійсно відрізняються, і можна зробити висновок про те, що у

досліджуваний індіанських родинах хлопчики справді народжуються достовірно рідше ніж вцілому по Англії. Однак на більш строгому рівні (α=0,01) такого висновку зробити не можна.

При застосуванні до даних функції MS Excel Хи2Тест(Фактический интервал; Ожидаемый интервал), за *фактичний інтервал* слід прийняти емпіричні частоти, а за *очікуваний* – теоретичні частоти. Однак значення критерію тут будуть обчислені без поправки на неперервність, тому в результаті отримаємо p-значення =0,016 замість 0,03. Проте, висновок залишиться тим самим, що і при обчисленнях за алгоритмом: розподіли можна вважати різними лише на рівні значущості α=0,05.

Критерій Колмогорова-Смирнова до даного прикладу застосувати не можна, оскільки категорії розподілу суто номінативні і не підлягають будь-якому впорядкуванню.

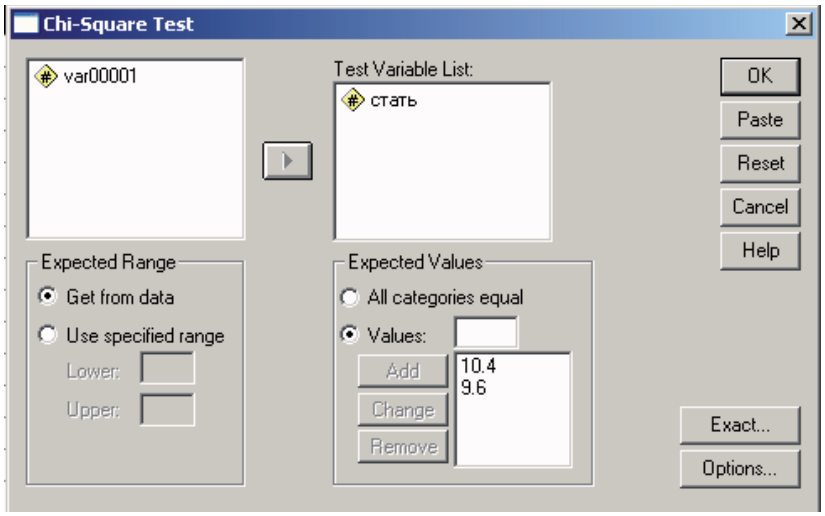


Рис. 27

Засобами пакету SPSS порівняння емпіричного ряду з теоретичним здійснюють за допомогою процедури Analyze –> Nonparametric Tests –> Chi-Square … Для її застосування дані прикладу слід подати у вигляді стовпця “стать” із значеннями 1 – хлопчик, 2 – дівчинка, наприклад (всього 20 значень). У вікні процедури Chi-Square Test (Рис. 27) слід

вказати змінну для тестування та очікувані частоти (Expected Values). Очікувати можна рівномірного розподілу (All categories equal) або заданого частотою кожного з можливих значень змінної. У даному прикладі можливі лише два значення змінної (1 або 2).

Результат застосування процедури буде таким:

стать	Observed N	Expected N	Residual
хлопчики	5	10,4	-5,4
дівчатка	15	9,6	5,4
Total	20		

Test Statistics

	стать
Chi-Square	5,841
df	1
Asymp. Sig.	,016

a 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 9,6.

Як можна бачити, значення критерію обчислено без поправки на неперервність. Отриманий результат такий самий, як і при застосуванні функції MS Excel. У примітці (a) зазначається, що у таблиці спряженості немає комірок з частотами менше за 5, тобто умови коректного застосування критерію виконано, отже результат є цілком достовірним.

Приклад 2а: Є відомості про зріст 53 дівчат – студенток першого курсу. З'ясувати, чи можна вважати, що зріст дівчат розподілено за нормальним законом?

178	170	167	165	164	164	163	161	157
175	170	167	165	164	164	163	161	157
174	168	166	165	164	164	162	160	155
173	168	166	165	164	163	162	160	154
173	167	166	165	164	163	161	158	
172	167	166	164	164	163	161	158	

Розв'язок. Спочатку слід обчислити середнє та стандартне квадратичне відхилення вибірки. Вони дорівнюють відповідно X=164,4 та S=5,14.

Дану вибірку зручно представити у вигляді інтервального статистичного ряду розподілу з величиною класового інтервалу i=5[14]. Початок першого інтервалу (152) та кінець останнього (181) замінено відповідно значеннями 0 та 200 для подальших обчислень.

	A	B	C	D	E
1	Границі інтервалів			Обчислені	Округлені
2	X_i	X_{i+1}	Емпіричні частоти	Теоретичні нормальні частоти	
3	0	156	3	2,728566	3
4	156	161	10	10,8109	11
5	161	166	26	19,52937	20
6	166	171	8	14,70097	15
7	171	176	5	4,602768	5
8	176	200	1	0,627425	1
9	Кількість	53			
10	Середнє	164,4			
11	Відхилення	5,14			

У стовці D (теоретичні нормальні частоти) записано формулу:

=(НОРМРАСП(A3;$A$10;$A$11;1)–НОРМРАСП(B3;$A$10;$A$11;1))*$A$9

Для застосування критерію χ^2 слід об'єднати деякі розряди і перейти до розподілу з чотирма розрядами, оскільки перший та останній розряди мають частоти менші за 5. В результаті матимемо:

Таблиця 23

Розряди		Частота емпірична	Частота теоретична	Частота теоретична округлена	$\frac{(f_e - f_t)^2}{f_t}$	d
152	161	13	13,5394686	14	0,071429	0,01018
162	166	26	19,5293696	20	1,8	0,11191
167	171	8	14,700968	15	3,266667	0,01452
172	181	6	5,23019379	6	0	0,00000
				Xi2=	5,138095 λ=	0,826655

В результаті застосування функції Хи2Тест до округлених значень теоретичних частот отримається р-значення = 0,16196. Такий же результат отримається при обчисленні за алгоритмом.

[14] Про порядок визначення класового інервалу див. с. 6.

Щоб переконатися в цьому слід застосувати функцію Хи2Расп до обчисленого значення критерія (χ^2 = 5,138095) або функцію Хи2Обр до отриманого р-значення, вказавши кількість степенів вільності *df=3*.

Однак слід пам'ятати, що для перевірки нормальності у даному випадку буде визначено лише одну степінь вільності (*df=4–3=1)*. В такому разі обчислене значення буде порівнюватися із критичним значенням $\chi^2_{0,05}$ = 3,841459 або $\chi^2_{0,01}$ = 6,634897. Тобто емпіричний розподіл не можна вважати нормальним.

Не складно, скориставшись отриманими частотами, обчислити різниці відносних накопичених частот (Таблиця 23, сьомий стовпець). Отримане таким чином максимальне значення d_{max}=0,11191 мало відрізняється від обчислених статистичними пакетами. Однак слід пам'ятати, що поскільки теоретичний розподіл розраховано з використанням вибіркового середнього та квадратичного відхилення, для визначення критичного значення та прийняття рішення буде застосовано статистику D^*. Обчислене значення дорівнює 0,826655. Згідно таблиці Таблиця 22 цьому значенню відповідає p<0,1 (що досить близько до результату теста нормальності процедури Explore пакета SPSS та процедури Distribution Fitting пакета Statistica). Отримане р-значення робить нормальність досліджуваної вибірки сумнівною (враховуючи "жорсткість" критерія).

Приклад 2б: Застосувати до даних з попереднього прикладу процедури пакету SPSS.

Застосування критерію Колмогорова-Смирнова не викликає труднощів. Однак потребує обережного використання. При застосуванні процедури Analyze –> Nonparametric Tests –> 1-Sample K-S за один прийом можна порівняти досліджувану ознаку з декількома теоретичними розподілами, а саме: з нормальним, рівномірним (Uniform), експоненційним та пуасонівським. У розглядуваному випадку результати будуть такими:

1) порівняння з рівномірним розподілом:

One-Sample Kolmogorov-Smirnov Test 2

		VAR00002
N		53
Uniform Parameters [a,b]	Minimum	152,00
	Maximum	178,00
Most Extreme Differences	Absolute	,234
	Positive	,234
	Negative	-,176
Kolmogorov-Smirnov Z		1,706
Asymp. Sig. (2-tailed)		,006

a. Test distribution is Uniform.

b. Calculated from data.

2) порівняння з нормальним розподілом:

One-Sample Kolmogorov-Smirnov Test

		VAR00002
N		53
Normal Parameters [a,b]	Mean	164,3774
	Std. Deviation	5,1374
Most Extreme Differences	Absolute	,116
	Positive	,116
	Negative	-,111
Kolmogorov-Smirnov Z		,846
Asymp. Sig. (2-tailed)		,472

a. Test distribution is Normal.

b. Calculated from data.

У першому випадку значення критерію Колмогорова-Смирнова = 1,706, а відповідне p-значення = 0,006. Тобто нульову гіпотезу слід відхилити, отже досліджуваний розподіл відрізняється від рівномірного.

У другому випадку наводиться значення критерію Колмогорова-Смирнова = 0,846, та відповідне p-значення = 0,472. Однак цей результат істотно відрізняється від отриманого процедурою Explore (див. с. 90) та отриманого згідно правила прийняття рішення (див. стор.74), оскільки неправомірно використовує просту гіпотезу.

Застосування процедури Analyze –> Nonparametric Tests –> Chi-square Test вимагає певних навичок, оскільки вона призначена для аналізу категоріальних, а не числових даних. Автоматично варіаційний ряд будуватиметься і частота вираховуватиметься для кожного значення досліджуваної змінної. Тому більшість з розрядів матимуть частоти менші за 5.

Для коректного застосування даної процедури слід попередньо підготувати (перекодувати) дані.

Поділ на категорії здійснюється дослідником (вікно процедури Transform –> Recode показано на Рис. 28).

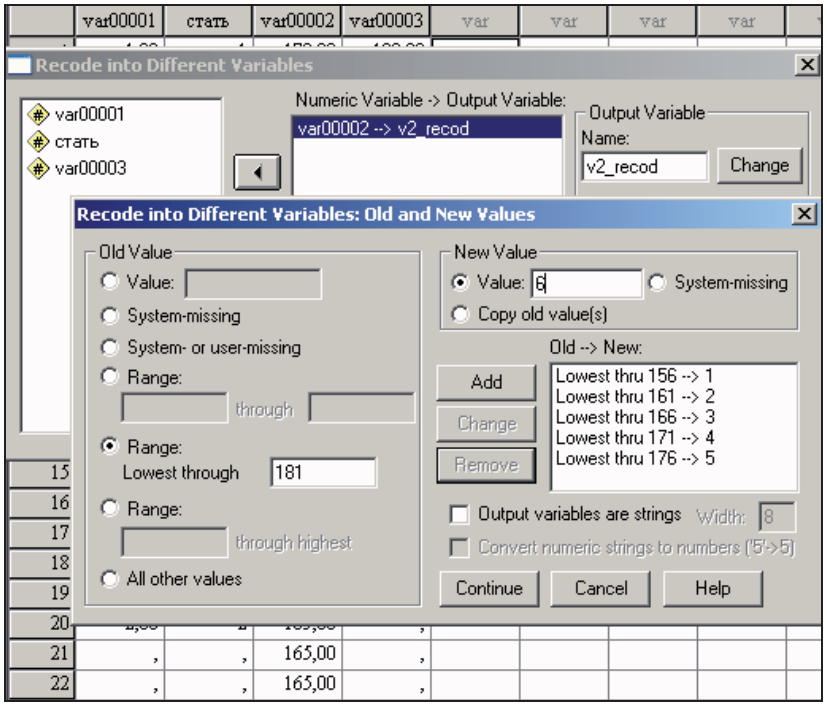


Рис. 28

Після поділу змінної, досліджуваної у даному прикладі, на категорії отримається нова змінна, до якої і слід застосувати процедуру Chi-square Test. При поділі на 6 категорії немає підстав відхиляти нульову гіпотезу:

V2_RECOD

	Observed N	Expected N	Residual
1,00	3	2,9	,1
2,00	10	10,6	-,6
3,00	26	19,3	6,7
4,00	8	14,5	-6,5
5,00	5	4,8	,2
6,00	1	1,0	,0
Total	53		

Test Statistics:	
	V2_RECOD
Chi-Square[a]	5,277
df	5
Asymp. Sig.	,383

a 3 cells (50,0%) have expected frequencies less than 5. The minimum expected cell frequency is 1,0.

При об'єднанні крайніх розрядів висновок буде менш категоричним:

V2_RECOD

	Observed N	Expected N	Residual
2,00	13	13,5	-,5
3,00	26	19,3	6,7
4,00	8	14,5	-6,5
,00	6	5,8	,2
Total	53		

Test Statistics

	V2_RECOD
Chi-Square[a]	5,257
df	3
Asymp. Sig.	,154

a 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 5,8.

Між тим тест нормальності з процедури Analyze –> Descriptive Statistics –> Explore показує значимі відмінності від нормальності:

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
VAR00002	,116	53	,072

a. Lilliefors Significance Correction

ідхилення від нормальності можна припустити також аналізуючи квантильний графік та графік остач нормального розподілу (Рис. 29).

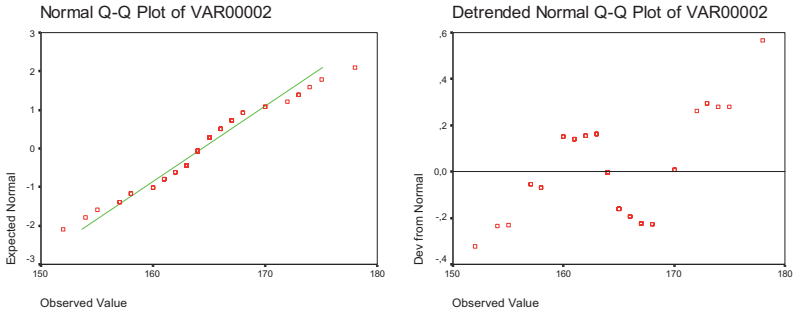


Рис. 29

Приклад 2г. Процедурою Statistics –> Distribution Fitting пакета Statistica автоматично буде визначено 17 інтервалів у діапазоні від 148 до 182. На Рис. 30 вказано також значення критеріїв χ^2-Пірсона та Колмогорова-Смирнова з поправкою Лілльєфорса. Обидва теста свідчать про сумнівність нульової гіпотези (значення критерія Колмогорова взагалі не визначено: p=n.s., тобто not specified – "не уточнене").

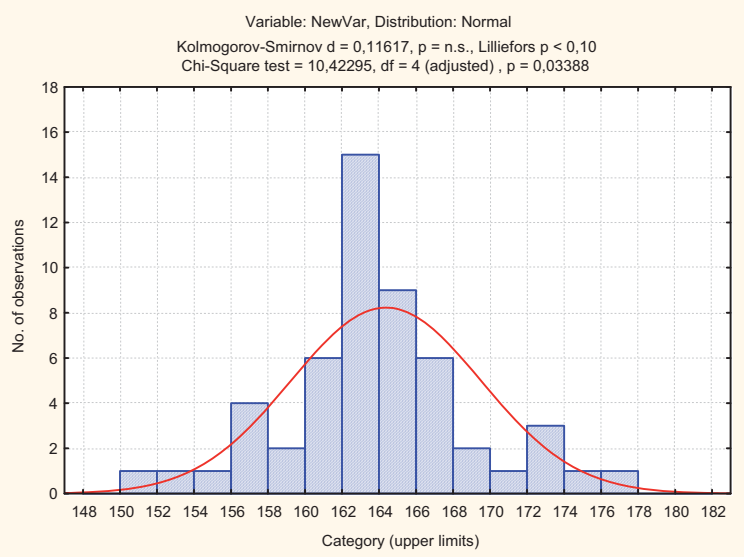


Рис. 30

При визначенні 6 інтервалів (як показано на Рис. 31) результати застосування критерія Колмогорова залишаться тими ж самими, а критерія χ^2 зміняться, що обумовлено особливостями його знаходження. На жаль, у наведеному прикладі довелося відмовитися від автоматичного об'єднання категорій для його обчислення (Рис. 32).

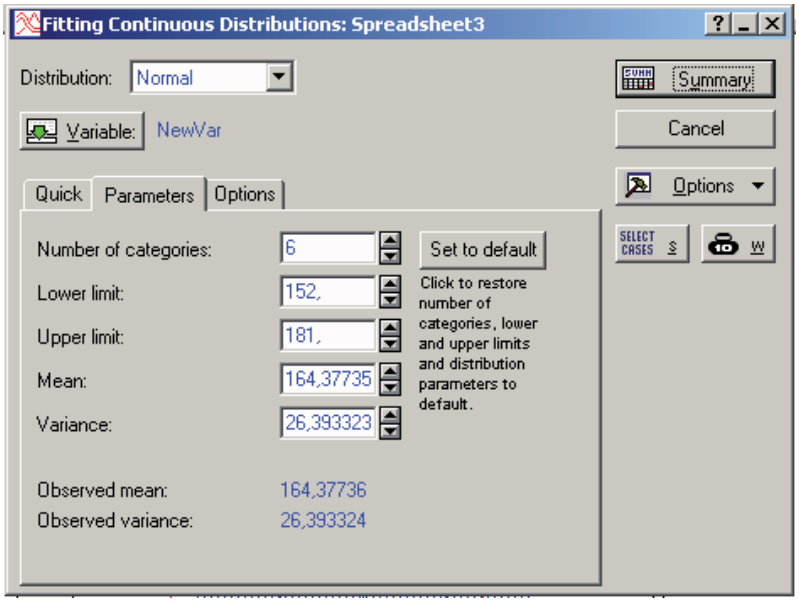

Fitting Continuous Distributions: Spreadsheet3
Distribution: Normal
Summary
Variable: NewVar
Cancel
Quick
Parameters
Options
Options
Number of categories: 6
Set to default
Lower limit: 152,
Click to restore number of categories, lower and upper limits and distribution parameters to default.
Upper limit: 181,
Mean: 164,37735
Variance: 26,393323
Observed mean: 164,37736
Observed variance: 26,393324

Рис. 31

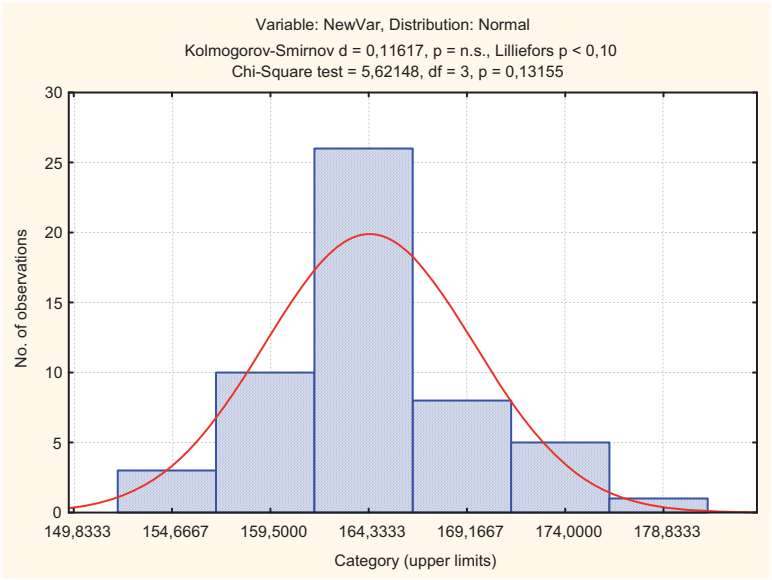

Variable: NewVar, Distribution: Normal
Kolmogorov-Smirnov d = 0,11617, p = n.s., Lilliefors p < 0,10
Chi-Square test = 5,62148, df = 3, p = 0,13155
No. of observations
30
25
20
15
10
5
0
149,8333
154,6667
159,5000
164,3333
169,1667
174,0000
178,8333
Category (upper limits)

Рис. 32

Якщо частоти емпіричного та теоретичного розподілів відомі, то до них можна також застосувати Chi-Square Test з пункту меню Statistics –> Nonparametric Statistics –> Observed versus expected X2. Для застосування такої процедури необхідно утворити дві змінні (ввести вручну), значеннями яких і будуть досліджувані частоти (див. Рис. 33).

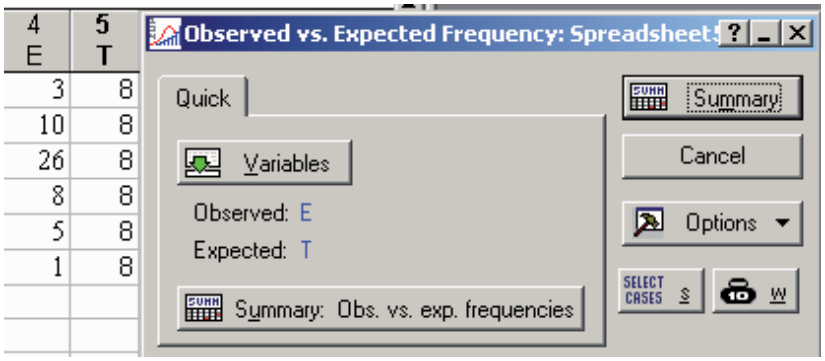


Рис. 33

Результати застосування процедури (Summary) представлено на таблиці Таблиця 24.

Таблиця 24

Case	Observed vs. Expected Frequencies (Spreadsheet5r) Chi-Square = 51,37500 df = 5 p < ,000000 NOTE: Unequal sums of obs. & exp. frequencies			
	observed E	expected T	O - E	(O-E)**2 /E
C: 1	3,00000	8,00000	-5,00000	3,12500
C: 2	10,00000	8,00000	2,00000	0,50000
C: 3	26,00000	8,00000	18,00000	40,50000
C: 4	8,00000	8,00000	0,00000	0,00000
C: 5	5,00000	8,00000	-3,00000	1,12500
C: 6	1,00000	8,00000	-7,00000	6,12500
Sum	53,00000	48,00000	5,00000	51,37500

Приклад 3: На Рис. 34 наведено приклад виконання завдання для 25 випадкових значень:

A1=ОКРУГЛ(СЛЧИС()*100+50;0);

A2=ОКРУГЛ(СЛЧИС()*2+1;0);

A3=ОКРУГЛ(СЛЧИС()+1;0).

Отримані значення (для A1 у діапазоні від 50 до 150, для A2 – 1, 2, 3; для A3 – 1, 2) можна інтепретувати, наприклад, як кількість тестових балів, вік та стать. У комірках A2:C26 наведено формули, а в комірках E2:G26 – скопійовані значення.

Зведену таблицю розміщено починаючи з комірки J1.

Для подальших обчислень значення цієї таблиці також зручно скопіювати, наприклад, у діапазон K10:M13.

У діапазоні K16:L18 розміщено формули для розрахунку теоретичних частот:

=$M10*K$13/$M$13	=$M10*L$13/$M$13
=$M11*K$13/$M$13	=$M11*L$13/$M$13
=$M12*K$13/$M$13	=$M12*L$13/$M$13

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	A1	A2	A3		A1	A2	A3			Количество по полю A1	A3						
2	99	2	2		59	2	2			A2	1	2	Общий итог				
3	82	3	2		56	1	1			1	3	3	6				
4	96	3	1		73	2	2			2	5	9	14				
5	108	3	1		64	2	1			3	3	2	5				
6	130	3	1		148	2	2			Общий итог	11	14	25				
7	55	3	1		117	2	2										
8	77	2	1		112	2	2										
9	102	3	1		51	3	2			**Емпіричний розподіл**	I	II			Fe*	Ft*	d
10	125	2	1		105	1	2			I	3	3	6		0,273	0,21429	0,0584
11	111	3	1		66	2	2			II	5	9	14		0,727	0,85714	0,1299
12	101	2	1		117	3	1			III	3	2	5		1	1	0
13	86	2	2		106	1	2				11	14	25				
14	134	3	1		140	2	2									dmax=	0,1299
15	142	1	2		68	2	1			**Теоретичний розподіл**	I	II				lambda=	0,3223
16	133	1	2		113	3	2			I	2,64	3,36	6				
17	81	2	2		72	1	1			II	6,16	7,84	14				
18	133	2	1		129	2	2			III	2,2	2,8	5				
19	141	2	2		70	3	1				11	14	25				
20	110	1	2		131	2	2										
21	140	1	1		139	3	1										
22	98	1	2		124	1	2				Xi2	p			V=	0,14122	
23	105	3	1		84	2	1				1	0,61					
24	79	2	1		136	2	1				5,99	0,05					
25	91	3	1		97	1	1				9,21	0,01					
26	53	2	2		104	2	1										

Рис. 34

В комірці L23 розміщено формулу =ХИ2ТЕСТ(K10:L12;K16:L18) для обчислення p-значення критерія χ^2, а у комірці K23 – формулу =ХИ2ОБР(L23;2). За цими результатами слід визнати, що розподіли, які відповідають

різним градаціям змінної A2 (або A3) статистично не відрізняються, тобто приймається нульова гіпотеза. Як наслідок, обчислене значення критерія Крамера (0,14122) є недостовірним, тобто зв'язку між досліджуваними ознакам не виявлено[15].

При застосуванні до тих самих змінних процедур пакета SPSS результати будуть такими, як показано нижче.

На Рис. 35 показано вікно процедури Crosstabs. На таблиці Таблиця 25 фактичні та очікувані (Expected) частоти. На таблиці Таблиця 26 – результату застосування критерія χ^2. А на Рис. 36 – гістограми "емпіричних" розподілів.

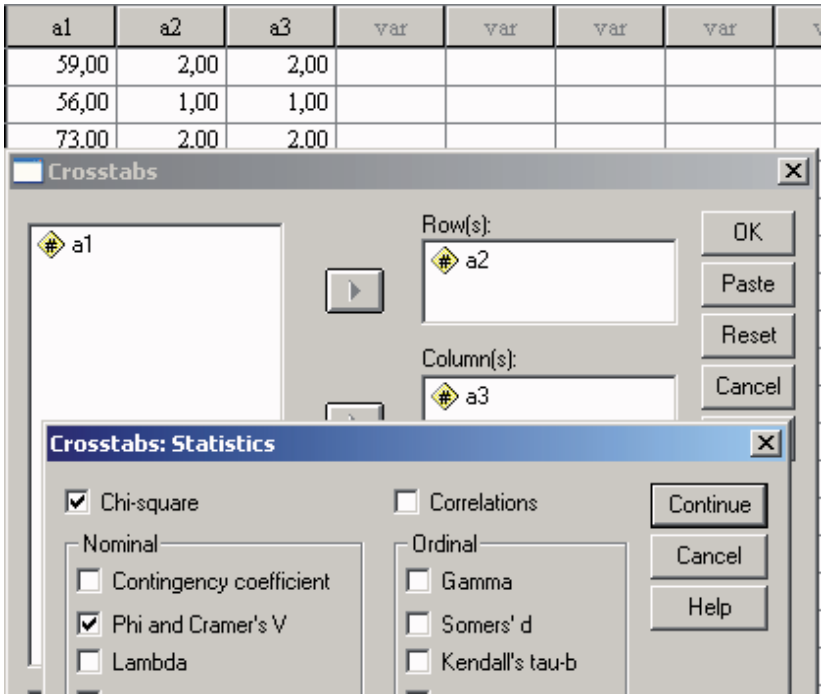


Рис. 35

[15] Для реальних вибірок такий результат може бути наслідком порушення репрезентативності, тому радять повторити дослід, збільшивши обсяг вибірки.

Таблиця 25

A2 * A3 Crosstabulation

			A3		
			1,00	2,00	Total
A2	1,00	Count	3	3	6
		Expected Count	2,6	3,4	6,0
	2,00	Count	5	9	14
		Expected Count	6,2	7,8	14,0
	3,00	Count	3	2	5
		Expected Count	2,2	2,8	5,0
Total		Count	11	14	25
		Expected Count	11,0	14,0	25,0

Таблиця 26

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	,997[a]	2	,607
Likelihood Ratio	,999	2	,607
Linear-by-Linear Association	,069	1	,793
N of Valid Cases	25		

a. 4 cells (66,7%) have expected count less than 5. The minimum expected count is 2,20.

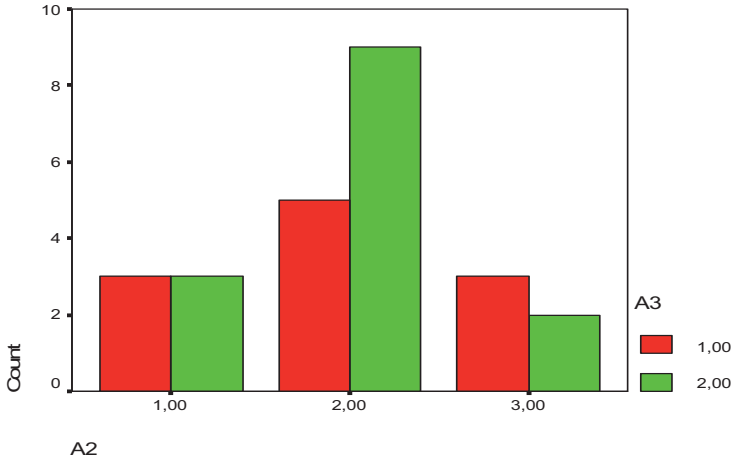


Рис. 36

У пакета Statistica таблиці спряженості будують та досліджують за допомогою процедур пункта меню Statistics –> Basic Statistics/Tables –> Tables and banners.

На закладці Crosstabulation за допомогою кнопки Specify tables (select variables) необхідно вибрати змінні-фактори (у даному прикладі це будуть змінні A2 та A3) та перейти далі, натиснувши OK.

У вікні Crosstabulation Tables Results на закладці Options вибрати тести, як буде застосовано до таблиці, наприклад, Pearson Chi-square та Cramer's V. Потім переглянути результати за допомогою послуг закладки Advanced: Summary – покаже отриману таблицю (Таблиця 27), Detailed two-way tables – результати застосування вказаних тестів (Таблиця 28), а решта – різноманітні графічні представлення досліджуваних розподілів.

Таблиця 27

	2-Way Summary Table: Observed F Marked cells have counts > 10		
A2	A3 1	A3 2	Row Totals
1	3	3	6
2	5	9	14
3	3	2	5
Totals	11	14	25

Таблиця 28

	Statistics: A3(2) x A2(3) (Spreadsheet5r)		
Statistic	Chi-square	df	p
Pearson Chi-square	,9972171	df=2	p=,60738
M-L Chi-square	,9994234	df=2	p=,60671
Phi	,1997215		
Contingency coefficient	,1958535		
Cramer's V	,1997215		

Приклад 4: За даними Міжнародної Організації Охорони Здоров'я серед людей, які не палять, захворюваність на рак легенів зустрічається у 12 випадках на 1000 осіб, а серед тих, хто палить – у 112 випадках. Перевірити, чи відрізняються дані розподіли? Чи можна вважати паління однією з причин захворювання на рак легенів (чи існує зв'язок між палінням та захворюванням на рак)?

Дані слід оформити таким чином:

	A	B	C	D
1	Експериментальні дані			
2		Хворі	Здорові	Разом
3	Палять	112	888	1000
4	Не палять	12	988	1000
5	Разом	124	1876	2000

Тут число здорових обчислюється за формулою:

=Разом–Хворі.

У даній таблиці значенням комірки С3 буде =D3-B3

Його слід скопіювати також до комірки С4.

Теоретичні частоти обчислюються із припущення про те, що якби розподіли відрізнялися випадково, тобто фактично не відрізнялися, то співвідношення хворих та здорових у кожній групі (тих, що палять, і тих, що не палять) було б пропорційним до їхнього співвідношення у загальній вибірці (тобто 124:1876).

До таблиці заносять формули. Результати обчислення наведено нижче.

Таблиця з формулами:

Теоретичні			
	Хворі	Здорові	Разом
Палять	=$D3*B$5/$D$5	=$D3*C$5/$D$5	=СУММ(G3:H3)
Не палять	=$D4*B$5/$D$5	=$D4*C$5/$D$5	=СУММ(G4:H4)
Разом	=СУММ(G3:G4)	=СУММ(H3:H4)	=СУММ(I3:I4)

Результати обчислення:

Теоретичні			
	Хворі	Здорові	Разом
Палять	62	938	1000
Не палять	62	938	1000
Разом	124	1876	2000

У формулах використано абсолютні та відносні адреси комірок.

Далі необхідно виконати обчислення за допомогою функції ХИ2ТЕСТ, параметрами якої є фактичний та теоретичний інтервали (масиви) даних, або обчислити за формулою: $\chi^2 = \sum \frac{\left(|f_e - f_t| - 0{,}5\right)^2}{f_t}$. Тут f_e – емпіричні частоти, f_t – теоретичні частоти, ***0,5*** – поправка Йєта (Yate) на неперервність, якщо df = 1 [Хили]. У даному випадку χ^2=84,26. А відповідне

значення коефіцієнта Крамера V = 0,145. Для коефіцієнта кореляції це значення невелике, однак при порівнянні емпіричного χ^2 з критичним значенням зв'язок досліджуваних ознак слід визнати достовірним, тобто гіпотезу H_0 відхилити.

У наведеній таблиці застосування функції =ХИ2ТЕСТ(B3:C4;G3:H4) дало результат 1,82165E-20. Тобто імовірність того, що випадкова величина може виявитися більшою за критичне значення критерію χ^2, дорівнює лише $1{,}8216 \cdot 10^{-20}$ (це значно менше за 0,05 та 0,01 – рівень значущості). Отже гіпотезу про однаковість розподілів (H_0) слід відкинути: вони відрізняються не випадково.

Оскільки досліджувався лише один впливаючий фактор – паління, – то треба визнати, що його вплив на захворюваність на рак досить значний.

Користуючись обчислювальними можливостями MS Excel, далі можна провести обчислювальний експеримент і, змінюючи кількість хворих або здорових, з'ясувати, при яких значеннях можна було б прийняти гіпотезу H_0.

На жаль, у пакетах SPSS та Statistica задачі до завдання №4 виконати не можна, оскільки ці пакети не працюють з готовими таблицями спряженості, а будують їх за даними (див. Приклад 3). Не можна також застосувати процедуру Statistics –> Nonparametric Statistics –> Observed versus expected X2 пакета Statistica, оскільки порівнювати фактично слід не емпіричні розподіли один з одним, а емпіричні частоти з теоретичними, які ще належить обчислити.

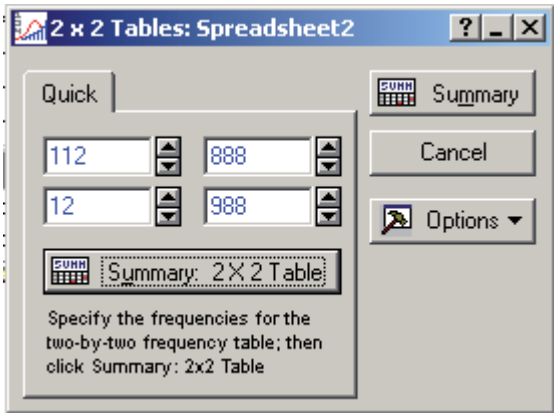


Рис. 37

Однак у пакеті Statistica є можливість досліджувати дані подані у вигляді таблиці 2x2 (див. **Приклад 4** на с. 108).

Для виконання дослідження таких таблиць застосовують процедуру Statistica –.> Nonparametric Statistics –> 2 x 2 Tables. Дані будуть вводитися так як показано на Рис. 37.

Результати представлено на таблиці нижче:

Таблиця 29

	2 x 2 Table (Spreadsheet2)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	112	888	1000
Percent of total	5,600%	44,400%	50,000%
Frequencies, row 2	12	988	1000
Percent of total	,600%	49,400%	50,000%
Column totals	124	1876	2000
Percent of total	6,200%	93,800%	
Chi-square (df=1)	85,98	p=0,0000	
V-square (df=1)	85,93	p=0,0000	
Yates corrected Chi-square	84,26	p=0,0000	
Phi-square	,04299		
Fisher exact p, one-tailed		----	
two-tailed		----	
McNemar Chi-square (A/D)	696,02	p=0,0000	
Chi-square (B/C)	850,69	p=0,0000	

Контрольні запитання

1. Для вирішення яких статистичних задач використовують критерій χ^2-Пірсона? Які обмеження (умови) його застосування?
2. Які гіпотези називають простими, а які складними?
3. Як визначити кореляцію якісних ознак (залежність/ незалежність декількох емпіричних розподілів)?
4. Які умови застосування критерія Колмогорова-Смирнова?
5. Для чого перевіряти нормальність розподілу? Як визначити, чи узгоджується емпіричний розподіл з нормальним?
6. Які критерії називають непараметричними? Чому?
7. Що таке розподіл? Які бувають розподіли? Як зобразити емпіричний розподіл графічно?
8. Назвіть характерні ознаки нормального та рівномірного розподілів.

Тема 6: “Непараметричні методи”

Мета:

Студенти повинні знати:

– особливості вимірювання номінативних та порядкових змінних;
– правили ранжування порядкових та метричних даних;
– класифікацію задач, в яких застосовують непараметричні методи ;
– переваги та особливості застосування непараметричних методів;
– способи непараметричного порівняння незв’язаних вибірок;
– способи непараметричного порівняння зв’язаних вибірок.

Студенти повинні уміти:

– подавати експериментальні дані у вигляді, зручному для обчислень;
– виконувати ранжування числових даних;
– аналізувати та інтерпретувати отримані результати;
– використовувати непараметричні методи, реалізовані у пакетах SPSS та Statistica.

Теоретичні відомості

Непараметричними називають статистичні методи, обчислювальні процедури яких не використовують параметри розподілу, а спираються на результати ранжування або простого підрахунку значень досліджуваної ознаки. Серед переваг непараметричних методів найбільш суттєвими є:

1) нечутливість до виду емпіричного розподілу імовірності випадкової величини (вид розподілу не має значення, тоді як переважна більшість параметричних методів коректно працює лише на нормально розподілених даних);
2) можливість застосування на малих вибірках, для яких неможливо перевірити відповідність нормальному розподілу;
3) досліджувану ознаку обчислено за номінативною або порядковою шкалою;
4) нечутливість до екстремальних значень.

Однак параметричні методи значно чутливіші за непараметричні, тому застосування непараметричних методів виправдане лише у випадках, коли параметричні методи застосувати не можна або для первинного аналізу даних.

При застосуванні непараметричних методів використовують таку ж класифікацію задач статистичного аналізу, як і для

параметричних методів:

Таблиця 30

Умови застосування	Непараметричні методи	Параметричні методи
1. Виявлення різниці рівнів досліджуваної ознаки (незв'язані вибірки)		
а) 2 вибірки	U – критерій Манна-Уітні	t-критерій Стьюдента
б) 3 та більшее вибірок	S – критерій тенденцій Джонкіра; H – критерій Крускала-Уолліса.	однофакторний дисперсійний аналіз
2. Оцінка зсуву значень досліджуваної ознаки (зв'язані вибірки)		
а) 2 заміри на одній вибірці	W-критерій Вілкоксона; G-критерій знаків;	t-критерій Стьюдента
б) 3 та більшее замірів на одній вибірці	χ_x^2 -критерій Фрідмана; L-критерій тенденцій Пейджа.	однофакторний дисперсійний аналіз
3. Виявлення відмінностей у розподілі ознаки		
а) при співставленні емпіричного розподілу з теоретичним	χ^2 -критерій Пірсона; λ -критерій Колмогорова-Смирнова	
б) при співставленні двох емпіричних розподілів	χ^2 -критерій Пирсона; λ -критерій Колмогорова-Смирнова;	
4. Виявлення міри узгодженості змін		
а) двох ознак	r_s -коефіцієнт рангової кореляції Спірмена; τ -Кендала.	r_{xy} -коефіцієєнт кореляції Пірсона
б) двох ієрархій або профілів		однофакторний дисперсійний аналіз
5.Аналіз змін ознаки під впливом контрольованих умов (факторів)		
а) під впливом одного фактора	S-критерій тенденцій Джонкіра; L-критерій тенденцій Пейджа.	однофакторний дисперсійний аналіз
б) під впливом двох факторів одночасно		двофакторний дисперсійний аналіз

Деякі з перелічених методів були розглянуті у попередніх лабораторних роботах. Тут розглянемо непараметричні методи до задач першої, другої та п'ятої груп.

Зсув досліджуваної ознаки (2 зв'язані вибірки)

Необхідність виявлення зсуву досліджуваної ознаки виникає тоді, коли дослідження проводиться на одній групі вибірці але у різних умовах.

G-критерій знаків призначений для виявлення загального напрямку зсуву: зменшилися чи збільшилися показники при переході від першого заміру до другого. Але критерій знаків не дозволяє оцінити інтенсивність зсуву (наскільки сильно збільшилися або зменшилися показники). Інтенсивність зсувів дозволяє оцінити ***W-критерій Вілкоксона***.

При застосуванні критерію знаків спочатку визначають напрямок "типового" зсуву, тобто такого який здається переважаючим. "Нульові" зсуви, тобто такі, при яких результати першого та другого заміру не змінилися, у критерії знаків не враховуються.

Кількість спостережень в обох замірах не менше 5 та не більше 300.

Статистичні гіпотези формулюються наступним чином:

H_0: Переважання типового напрямку зсуву є випадковим.

H_1: Переважання типового напрямку зсуву не випадкове.

Алгоритм 1: *G-критерій знаків*

1. Для кожного досліджуваного визначити напрямок змін. Збільшення позначити знаком "+", зменшення – знаком "–", відсутність змін знаком "0".
2. Виключити нульові зсуви з подальших обрахунків та зменшити обсяг вибірки на кількість нулів.
3. Визначити "типовий" (переважний) напрямок зсуву.
4. Підрахувати кількість "нетипових" зсувів. Вважати це число емпіричним значенням G.
5. Для даного n визначити критичні значення G за спеціальною таблицею.
6. Якщо $G_{емп} > G_{кр}$, то зсув недостовірний, гіпотезу H_0 відхилити не можна.

W-критерій Вілкоксона більш потужний за критерій знаків, оскільки дозволяє визначати не лише напрямок, а й інтенсивність (вираженість) змін.

За типовий напрямок змін (збільшення/зменшення) також приймають той, який частіше зустрічається у вибірці. Обсяг вибірки (від 5 до 50 значень) обмежений наявним таблицями критичних значень.

Статистичні гіпотези формулюються наступним чином:

H_0: Інтенсивність зсуву у типовому напрямку не більша за інтенсивність зсуву у нетиповому напрямку.

H_1: Інтенсивність зсуву у типовому напрямку більша за інтенсивність зсуву у нетиповому напрямку.

Нульові зсуви виключаються (якщо нульові зсуви не виключати, то напрямлену гіпотезу слід замінити ненапрямленою: наприклад, "зсув у бік збільшення значень не перевищує зсув у бік зменшення або збереження значень на тому ж рівні").

Алгоритм 2. *W-критерій Вілкоксона*

1. Скласти таблицю значень досліджуваних у першому та другому замірах.
2. Обчислити різниці між індивідуальними результатами першого та другого заміру. Визначити "типовий" напрямок та сформулювати гіпотези.
3. В окремий стовпець виписати абсолютні значення різниць.
4. Проранжувати абсолютні значення, приписуючи найменшому значенню ранг 1. Перевірити правильність нарахування рангів.
5. Відмітити ранги, що відповідають "нетиповим" зсувам.
6. Підрахувати їхню суму (W).
7. Для даного n визначити за таблицею критичних значеннь $W_{кр}$.
8. Якщо $W_{емп} \leq W_{кр}$, то гіпотезу H_0 можна відхилити: зсув у "типовому" напрямку за інтенсивністю достовірно переважає.

Однофакторний аналіз (зв'язані вибірки)

Критерій χ_r^2 Фрідмана дозволяє встановити, що величини показників від умови до умови змінюються, але не вказує напрямок змін.

Критерій застосовують, коли вибірок не менше 3-х, причому кожна містить не менше 2-х значень.

При великих обсягах вибірок отримані значення χ_r^2 співставляють з критичними значеннями χ^2-Пірсона для кількості степенів вільності df = m – 1. При кількості вибірок m>3 та кількості значень в них n≤9, або при m =4 та n≤4 – за спеціальними таблицями.

Статистичні гіпотези формулюються наступним чином:

H_0: між показниками, отриманими у різних умовах існують лише випадкові відмінності.

H_1: відмінності між показниками, отриманими у різних умовах, не випадкові.

Алгоритм 3. *Критерій χ_r^2 Фрідмана*

1. Проранжувати індивідуальні значення першого досліджуваного, отримані у першому, другому і т.д. замірах.
2. Виконати таке ж ранжування для усіх інших досліджуваних.
3. Підрахувати суми рангів окремо для кожного заміру (стовпця). Перевірити правильність нарахування рангів.
4. Обчислити емпіричне значення критерію за формулою: $\chi_r^2 = \left[\frac{12}{n \cdot m \cdot (c+1)} \cdot \sum\left(T_j^2\right)\right] - 3 \cdot n \cdot (m+1)$, де c – кількість умов, n – кількість досліджуванихоб'єктів, T_j – суми рангів за кожною з умов.
5. Визначити рівні статистичної значущості для χ_r^2 відповідно до обмежень методу.
6. Якщо $\chi_{r\,емп}^2 \leq \chi_{r\,кр}^2$, то відмінності достовірні, тобто H_0 можна відкинути.

L-критерій тенденцій Пейджа крім констатації відмінностей між декількома замірами, дозволяє визначити напрямок змін.

Мінімальна кількість досліджуваних об'єктів – 2 не менш як для трьох замірів. Максимальна кількість – 12 досліджуваних та 6 умов (n≤12, m≤6).

При обчислення стовпці слід впорядкувати за зростанням рангових сум. За першу вибірку беруть ту, у якій рівень ознаки здається вищим.

Статистичні гіпотези формулюються наступним чином:

H_0: Збільшення індивідуальних показників при переході від першої умови до другої і далі – випадкове.

H_1: Збільшення індивідуальних показників при переході від першої умови до другої і далі – не випадкове.

Алгоритм 4. *L-критерій тенденцій Пейджа*

1. Проранжувати індивідуальні значення першого досліджуваного об'єкта, отримані ним у першому, другому, третьому і т.д. замірах (порядок досліджуваних у таблиці – довільний).
2. Виконати те саме для інших об'єктів.
3. Додати отримані ранги по стовпцях (умовах замірів). Перевірити правильність нарахування рангів.
4. Розташувати стовпці (умови) у порядку зростання рангових сум.
5. Обчислити емпіричне значення L за формулою: $L=\Sigma(\mathbf{T_j \cdot j})$, де j – порядковий номер стовпця (умови), T_j – сума рангів за даною умовою (стовпцем).
6. За спеціальною таблицею визначити критичне значення L. Якщо $L_{емп} \geq L_{кр}$, тенденція достовірна, тобто H_0 можна відхилити.

Порівняння двох незалежних вибірок

U-критерій Манна-Уітні виявляє відмінності між вибірками за оцінкою ширини спільної для обох вибірок зони значень. Чим вужча ця зона, тим імовірніші відмінності.

Застосовується до змінних, вимірюється за шкалою не нижче порядкової. При обсягах вибірок n_1, $n_2>20$ доцільніше застосовувати кутове перетворення Фішера у комбінації з критерієм λ-Колмогорова-Смирнова [19].

За першу вибірку беруть ту, у якій рівень ознаки здається вищим, а за вибірку 2 – ту, в якій він за попередньою оцінкою здається нижчим.

Статистичні гіпотези формулюються наступним чином:

H_0: Рівень ознаки у вибірці 2 не нижчий за рівень ознаки у вибірці 1.

H_1: Рівень ознаки у вибірці 2 нижчий за рівень ознаки у вибірці 1.

Алгоритм 5: ***U-критерій Манна-Уітні***

1. Впорядкувати значення в кожній вибірці за зростанням.

2. Проранжувати значення у двох вибірках так, ніби то вони утворюють одну загальну, приписуючи найменшому значенню ранг 1, а найбільшому – (n_1+n_2).
3. Підрахувати суму рангів окремо для кожної вибірки. Перевірити правильність нарахування рангів.
4. Визначити більшу з двох рангових сум (T_x).
5. Обчислити значення критерію U за формулою: $U = (n_1 \cdot n_2) + \frac{n_x \cdot (n_x + 1)}{2} - T_x$, де n_1 – кількість досліджуваних об'єктів у вибірці 1, n_2 – кількість досліджуваних об'єктів у вибірці 2, T_x – більша з двох рангових сум, n_x – кількість досліджуваних у групі з більшою сумою рангів.
6. Визначити критичні значення U за спеціальною таблицею. Якщо $U_{емп} > U_{кр\ 0,05}$, то приймається гіпотеза H_0. Чим значення U менше, тим достовірніша різниця між вибірками.

Однофакторний аналіз (незв'язані вибірки)

H-критерій Крускала-Уолліса є продовженням критерія *U-критерій Манна-Уітні*: виявляє відмінності між трьома, чотирма і т.д. вибірками за оцінкою ширини спільної для них зони значень. Чим вужча ця зона, тим імовірніші відмінності.

Застосовується до змінних, вимірюється за шкалою не нижче порядкової. При великих обсягах та кількостях вибірок критерій асимтотично наближається до критерія χ^2-Пірсона, тому отримані значення співставляють з критичними значеннями χ^2-Пірсона для кількості степенів вільності df = m – 1, де m – кількість вибірок.

Мінімальна кількість об'єктів у групі – 3 (при трьох вибірках – 2).

Статистичні гіпотези формулюються наступним чином:

H_0: Відмінності рівня досліджуваної ознаки між групами випадкові.

H_1: Відмінності рівня досліджуваної ознаки між групами не випадкові (істотні).

Алгоритм 6: *H-критерій Крускала-Уолліса*

1. Впорядкувати значення в кожній вибірці за зростанням.

2. Проранжувати значення всіх вибірок так, ніби то вони утворюють одну загальну, приписуючи найменшому значенню ранг 1, а найбільшому – ($n_1+n_2+\ldots+n_m$).
3. Підрахувати суму рангів окремо для кожної вибірки. Перевірити правильність нарахування рангів.
4. Визначити більшу з двох рангових сум (T_x).
5. Обчислити значення критерію H за формулою: $H = \left[\frac{12}{N(N+1)} \cdot \sum \frac{T_j^2}{n_j} \right] - 3 \cdot (N+1)$, де N – загальна кількість об'єктів в об'єднаній вибірці, n_j – кількість досліджуваних об'єктів у j-тій вибірці, T_j – сума рангів у j-тій вибірці.
6. Якщо кількість груп більша 3 та обсяги вибірок $n_j > 5$, то порівняти отримані значення H з критичними значеннями χ^2-Пірсона для df=m–1 (m – кількість вибірок), інакше – з критичними значеннями H за спеціальною таблицею. Якщо отримане значення більше за критичне, то нульову гіпотезу слід відкинути.

S-критерій тенденцій Джонкіра дозволяє виявити тенденцію змін ознаки від вибірки до вибірки.

Якщо вибірки сформовані на основі деякої кількісної харакеристики, то впорядковуючи їх за допомогою S-критерія тенденцій Джонкіра, можна встановити ще й міру зв'язку між двома кількісними змінними, що доцільно, коли одна зознак варіює у вузькому діапазоні значень.

Обсяги вибірок мають бути однаковими: якщо обсяги різні, то їх вирівнюють, використовуючи таблицю випадкових значень[16].

Статистичні гіпотези формулюються наступним чином:

H_0: Тенденція зростання значень ознаки від вибірки до вибірки випадкова.

---

[16] Випадкові числа можна згенерувати в MS Excel за допомогою функції СЛЧИС, в SPSS обчисленням нової змінної (Transform –> Compute) за допомогою виразу RND(UNIFORM(N)) та в пакеті Statistica при створенні нової змінної задати математичну функцію (Function) *Rnd(x*) або *Uniform(x)* – випадкове число з діапазону від 0 до x.

H_1: Тенденція зростання значень ознаки від вибірки до вибірки не випадкова.

Алгоритм 7: *S-критерій тенденцій Джонкіра*

1. Вирівняти обсяги порівнюваних вибірок, відкинувши випадковим чином зайві значення.
2. Впорядкувати кожну групу за зростанням досліджуваної ознаки.
3. Впорядкувати між собою групи за зростанням досліджуваної ознаки (за критерій можна взяти суму значень або середнє арифметичне).
4. Для кожного значення (крім значень з найбільшої за критерієм групи) підрахувати кількість значень з більших груп, що його перевищують (S_i).
5. Підрахувати суму усіх S_i: $A=\sum S_i$, – та максимально можливу кількість перевищуючих значень: $B = \frac{m(m-1)}{2} n^2$, де m – кількість груп, n – обсяг кожної вибірки.
6. Визначити емпіричне значення критерія: $S = 2 \cdot A - B$.
7. За спеціальною таблицею визначити критичне значення. Нульову гіпотезу слід відхилити, якщо обчислене значення неменше за критичне.

У розділі використано матеріали з посібників [8, 19, 21].

Параметричний дисперсійний аналіз для зв'язаних вибірок (повторні вимірювання)

При дослідженні зв'язаних вибірок (тобто, коли вимірювання досліджуваної ознаки проводиться в різних умовах на одні вибірці об'єктів), крім впливу досліджуваного фактора слід враховувати індивідуальні впливи окремих елементів вибірки. Відповідно лінійну модель дисперсійного аналізу буде доповнено. У разі однофакторного аналізу матимемо $x_{ij} = M_{виб} + (M_j - M_{виб}) + P_i + a_{ji} + e_{ij}$. Тут $M_{виб}$ – математичне сподівання генеральної сукупності (його оцінкою є вибіркове середнє), M_j – середнє у групі, що відповідає j-тій градації фактора, P_i – вплив індивідуальних особливостей і-того об'єкта вибірки, a_{ji} – вплив взаємодії індивідуальної та факторної компонент, e_{ij} – "похибка" лінійної моделі, пов'язана з j-тою градацією фактора та і-тим об'єктом.

Лінійна модель двофакторного аналізу у випадку зв'язаних вибірок також доповнюється компонентами індивідуальних особливостей та їхніх взаємодій з досліджуваними факторами:

$$x_{ij} = M_{виб} + P_j + P_g + P_i + a_{jg} + a_{ji} + a_{gi} + a_{jgi} + e_{ij}.$$

Тут P_j, P_g та a_{jg} – вплив досліджуваних факторів та їхньої взаємодії; a_{ji}, a_{gi}, a_{jgi} – відповідно взаємодія досліджуваних факторів з індивідуальними особливостями об'єктів.

В однофакторному випадку основними припущеннми є 1) припущення про те, що кожне значення досліджуваної ознаки, отримане для одного об'єкта, є вибіркою з багатовимірного нормального розподілу; 2) для коректності визначення F-відношення необхідним є припущення про складену симетрію (compound symmetry) коваріаційної матриці, утвореної змінними, які відповідають різним градаціям міжгрупових факторів (тобто елементи головної діагоналі коваріаційної матриці мають бути однаковими, а матриця – симетричною). Відповідно повинна виконуватись умова про гомогенність вибіркових дисперсій.

Припущення про складену симетричність коваріаційної матриці є частиною більш загального припущення про сферичність коваріаційно-дисперсійної матриці. Сферичність означає, що дисперсійно-коваріаційні матриці, отримані за міжгруповими факторами – однакові, а за залежними змінними – це одинична матриця. Лише за такої умови застосування F-критерія є надійним засобом статистичної оцінки факторних ефектів та міжфакторної взаємодії у дисперсійному аналізі з повторними вимірюваннями (зв'язаними вибірками).

Сферичність перевіряють за допомогою теста Моучлі (Mauchly). Однак для малих вибірок цей тест недостатньо потужний, а для великих – навіть малі відхилення від сферичності може прийняти за статистично значущі. У статистичних пакетах при порушенні сферичності вводиться поправка (epsilon ajustment) на кількість степенів вільності для F-критерія.

У багатовимірному випадку також вимірювання, що відповідають кожному об'єкту повинні бути вибіркою, утвореною з багатовимірного нормального розподілу (перше припущення). Другим припущенням є припущення про ідентичність дисперсійно-коваріаційних матриць, утворених міжгруповими

факторами. Це припущення перевіряється за допомогою M-теста Бокса (Box's M-test).

Завдання 1: дві вибірки

1. Застосувати непараметричні методи (U-критерій Манна-Уітні) до двох незв'язаних вибірок (див. завдання 1 теми 2, с. 20) засобами пакетів SPSS та Statistica.
2. Порівняти з результатами застосування параметричних методів. Зробити висновки.
3. Застосувати непараметричні методи (критерії знаків та Вілкоксона) до даних двох зв'язаних вибірок (див. варіанти завдання 2 з теми 2 на с. 202) засобами пакетів SPSS та Statistica.
4. Порівняти з результатами застосування параметричних методів. Зробити висновки.

Завдання 2: декілька вибірок

1. Застосувати непараметричні методи (критерії Крускала-Уолліса та Джонкіра) до задачі з теми 3 (варіанти до завдання 1 на с. 203) засобами пакетів SPSS та Statistica.
2. Порівняти з результатами однофакторного дисперсійного аналізу. Зробити висновки.
3. Підібрати дані до аналізу декількох зв'язаних вибірок.
4. Виконати їхнє порівняння засобами параметричного однофакторного аналізу.
5. Виконати їхнє порівняння засобами непараметричного однофакторного аналізу (критерія χ_r^2 Фрідмана).
6. Порівняти отримані результати. Зробити висновки.

Приклади виконання

Приклад 1: дві незв'язані вибірки.

Застосуємо непараметричний критерії до даних з прикладу 1 до теми 2 (див. с. 22).

У пакеті MS Excel непараметричні методи не реалізовано (крім критерія χ^2-Пірсона), тому доведеться виконувати обчислення за алгоритмом Алгоритм 5. Фрагмент електронної таблиці наведено на Рис. 38. Ранги обчислено вручну, суму рангів – за допомогою стандартної функції СУММ, а значення U – за формулою, наведеною в алгоритмі.

	A	B	C	D
1		n=	14	12
2	*група1*	*група2*	*Ранг1*	*Ранг2*
3	90	102	1	4,5
4	95	104	2	6,5
5	99	105	3	8
6	102	107	4,5	11,5
7	104	108	6,5	14
8	106	111	9	15,5
9	107	112	11,5	17
10	107	113	11,5	18
11	107	114	11,5	19
12	111	117	16,5	23
13	115	122	21,5	24
14	115	123	21,5	25
15	116		22	
16	127		26	
17	Сума рангів=		168	186
18	U=		**60**	

Рис. 38

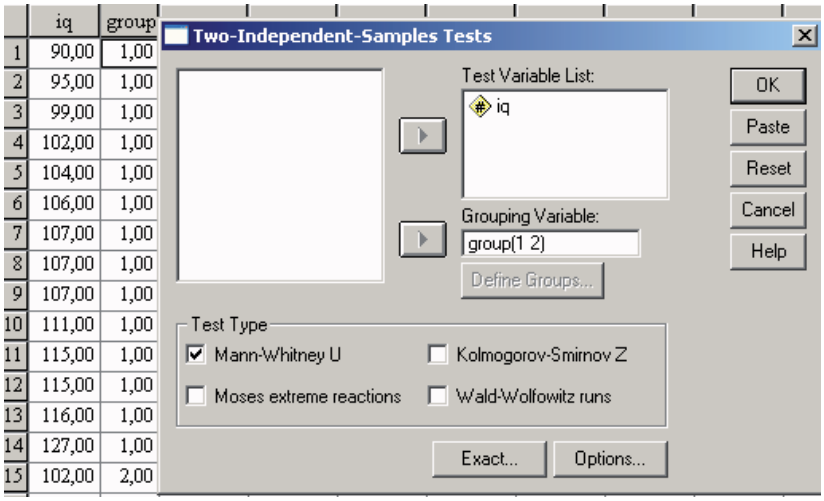


Рис. 39

У статистичних пакетах SPSS та Statistica вхідні дані до задач подібного типу слід подавати у вигляді двох змінних – групуючої (зі значеннями номерів груп, до яких належать відповідні досліджувані об’єкти) та змінної, що містить числові значення досліджуваної ознаки.

У пакеті SPSS для аналізу двох незалежних змінних створимо змінні IQ та GROUP, застосуємо процедуру Analyze –> Nonparametric Tests –> 2 Independent Samples та зробимо у діалоговому вікні настройки, як показано на Рис. 39.

В результаті отримаємо дві таблиці: Ranks – з підсумковими результатами ранжування, – та Test Statistics – з аналізом достовірності критерія.

Ranks

GROUP		IQ
1,00	N	14
	Mean Rank	11,79
	Sum of Ranks	165,00
2,00	N	12
	Mean Rank	15,50
	Sum of Ranks	186,00
Total	N	26

Test Statistics[b]

	IQ
Mann-Whitney U	60,000
Wilcoxon W	165,000
Z	-1,237
Asymp. Sig. (2-tailed)	,216
Exact Sig. [2*(1-tailed Sig.)]	,231[a]

a. Not corrected for ties.

b. Grouping Variable: GROUP

Як видно з таблиць, отримані рангові суми та значення U співпадають з обчисленими за алгоритмом. А p-значення критерія (Asymp. Sig) співпадає з табличним та у даному випадку свідчить про відсутність різниці між двома вибірками (тобто нульову гіпотезу не можна відхилити).

У пакеті Statistica необхідно виконати процедуру Statistics –> Nonparametric Statistics –> Comparing two independent samples. У діалоговому вікні Variables вказати залежну змінну (IQ) та групуючу (group). Результати порівняння двох вибірок будуть такими самими, як і у пакеті SPSS:

Mann-Whitney U Test (Spreadsheet1) By variable group Marked tests are significant at p <,05000	
Rank Sum Group1	165,0000
Rank Sum Group2	186,0000
U	60,00000
Z	-1,23443
p-level	0,217045
Z adjusted	-1,23739
p-level	0,215943
Valid N Group1	14
Valid N Group2	12
2*1sided exact p	0,231155

При обсягах вибірок більше 20 розподіл величини U наближається до нормального, тому результати тесту супроводжуються значенням z-апроксимації для розподілу статистики критерія та її двостороннім p-значенням.

Для малих вибірок (обсягом менше 20) значення імовірності похибки уточнюється (2*1 sided exact p). За p тут приймають значення *1 мінус накопичена імовірність відповідної U-статистики*. Уточнене таким чином значення похибки завжди більше асимптотичного, що знижує його потужність на малих вибірках.

Різницю між двома вибірками можна вважати достовірною, коли рівень похибки виявиться меншим за встановлений рівень значущості. В пакеті Statistica такі результати відмічають кольором.

У даному випадку висновок про статистичну гіпотезу буде такм самим, як і при застосуванні параметричного критерія Стьюдента: нульову гіпотезу не можна відхилити, тобто різниця у рівнях інтелекту недостовірна (неістотна).

Приклад 2: дві зв'язані вибірки (повторні вимірювання).

Застосуємо непараметричні критерії до даних прикладу 2 з теми 2 (див. с. 22).

	A	B	C	D	E
1		X	Y	Y-X	Ранг
2	учень 1	3400	3800	**400**	9
3	учень 2	3600	3700	**100**	3
4	учень 3	3000	3300	**300**	7,5
5	учень 4	3660	3600	**-60**	**1**
6	учень 5	2900	3100	**200**	5,5
7	учень 6	3100	3200	**100**	3
8	учень 7	3200	3200	0	
9	учень 8	3400	3300	**-100**	**3**
10	учень 9	3200	3500	**300**	7,5
11	учень 10	3400	3600	**200**	5,5
12			G =	2	
13			n =	9	
14				W =	4

Рис. 40

Виконаємо спершу обчислення значень G-критерія знаків та W-критерія Вілкоксона за алгоритмом у пакеті MS Excel. На Рис. 40 у стовпцях B та C містяться вхідні дані – результати першого (X) та другого (Y) вимірювань, у стовпці D – їхні різниці. У комірці D19 – загальна кількість ненульових зсувів (комірка містить формулу =СЧЁТЕСЛИ(D8:D17;"<>0")). У комірці D18 – обчислене значення G-критерія – кількість нетипових, у даному випадку від'ємних, зсувів, яке підраховано за допомогою формули =СЧЁТЕСЛИ(D8:D17;"<0").

У стовпці E вручну, відповідно до правила, абсолютним значенням різниць (стовбець D) нараховано ранги. У комірці E20 обчислено значення W-критерія Вілкоксона, тобто суму рангів, які відповідають нетиповим, у даному випадку від'ємним, зсувам. Значення отримано за допомогою формули =СУММЕСЛИ(D8:D17;"<0";E8:E17).

Висновок стосовно висунутих статистичних гіпотез можна отримати, порівнявши обчислені значення критеріїв з відповідними критичними для прийнятого рівня значущості.

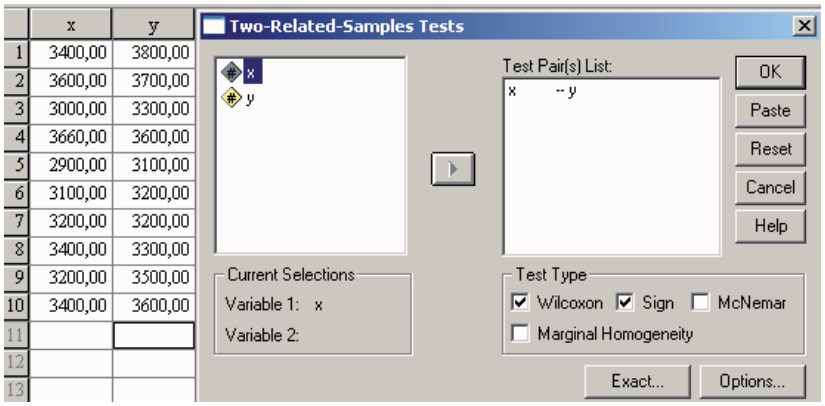


Рис. 41

У пакеті SPSS порівняння зв'язаних вибірок здійснюється процедурою Analyze –> Nonparametric Tests –> 2 Related Samples. У діалоговому вікні процедури пов'язані між собою змінні парами переносяться в розділ Test Pair(s) List; здійснюється вибір потрібних непараметричних тестів (Рис. 41).

Для G-критерія знаків (Sign Test) отримаємо (Таблиця 31) таблицю частот додатних, від'ємних та нульових зсувів і p-значення критерія, яке у даному випадку більше за прийнятий

рівень значущості (наприклад, 0,05), тому нульову гіпотезу відкинути не можна.

Таблиця 31

Frequencies

		N
Y - X	Negative Differences[a]	2
	Positive Differences[b]	7
	Ties[c]	1
	Total	10

a. Y < X
b. Y > X
c. X = Y

Test Statistics[b]

	Y - X
Exact Sig. (2-tailed)	,180[a]

a. Binomial distribution used.
b. Sign Test

Для W-критерія Вілкоксона (Wilcoxon Signed Ranks Test) отримаємо для додатних, від'ємних та нульових зсувів частоти і суми рангів. Вони такі ж, як і при обчисленні за алгоритмом. За результатими z-апроксимації асимптотичне p-значення критерію дорівнює 0,028, тобто для рівня значущості $\alpha = 0{,}05$ є підстави відхилити нульову гіпотезу.

Таблиця 32

Ranks

	Y - X		
	N	Mean Rank	Sum of Ranks
Negative Ranks	2[a]	2,00	4,00
Positive Ranks	7[b]	5,86	41,00
Ties	1[c]		
Total	10		

a. Y < X
b. Y > X
c. X = Y

Test Statistics[b]

	Y - X
Z	-2,203[a]
Asymp. Sig. (2-tailed)	,028

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

Якщо порівняти результати G-критерія знаків, W-критерія Вілкоксона та t-критерія Стьюдента, то виявиться, що згідно до двох останніх нульову гіпотезу буде відхилено, хоча при цьому $P_w > P_t$ $(0{,}028 > 0{,}022)$, – тобто з трьох критеріїв найменш потужним є критерій знаків, а найбільш потужним – критерій Стьюдента. Не слід також забувати, що застосування критерія Стьюдента до даних розглядуваного прикладу є найменш

коректним – адже не доведено, що вони відповідають нормальному закону.

У пакеті Statistica обробка двох зв'язаних вибірок здійснюється процедурою Statistics –> Nonparametric Statistics –> Comparing two dependent samples. У діалоговому вікні процедури (Рис. 42) необхідно вказати залежні змінні та по черзі переглянути результати критерію знаків (Sign test –

Таблиця 33), критерію Вілкоксона (Wilcoxon matched pairs test –

Таблиця 34) та блочні діаграми для порівнюваних змінних. Як видно з таблиць, результати такі самі, як і отримані в SPSS.

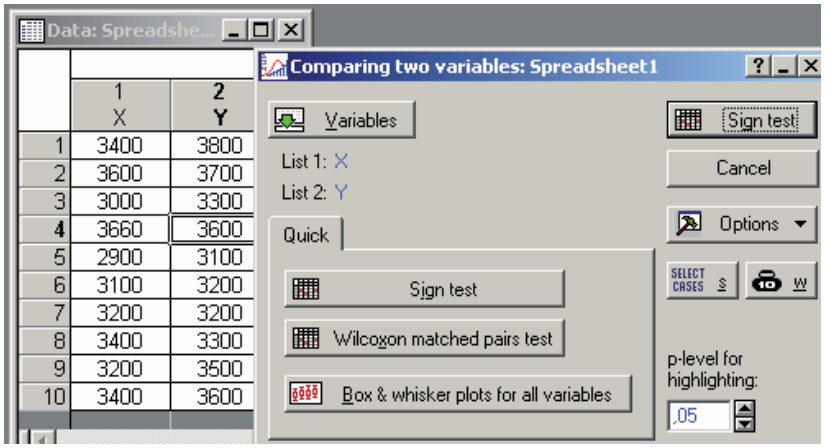


Рис. 42

Таблиця 33

Pair of Variables	Sign Test (Spreadsheet1) Marked tests are significant at p <,05000			
	No. of Non-ties	Percent v < V	Z	p-level
X & Y	9	77,77778	1,333333	0,182422

Таблиця 34

Pair of Variables	Wilcoxon Matched Pairs Test (Spreadsheet1) Marked tests are significant at p <,05000			
	Valid N	T	Z	p-level
X & Y	10	4,000000	2,191691	0,028403

Приклад 3: декілька незв'язаних вибірок

Застосуємо непараметричні методи до даних прикладу 1 з теми 3 (с. 36).

Як і при порівнянні двох незалежних вибірок, дані слід представити у вигляді двох змінних – групуючої та для значень досліджуваної ознаки. Тут назвемо групуючу змінну ДОСВІД, її значеннями будуть 1, 2, 3 – номери груп досліджуваних. Змінна ПОМИЛКИ міститиме дані про середню кількість допущених помилок.

У пакеті SPSS для порівняння декількох незалежних змінних застосовують процедуру Analyze –> Nonparametric Tests –> K Independent Samples. На Рис. 43 представлено діалогове вікно процедури та фрагмент даних.

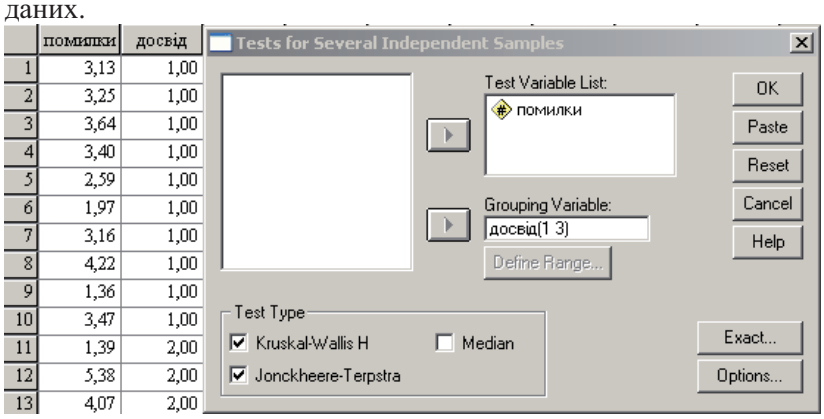


Рис. 43

За тестом Крускала-Уоліса буде отримано середні значення рангів для кожної умови, значення статистики критерія (9,809) та асимптотичний рівень значущості (0,007), відповідно до якого відмінності між групами слід визнати достовірними, тобто відхилити нульову гіпотезу.

Ranks

	ДОСВІД	N	Mean Rank
ПОМИЛКИ	1,00	10	9,10
	2,00	10	16,00
	3,00	10	21,40
	Total	30	

Test Statistics[a,b]

	ПОМИЛКИ
Chi-Square	9,809
df	2
Asymp. Sig.	,007

a. Kruskal Wallis Test

b. Grouping Variable: ДОСВІД

Для обчислення критерія Джонкіра групи слід впорядкувати за зростанням досліджуваної ознаки. На щастя, у наведеному прикладі групи впорядковано.

Таблиця 35 містить результати. Отже за критерієм Джонкіра, так само, як і за результатими дисперсійного аналізу та критерія Крускала-Уоліса, нульову гіпотезу слід відхилити, але з уточненням, що кількість помилок з зменшується досвідом.

Якщо порівняти два непараметричні критерії, то останній виявиться більш потужним ($P_j<P_H$), але його можна застосувати лише тоді, коли є сенс досліджувати "динаміку" змін.

Таблиця 35

Jonckheere-Terpstra Test[a]

	ПОМИЛКИ
Number of Levels in ДОСВІД	3
N	30
Observed J-T Statistic	239,000
Mean J-T Statistic	150,000
Std. Deviation of J-T Statistic	26,300
Std. J-T Statistic	3,384
Asymp. Sig. (2-tailed)	,001

a. Grouping Variable: ДОСВІД

У пакеті Statistica для порівняння незв'язаних вибірок слід виконати процедуру Statistics –> Nonparametric Statistics –>

Comparing multiple indep. samples (groups). У наступному діалоговому вікні слід вказати залежну та групуючу змінні і отримати результат (Summary: Kruskal-Wallis ANOVA & Median Test).

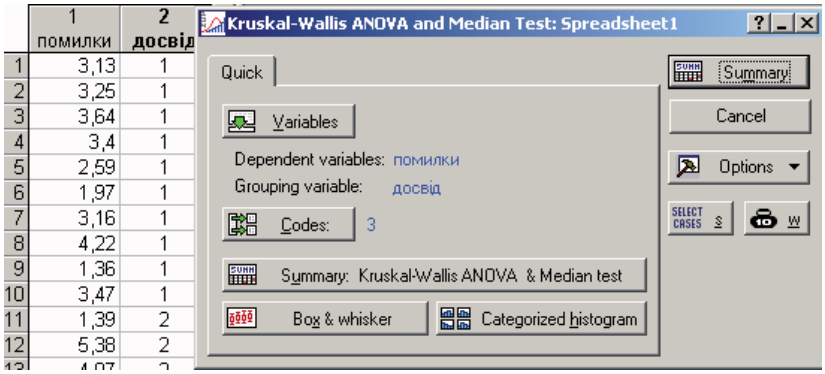


Рис. 44

На першій з таблиць будуть подані результати аналізу за критерієм Крускала-Уоліса (Таблиця 36). Вони співпадають з отриманими в SPSS.

Таблиця 36

	Kruskal-Wallis ANOVA by Ranks; помилки (Spreadsheet1) Independent (grouping) variable: досвід Kruskal-Wallis test: H (2, N= 30) =9,809032 p =,0074		
Depend.: помилки	Code	Valid N	Sum of Ranks
Grp.1	1	10	91,0000
Grp.2	2	10	160,0000
Grp.3	3	10	214,0000

Крім того можна побудувати блочні діаграми для групових середніх (Box & whisker) – Рис. 45.

Як і в пакеті SPSS, порівняння декількох незв'язаних вибірок можна виконати за допомогою медіанного критерія. Він менш потужний порівняно з критерієм крускала-Уоліса і тут не розглядається. Критерій Джонкіра в пакеті Statistica не обчислюють.

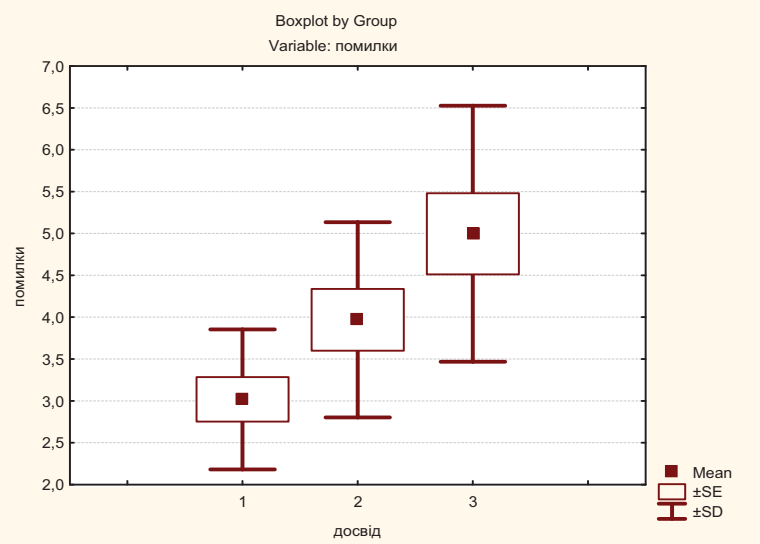


Рис. 45

Приклад 4а: декілька зв'язаних вибірок (один фактор).

На групі з 9 пацієнтів, які страждають мігренню, вивчали вплив тренінга релаксації на частоту головного болю. Досліджуваних просили реєструвати кількість годин мігренеподібних нападів протягом 5 тижнів: 2 тижні до тренінга та 3 тижні під час тренінга. З'ясувати, чи впливає тренінг на частоту головного болю? Чи однакова частота головного болю в усіх досліджуваних? Чи залежить ефективність тренінгу від індивідуальних особливостей пацієнтів?

Вхідні дані подано на таблиці Таблиця 37.

Таблиця 37

	A	B	C	D	E	F
1		Фон		Тренінг		
2		тиждень1	тиждень2	тиждень3	тиждень4	тиждень5
3	пацієнт1	21	22	8	6	6
4	пацієнт2	20	19	10	4	4
5	пацієнт3	17	15	5	4	5
6	пацієнт4	25	30	13	12	17
7	пацієнт5	30	27	13	8	6
8	пацієнт6	19	27	8	7	4
9	пацієнт7	26	16	5	2	5
10	пацієнт8	17	18	8	1	5
11	пацієнт9	26	24	14	8	9

Відповідні статистичні гіпотези будуть такими:

H_0: між показниками, отриманими в різних умовах, існують лише випадкові відмінності.

H1: між показниками, отриманими в різних умовах, існують не випадкові відмінності.

Застосування непараметричних критеріїв у даному випадку є абсолютно доречним і доцільним.

На Рис. 46 показано результати обчислення критерію χ_r^2 Фрідмана за алгоритмом в MS Excel. У комірці H14 знаходиться значення критерію, обчислене за формулою (Алгоритм 3), та у комірці H15 – його p-значення, обчислене за допомогою функції =ХИ2РАСП(H14;4). За отриманим результатом також можна зробити висновок, що тренінг впливає на частоту головних болей.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Фон		Тренінг				Ранги					
2		тиждень 1	тиждень 2	тиждень 3	тиждень 4	тиждень 5		тиждень 1	тиждень 2	тиждень 3	тиждень 4	тиждень 5	
3	пацієнт1	21	22	8	6	6		5	4	3	1,5	1,5	
4	пацієнт2	20	19	10	4	4		5	4	3	1,5	1,5	
5	пацієнт3	17	15	5	4	5		5	4	2,5	1	2,5	
6	пацієнт4	25	30	13	12	17		4	5	2	1	3	
7	пацієнт5	30	27	13	8	6		5	4	3	2	1	
8	пацієнт6	19	27	8	7	4		4	5	3	2	1	
9	пацієнт7	26	16	5	2	5		5	4	2,5	1	2,5	
10	пацієнт8	17	18	8	1	5		4	5	3	1	2	
11	пацієнт9	26	24	14	8	9		5	4	3	1	2	
12						Сума R		42	39	25	12	17	
13						Сума R)^2		1764	1521	625	144	289	4343
14							Xi2r=	31,02					
15							p=	3E-06					

Рис. 46

Такі ж самі результати будуть отримані при застосуванні процедури Analyze –> Nonparametric Tests –> K Related Samples в пакеті SPSS та процедури Statistics –> Nonparametrics –> Comparing multiple dep. samples (variables). В обох пакетах одночасно з тестом Фрідмана виконується обчислення W – коефіцієнта конкордації (узгодженості) Кендала, який приймає значення від 0 до 1. Значення W близьке до 1 (у даному прикладі W = 0,876) свідчить про значну узгодженість між змінними. Таблиця 38 представляє результати, отримані у пакеті SPSS.

Таблиця 38

Friedman Test

Test Statistics[a]

N	9
Chi-Square	31,545
df	4
Asymp. Sig.	,000

a. Friedman Test

Kendall's W Test

Test Statistics

N	9
Kendall's W[a]	,876
Chi-Square	31,545
df	4
Asymp. Sig.	,000

a. Kendall's Coefficient of Concordance

Таблиця 39 представляє результати, отримані у пакеті Statistica, а на Рис. 47 зображено основне діалогове вікно процедури.

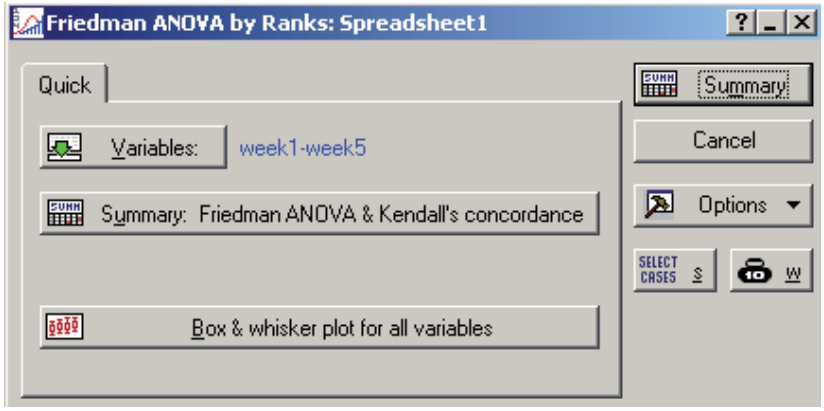


Рис. 47

Таблиця 39

Variable	Friedman ANOVA and Kendall Coeff. of Concordance (Spreadsheet1 ANOVA Chi Sqr. (N = 9, df = 4) = 31,54545 p < ,00000 Coeff. of Concordance = ,87626 Aver. rank r = ,86080			
	Average Rank	Sum of Ranks	Mean	Std.Dev.
week1	4,555556	41,00000	22,33333	4,582576
week2	4,444444	40,00000	22,00000	5,338539
week3	2,777778	25,00000	9,33333	3,391165
week4	1,333333	12,00000	5,77778	3,419714
week5	1,888889	17,00000	6,77778	4,116363

Приклад 4б: застосуємо до даних прикладу 4а параметричний метод – однофакторний дисперсійний аналіз для повторних вимірювань.

Для даного методу буде утворено дві пари статистичних гіпотез:

$H_{0(A)}$: вплив тренінгу (фактору ТИЖДЕНЬ) на частоту головних болей випадковий.

$H_{1(A)}$: вплив тренінгу на частоту головних болей не випадковий.

$H_{0(I)}$: вплив індивідуальних відмінностей (фактору ПАЦІЄНТ) на частоту головних болей випадковий (виражений не більше ніж відмінності, обумовлені випадковими причинами).

$H_{1(I)}$: вплив індивідуальних відмінностей на частоту головних болей не випадковий.

У пакеті MS Excel відповідна процедура відсутня, однак правильний результат можна отримати, використовуючи результати наявних процедур та враховуючи модель дисперсійного аналізу (с. 120).

Для цього необхідно виконати дисперсійний аналіз вважаючи основним впливаючим фактором спершу фактор ТИЖДЕНЬ, а потім виконати аналіз з фактором ІНДИВІДУАЛЬНІ ВІДМІННОСТІ.

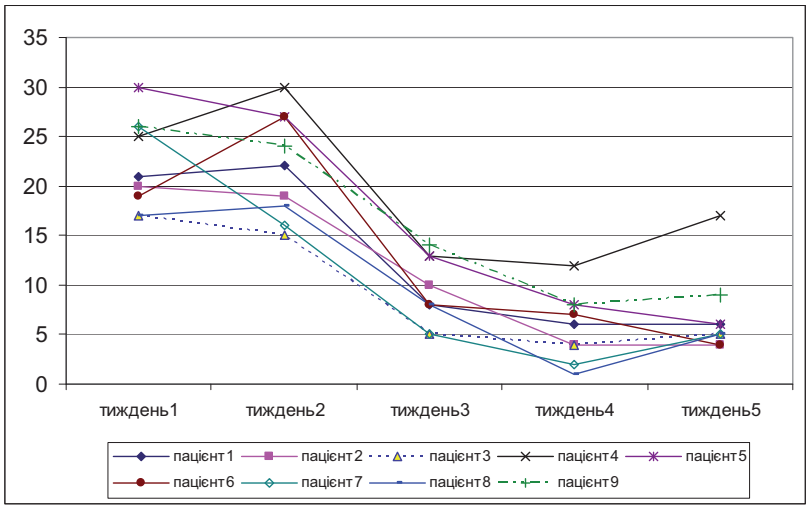


Рис. 48

Результати однофакторного дисперсійного аналізу, отримані в пакеті MS Excel для фактора ТИЖДЕНЬ, тобто до

стовпців таблиці (в пакеті Анализ Данных виділити діапазон B2:F11 по стовпцях, мітки у першому рядку), свідчать про те, що частота головного болю за час проведення дослідження змінилася. Оскільки основним впливаючим фактором у цей період був тренінг, то є підстави стверджувати, що вплив тренінга не випадковий (Таблиця 40). Однак вплив індивідуальних особливостей пацієнтів тут не враховано, хоча, як видно на Рис. 48, їхні результати помітно відрізняються.

Таблиця 40

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
тиждень1	9	201	22,33333	21
тиждень2	9	198	22	28,5
тиждень3	9	84	9,333333	11,5
тиждень4	9	52	5,777778	11,69444
тиждень5	9	61	6,777778	16,94444

Дисперсионный анализ

Источник в	*SS*	*df*	*MS*	*F*	*P-Значени*	*F критическое*
Между группами	2449,2	4	612,3	34,1537	2,08E-12	2,605975
Внутри групп	717,1111	40	17,92778			
Итого	3166,311	44				

З’ясуємо для цих же даних вплив фактора індивідуальних відмінностей (без урахування тренінгу), тобто виконаємо однофакторний аналіз для рядків таблиці (в пакеті Анализ Данных виділити діапазон A3:F11 по рядках, мітки у першому стовпці). У таблиці Таблиця 41 наведено результати, згідно до яких вплив фактора індивідуальних відмінностей слід визнати недостовірним.

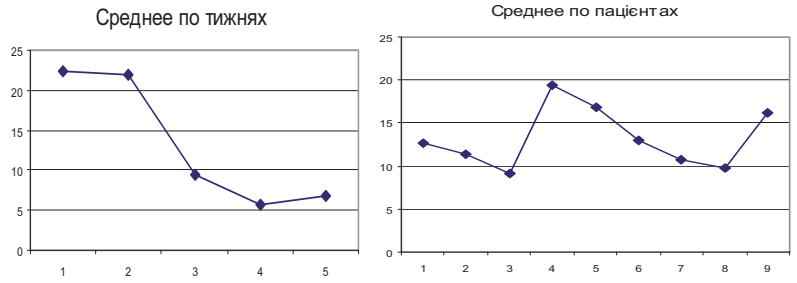


Рис. 49

На Рис. 49 представлено графіки, побудовані за середніми. На них висновки, зроблені за таблицями Таблиця 40 та Таблиця 41, підтверджуються графічно.

Таблиця 41

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
пацієнт1	5	63	12,6	66,8
пацієнт2	5	57	11,4	60,8
пацієнт3	5	46	9,2	39,2
пацієнт4	5	97	19,4	61,3
пацієнт5	5	84	16,8	121,7
пацієнт6	5	65	13	93,5
пацієнт7	5	54	10,8	100,7
пацієнт8	5	49	9,8	55,7
пацієнт9	5	81	16,2	70,2

Дисперсионный анализ

Источник	*SS*	*df*	*MS*	*F*	*P-Значени*	*F критическое*
Между группами	486,7111	8	60,83889	0,817361	0,592361	2,208518
Внутри групп	2679,6	36	74,43333			
Итого	3166,311	44				

Виконаємо перерахунки з урахуванням лінійної моделі для повторних вимірювань (с. 120). Для цього використаємо отримані у попередніх обчисленнях загальну суму квадратів $Q_{total} = 3166{,}31$, суму квадратів по фактору ТИЖДЕНЬ $Q_A = 2449{,}2$, суму квадратів по фактору індивідуальних відмінностей $Q_I = 486{,}71$. Обчислимо суму квадратів, обумовлену випадковими причинами: $Q_Z = Q_{total} - Q_A - Q_I$.

Заповнимо таблицю дисперсійного аналізу:

Джерело варіації	Суми квадратів	степені вільності	Дисперсії	F
Досліджуваний фактор (міжгруповий)	Q_A	m–1	$S_A^2 = \frac{Q_A}{m-1}$	$F_A = \frac{S_A^2}{S_z^2}$
Індивідуальні відмінності	Q_I	n–1	$S_I^2 = \frac{Q_I}{n-1}$	$F_I = \frac{S_I^2}{S_z^2}$
Залишкова	Q_z	(m-1)(n-1)	$S_z^2 = \frac{Q_z}{(m-1)(n-1)}$	
Загальна	Q_{total}	mn–1		

Для даного прикладу отримаємо:

	Q	df	S^2	F	P	$F_{0,05}$
ТИЖДЕНЬ	2449,20	4	612,30	85,042	1,39E-16	2,668
ПАЦІЄНТ	486,71	8	60,84	8,450	7,13E-07	2,244
Випадкова	230,40	32	7,20			
Загальна	3166,31	44	71,96			

Згідно наведеної таблиці слід вважати достовірними як вплив фактора ТИЖДЕНЬ (результати тренінгу), так і індивідуальні відмінності між пацієнтами. Причому урахування індивідуальних особливостей дає можливість краще оцінити факторний ефект.

У пакеті SPSS однофакторний дисперсійний аналіз з повторними вимірюваннями проводиться за допомогою процедури Analyze –> General Linear Model –> Repeated Measures. У діалоговому вікні, що відкриється (Рис. 50), слід вказати назву внутрігрупового фактора (Within-Subject Factor Name) та кількість його градацій (ця назва ніяким чином не пов'язана з назвами змінних, у даному прикладі доречно використати назву ТИЖДЕНЬ – week). Натиснути Add.

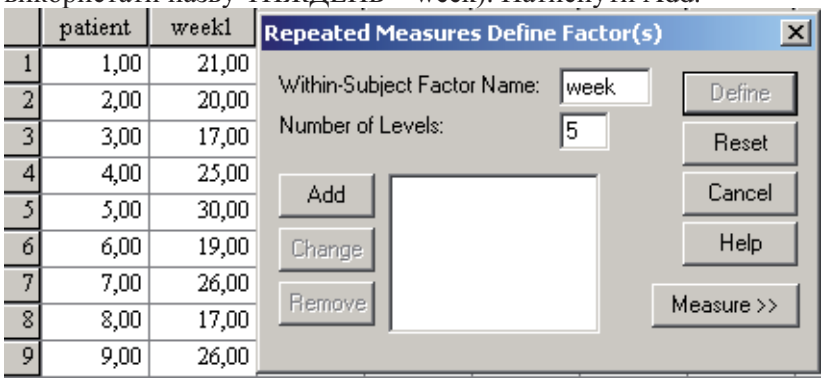


Рис. 50

У наступному діалоговому вікні (Repeated Measures), яке відкриється кнопкою Define слід вказати, яка змінна якому рівню фактора відповідатиме. У даному випадку рівень 1 це week1 і так далі (фрагмент вікна представлено на Рис. 51).

Додатково у цьому ж вікні можна вибрати модель обробки даних (Model). За замовченням виконується Full Factorial. Добавити у звіт графіки (Plots) та деякі додаткові показники

(Options), наприклад, силу факторного впливу (Estimates of effect size).

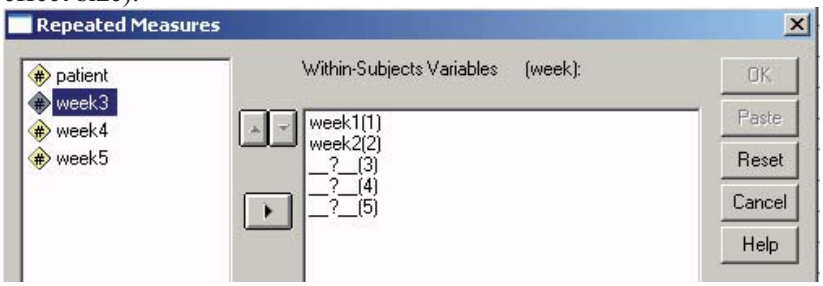


Рис. 51

Основні результати однофакторного аналізу будуть представлені у таблиці Tests of Within Subjects Effects) Таблиця 42). В цій таблиці подано суми квадратів Q_A та Q_z (у стовпці Sum of Squares), відповідні дисперсії та значення F_A.

Таблиця 42

Tests of Within-Subjects Effects

		Measure					
		MEASURE_1					
Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
WEEK	Sphericity Assumed	2449,200	4	612,300	85,0	,000	,914
	Greenhouse-Geisser	2449,200	2,738	894,577	85,0	,000	,914
	Huynh-Feldt	2449,200	4,000	612,300	85,0	,000	,914
	Lower-bound	2449,200	1,000	2449,200	85,0	,000	,914
Error (WEEK)	Sphericity Assumed	230,400	32	7,200			
	Greenhouse-Geisser	230,400	21,903	10,519			
	Huynh-Feldt	230,400	32,000	7,200			
	Lower-bound	230,400	8,000	28,800			

Таблиця 43

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	7893,689	1	7893,689	129,7	,000	,942
Error	486,711	8	60,839			

У таблиці Tests of Between-Subjects Effects (

Таблиця 43) знайдемо суму квадратів та дисперсію, що відповідає фактору індивідуальних відмінностей (Q_I та S^2_I).

Одновимірний підхід (Univariate approach) базується на припущенні про сферичність коваріаційно-дисперсійної матриці, Сферичність означає однакові дисперсії для різних рівнів внутрігрупового фактора та додатні кореляції між повторніми вимірюваннями. Перевірка сферичності здійснюється тестом Моучлі (Mauchly's Test of Sphericity). Якщо результати теста статистично значимі, то одновимірний підхід не можна застосувати, тому при порушеннях сферичності пропонується поправка (Epsilon Corrected) степенів вільності та рівня значущості.

У даному прикладі результати теста Моучлі не значимі (Sig. = 0,537), отже застосування одновимірного підходу цілком коректне (Таблиця 44).

Таблиця 44

Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
WEEK	,282	8,114	9	,537	,684	1,000	,250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.
Design: Intercept
Within Subjects Design: WEEK

Іншими результатами однофакторного аналізу є дослідження сили та виду нелінійного зв'язку (кореляції) між факторами ТИЖДЕНЬ та індивіуальних відмінностей. Найбільш достовірною ц наведеному прикладі виявляється лінійна модель (Linear), саме вона забезпечує найбільшу силу зв'язку (η^2=0,96).

Таблиця 45

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	WEEK	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
WEEK	Linear	2016,400	1	2016,4	190,2	,000	,960
	Quadratic	89,175	1	89,175	12,011	,008	,600
	Cubic	256,711	1	256,711	53,918	,000	,871
	Order 4	86,914	1	86,914	14,451	,005	,644
Error (WEEK)	Linear	84,800	8	10,600			
	Quadratic	59,397	8	7,425			
	Cubic	38,089	8	4,761			
	Order 4	48,114	8	6,014			

Щоб отримати графіки, побудовані за індивідуальними значеннями можна скористатися процедурою Graph –> Line, вибрати тип графіка Multiple та Values of Individual Cases. Далі, як показано на Рис. 52, необхідно визначити, які змінні представлятимуть графіки та підписи для горизонтальної вісі. Результат побудови представлено на Рис. 53.

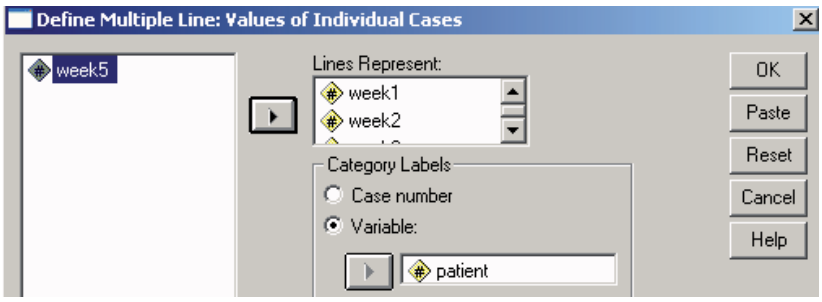


Рис. 52

Щоб отримати графік, подібний до Рис. 48, вхідні дані слід транспонувати за допомогою процедури Data –> Transpose.

У пакеті Statistica для малих дисперсійних комплексів (до п’яти градацій змінних та малої кількості досліджуваних об’єктів) можна застосувати процедуру Statistics –> ANOVA –> Repeated measures ANOVA.

Більш універсальною є процедура Statistics –> General Linear/ Nonlinear Models –> General Linear Models –> Repeated measures ANOVA.

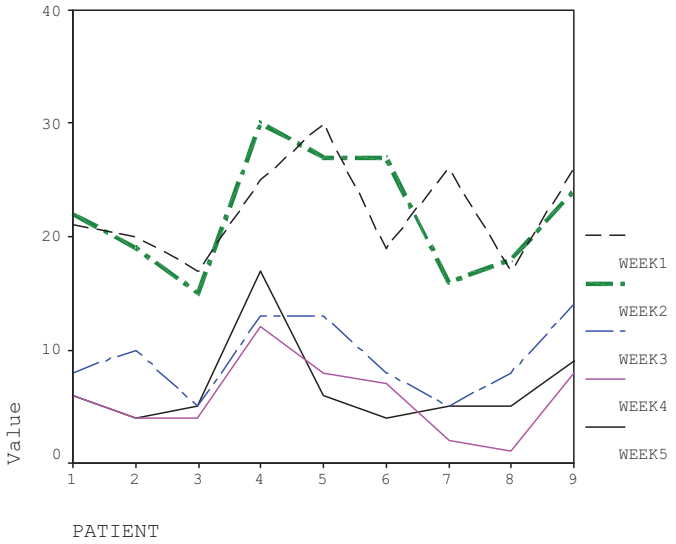


Рис. 53

Для однофакторної моделі необхідно вказати залежні змінні (у даному прикладі week1-week5 – що містять дані дослідження) та вказати назву і кількість градацій досліджуваного фактора (так само, як і в SPSS, назва може бути довільною). У даному прикладі фактору дано назву WEEK (Within effects) (Рис. 54).

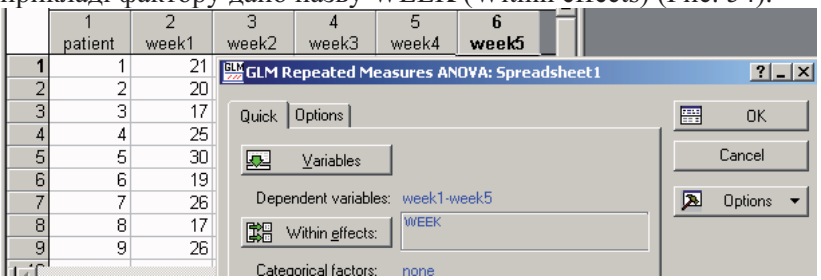
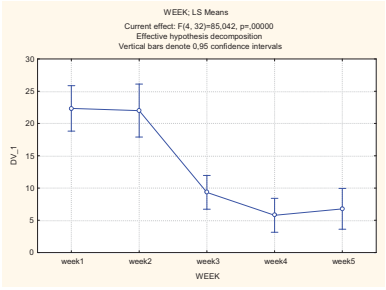


Рис. 54

На закладці Summary діалогового вікна GLM results (Рис. 55) отримаємо наступні звіти:

1. All effects/ Graphs – графік середніх, що відповідають різним градаціям фактора:



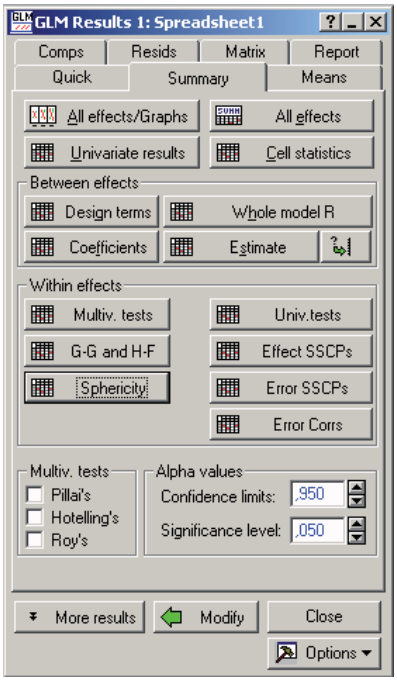


Рис. 55

2. Sphericity – результати теста сферичності:

Effect	Mauchley Sphericity Test (Spreadsheet1) Sigma-restricted parameterization Effective hypothesis decomposition			
	W	Chi-Sqr.	df	p
WEEK	0,282355	8,114462	9	0,52265

3. All effects – основна таблиця дисперсійного аналізу:

Effect	Repeated Measures Analysis of Variance (Spreadsheet1) Sigma-restricted parameterization Effective hypothesis decomposition				
	SS	Degr. of Freedom	MS	F	p
Intercept	7893,689	1	7893,689	129,7474	0,000003
Error	486,711	8	60,839		
WEEK	2449,200	4	612,300	85,0417	0,000000
Error	230,400	32	7,200		

Контрольні запитання

1. Які критерії називають непараметричними? Чому?
2. Назвіть переваги та недоліки параметричних та непараметричних критеріїв.
3. Які обмеження параметричних методів?
4. Наведіть приклади непараметричних та параметричних критеріїв.

5. Що таке потужність критерія. Які критерії (параметричні чи непараметричні) є більш потужними?
6. Які критерії застосовують для порівняння зв'язаних вибірок? Наведіть приклади задач порівняння зв'язаних вибірок.
7. Які критерії застосовують для порівняння незв'язаних вибірок? Наведіть приклади задач порівняння незв'язаних вибірок.
8. Наведіть класифікацію задач статистичного аналізу даних.
9. Які параметричні методи застосовують для порівняння зв'язаних вибірок (повторних вимірювань)?
10. Які статистичні гіпотези при цьому перевіряють?
11. Які складові лінійної моделі дисперсійного аналізу у разі повторних вимірювань?

Тема 7:
“Дискримінантний аналіз”

Мета:

Студенти ***повинні знати***:
– призначення дискримінантного аналізу;
– умови та обмеження застосування дискримінантного аналізу;
– представлення даних для дискримінантного аналізу;
– геометричну інтерпретацію правила класифікації;
– призначення канонічних дискримінантних функцій, порядок їх визначення;
– критерії оцінювання якості класифікації;
– критерії відсіювання дискримінантних змінних при покроковому аналізі;
– основні результати дискримінантного аналізу.

Студенти ***повинні уміти***:
– подавати експериментальні дані у вигляді, зручному для обчислень;
– виконувати перерахування класифікуючої змінної у номінативну шкалу (при потребі);
– виконувати дискримінантний аналіз засобами пакетів SPSS та STATISTICA;
– будувати необхідні графіки;
– оцінювати якість отриманого прогнозу класової приналежності об’єктів;
– аналізувати та інтерпретувати отримані канонічні функції.

Теоретичні відомості.

Класифікація багатовимірних методів

Багатовимірні методи, одним з яких є дискримінантний аналіз (далі ДкА), представляють собою засновані на аналізі багатьох змінних емпіричні математичні моделі, прийнятні для інтерпретації у термінах досліджуваної предметної галузі. Однак теорію багатовимірних статистичних даних ще до кінця не розроблено, тому деякі з використовуваних багатовимірних методів не мають чіткої статистичної трактовки у плані концепції перевірки гіпотез, побудови довірчих інтервалів тощо [8].

Багатовимірні методи можна класифікувати за трьома критеріями: призначенням, способом співставлення даних та виглядом початкових емпіричних даних.

Класифікація за призначенням:

1. Методи прогнозування (ДкА та множинний регресійний аналіз) – дозволяють за множиною "відомих" змінних визначити значення "невідомої" змінної. У випадку ДкА "невідома" змінна номінативна, а у випадку множинного регресійного аналізу – метрична.
2. Методи класифікації (варіанти кластерного аналізу та ДкА).
3. Структурні методи (факторний аналіз та багатовимірне шкалювання).

Класифікація за методом співставлення (початковим припущенням про структуру) даних:

1. Методи засновані на кореляційній моделі даних (припущенні, про узгоджену мінливість ознак у множини об'єктів) –факторний, множинний регресійний аналіз, частково ДкА.
2. Методи, побудовані на дистантній моделі (відмінності між об'єктами розглядаються як відстані між ними) – кластерний аналіз, багатовимірне шкалювання, частково ДкА.

Класифікація за виглядом вхідних даних:

1. Методи, що використовують тільки виміряні ознаки (множинний регресійний аналіз, факторний, ДкА).
2. Методи, вхідними даними для яких можуть бути попарні відмінності (схожість) об'єктів, оцінені групою експертів (кластерний аналіз, багатовимірне шкалювання).

Призначення та основні припущення ДкА

Задачі ДкА полягають у тому, щоб визначити правила, які б дозволили за значенням дискримінантних (метричних) змінних віднести кожен з досліджуваних об'єктів (в тому числі об'єктів, класова приналежність яких невідома) до одного з відомих класів, та визначити "вагу" кожної дискримінантної змінної у такому поділі на класи. Тобто ДкА – це група статистичних процедур, які відносять до методів "класифікації з навчанням" або "розпізнавання образів". В результаті виконання ДкА можна отримати корисні відомості про окремі об'єкти, про відмінності між класами та про здатність змінних як таких точно розрізняти класи.

Початкові припущення ДкА

Позначимо через g кількість класів, p – кількість дискримінантних змінних, n_i – кількість об'єктів у класі i, N – загальну кількість об'єктів, тоді у моделі ДкА має бути:

– кількість класів не менше двох: $g \geq 2$;
– принаймні два об'єкти у кожному класі: $n_i \geq 2$;
– довільна кількість дискримінантних змінних, за умови, що їх не більше ніж кількість об'єктів мінус 2: $0 < p < (N–2)$;
– вимірювання дискримінантних змінних за шкалою не нижче інтервальної;
– лінійна незалежність дискримінантних змінних;
– приблизна рівність між коваріаційними матрицями для кожного класу;
– багатовимірна нормальність закону розподілу дискримінантних змінних для кожного класу.

Геометрична інтерпретація методу

Геометрична інтерпретація методу полягає у тому, що кожен об'єкт дослідження можна представити у p-вимірномі просторі дискримінантних змінних точкою. Точки, координатами яких є середні значення дискримінантних змінних для класу, називають "*центроїдами*" класів. Об'єкт приписують до того класу, до центроїда якого він найближчий.

Для того, щоб розрізняти відносне положення центроїдів достатньо розмірності, на одиницю меншої за кількість класів. Отже одним із завдань ДкА є перехід від h-вимірного простору дискримінантних змінних до (g–1)-вимірного простору канонічних дискримінантних функцій. Початком координат нового простору є головний центроїд – точка, координатами якої є середні значення усіх дискримінантних змінних. Першу канонічну вісь орієнтують у напрямку, в якому усі центроїди класів розрізняються максимально, другу – перпендикулярно до першої також у напрямку максимальної відмінності класів (центроїдів) і т.д.

Кожна канонічна функція є лінійною комбінацією дискримінантних змінних і має вигляд:

$f_{km}=u_0+u_1X_{1km}+u_2X_{2km}+\ldots+u_pX_{pkm}$, де f_{km} – значення канонічної дискримінантної функції для m-ного об'єкта у групі k; X_{ikm} – значення дискримінантної змінної X_i для m-ного об'єкта у групі k; u_i – канонічні коефіцієнти.

Відстані між об'єктами та центроїдами класів визначають за мірою Махаланобіса, для якої необхідне виконання умови рівності для усіх класів коваріаційних матриць:

$$D^2(X \mid G_k) = (N-g)\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}(X_i - X_{ik})(X_j - X_{jk}),$$ де

$D^2(X|G_k)$ – квадрат відстані від даного об'єкта X до центроїда класу k.

Критерії оцінки якості класифікації

Якість класифікації визначається за власним значенням канонічної функції, Λ-статистикою Уілкса, χ^2-тестом та імовірністю приналежності об'єкта до класу.

Власне значення канонічної функції є показником її інформативності: частка від ділення власного значення функції на кількість класів показує долю сумарної дисперсії усіх об'єктів по всіх змінних, яка вичерпується цією канонічною функцією. За власним значенням також визначають дискримінуючу здатність канонічної функції, додаючи усі власні значення та порівнюючи долю кожної канонічної функції у загальній сумі власних значень. Однак правила відбору функції за цим показником немає.

Λ-статистика Уілкса – це міра відмінностей між класами за декількома змінними, яку обчислюють так: $\Lambda = \prod_{i=k+1}^{g} \frac{1}{1+\lambda_i}$, де k – число вже обчислених канонічних функцій, П – символ добутку, λ_i – власне значення і-тої канонічної функції, g – кількість класів. Оскільки це "обернена" міра, то чим вона менша (ближча до нуля), тим краще розрізняються класи. При збільшенні Λ до максимального значення 1, центроїди класів співпадають.

На основі Λ-статистики Уілкса отримують тест значимості, апроксимуючи розподіл деякої функції від неї розподілом χ^2 із степенями вільності (p–k)(g–k–1): $\chi^2 = -\left[N - \left(\frac{p+g}{2}\right) - 1\right] ln\,\Lambda_k$.

Якщо отримана похибка менша за прийнятий рівень значущості (0,05 або 0,01), то центроїди класів розрізняються достовірно; якщо більша, то відмінності вважають не значущими, тобто таку канонічну функцію можна не визначати.

Імовірність приналежності об'єкта до класу обчислюється як: $Pr(X|G_k)$ – імовірність того, що об'єкт віддалений від центроїда на деяку відстань, належить цьому класу. Її обчислюють як відношення кількості "відомих" об'єктів класу до кількості усіх "відомих" об'єктів і називають апріорною імовірністю. Тоді для будь-якого об'єкта сума апріорних імовірностей по всіх класах може відрізнятися від 1, адже в ситуації, коли класи погано розрізняються, один об'єкт може бути близьким до декількох центроїдів. (Апріорна імовірність оцінює долю об'єктів класу, які знаходяться від центроїда далі ніж X).

Припускаючи, що кожен об'єкт може належати лише одному класу, обчислюють апостеріорну імовірність $Pr(G_k \mid X) = \frac{Pr(X \mid G_k)}{\sum_{i=1}^{g} Pr(X \mid G_i)}$. Сума цих імовірностей по всіх класах дорівнює 1. Апостеріорна імовірність оцінює імовірність, з якою об'єкт належить даному класу.

Точність прогнозу

Нарівні з вищевказаними критеріями найбільш прийнятною для інтерпретування мірою оцінки дискримінантної інформації є точність прогнозу (ТП). Її обчислюють як відношення сумарної кількості правильно спрогнозованих об'єктів до загальної кількості "відомих" об'єктів. Однак про її величину можна судити лише порівняно з відсотком випадкової класифікації. Наприклад, для двох класів при випадковій класифікації можна отримати 50% правильно спрогнозованих значень, для чотирьох класів – 25%. Тому ТП=60% для двох класів результат недостатній (прогноз малоефективний), а для чотирьох – досить хороший.

Стандартизованою мірою ефективності для довільної кількості класів буде τ-статистика похибок: $\tau = \frac{n_c - \sum_{i=1}^{g} p_i n_i}{N - \sum_{i=1}^{g} p_i n_i}$, де n_c – кількість правильно класифікованих об'єктів, p_i – апріорна імовірність приналежності до класу. Вираз $\sum_{i=1}^{g} p_i n_i$ обчислює кількість об'єктів, класову приналежність яких буде правильно спрогнозовано при випадковій класифікації пропорційно апріорним імовірностям. У випадку безпомилкового прогнозу τ досягає значення 1. Значення $\tau \leq 0$ – свідчить про погане розрізнення класів або виродження. Наприклад, значення τ=0,93 слід інтерпретувати так, що класифікація за допомогою дискримінантних функцій робить на 93% помилок менше, ніж могло бути при випадковій класифікації (тобто приблизно 1 помилка замість 14).

Зазвичай ТП та τ-статистика переоцінюють ефективність класифікаційної процедури, оскільки обґрунтування прийнятих висновків здійснюється на тій же вибірці, на якій будувалися класифікаційні функції. Тому для великих вибірок застосовують розбиття на дві підмножини, одна з яких використовується для отримання функцій, а друга – для перевірки класифікації. Виділені підмножини мають різні вибіркові похибки, тому тестова підмножина дасть кращу оцінку прогнозуючої здатності властивостей генеральної сукупності. При поділі слід приділяти увагу тому, щоб підмножина, на якій будуть створюватися функції, була досить великою для забезпечення стабільності коефіцієнтів.

Критерії відбору дискримінантних функцій

При покроковому виконанні ДкА в першу чергу виключаються змінні з низькою *толерантністю*. Толерантність обчислюється як одиниця зменшена на квадрат множинної кореляції даної змінної з усіма іншими. Вона є показником лінійної залежності змінної. Значення 0 свідчить про те, що дана змінна є лінійною комбінацією однієї або декількох змінних.

Статистика F-вилучення оцінює значимість погіршення розрізнення (дискримінації) після вилучення змінної зі списку дискримінантних. Вона також використовується для ранжування внеску змінних у дискримінацію.

Статистика F-включення оцінює покращення дискримінації при внесенні змінної до списку раніше відібраних дискримінантних змінних.

При малих значення статистик F-вилучення та F-включення змінну можна вилучити з переліку дискримінантних, оскільки її внесок у дискримінацію незначний.

Детальніше про ДкА та відповідний інструментарій статистичних пакетів можна ознайомитися у [24, 9, 10, 12].

Завдання 1: виконання дискримінантного аналізу засобами пакета SPSS

1. Серед прикладів до програми SPSS вибрати файл, який містить декілька числових та принаймні одну номінативну функцію (наприклад, cars.sav, world95.sav).
2. У головному меню SPSS вибрати пункт Analyse–>Classify–>Discriminant.
3. Вказати номінативну змінну як змінну, за якою здійснюється групування (Grouping Variable). Зазначити її мінімальне та максимальне можливі значення.
4. Визначити усі числові змінні, для яких перевіряється їхній внесок у прогнозування, як незалежні (Independent Variables).
5. Вибрати покроковий метод аналізу незалежних змінних (Use stepwise method).
6. У діалоговому вікні, яке відкривається кнопкою Method, вибрати критерії оцінки прогнозу Wilk's lambda, та встановити значення для статистик F-включення (Entry) та F-вилучення (Removal).
7. У діалоговому вікні, яке відкривається кнопкою Statistics вибрати методи статистичного аналізу вхідних даних, наприклад, аналіз середніх (Means) та однофакторний дисперсійний аналіз (Univariate ANOVAs), а також включити прапорець Unstandardized Function Coefficients (для виводу коефіцієнтів канонічних функцій).
8. У діалоговому вікні, яке відкривається кнопкою Classify, вказати,

– що апріорні імовірності слід розраховувати на основі розмірів груп (Prior probabilities: Compute of group size);
– що показувати слід результати для всіх об'єктів (Display: Casewise results) та вивести підсумкову таблицю (Display: Summary table);
– що в системі координат канонічних функцій слід виводити графічні зображення усіх об'єктів (Plots: Combined groups). Якщо канонічних функцій дві, то цікавим буде графік Territorial map.

9. За отриманими результатими класифікації та апріорними імовірностями оцінити точність прогнозу.
10. Проаналізувати результати дискримінантного аналізу. Зробити висновки про якість прогнозування значень групуючої змінної від незалежних: яка із незалежних змінних має найбільший вплив на класифікацію? Чи всі обрані змінні мають вплив на групуючу змінну? Наскільки точний отриманий прогноз? Чи можна в подальшому класифікувати об'єкти лише за допомогою обраних змінних? Спробувати проінтерпретувати отримані канонічні функції.

Завдання 2: виконання дискримінантного аналізу засобами пакета Statistica

1. У пакеті Statistica виконати дискримінантний аналіз для тих самих даних: вибрати пункт меню Statistics –> Multivariate Exploratory Techniques –> Discriminant Function Analysis. Вказати групуючу (Grouping) та незалежні (Independent) змінні. Вибрати покроковий аналіз (Advanced options (stepwise analysis)).
2. У наступному діалоговому вікні (Model Definition) на закладці Advanced налаштувати параметри дискримінантного аналізу (див. Рис. 60):

а) вибрати метод Forward stepwise – при якому на кожному кроці до моделі залучатиметься змінна з найбільшим значенням статистики F-включення (процес буде припинено, коли не знайдеться жодної змінної зі значенням F-включення більшим за порогове).

При виборі методу Standart – до моделі буде включено усі незалежні змінні. При методі Backward stepwise змінні зі значенням статистики F-виключення меншим за порогове будуть виключатися з моделі.

б) встановити порогове значення толерантності – якщо хоча б одна незалежна змінна матиме значення толерантності менше за порогове, дискримінантний аналіз не проводитиметься.

в) встановити порогові значення F-включення та F-виключення (завжди обов'язково $F_{\text{to enter}} > F_{\text{to remove}}$).

г) встановити кількість кроків для аналізу (якщо кількість кроків буде мала, то аналіз припиниться раніше, ніж змінні досягнуть визначених у попередньому пункті порогових значень F включення/виключення).

д) виводити результати аналізу покрокові (At each step) або тільки підсумки (Summary only).

3. У наступному діалоговому вікні (Discriminant Function Analysis Results) переглянути результати дискримінантного аналізу. Порівняти їх з результатами, отриманими у пакеті SPSS. Зробити висновки.

Приклад виконання

Розглянемо таблицю результатів психологічного тестування (Таблиця 46).

Таблиця 46

код	лідерство	впевненість	вимогливість	скептицизм	поступливість	довірливість	добросердя	чуйність	група
1	2	2	3	3	10	10	8	7	3
2	15	9	8	8	8	7	13	11	2
3	9	5	9	9	8	3	9	4	1
4	5	5	9	9	3	9	8	4	
5	8	6	10	10	7	7	9	7	1
6	1	4	2	2	9	7	8	9	3
7	7	4	5	15	15	14	12	12	2
8	10	10	10	10	15	13	14	14	2
9	11	7	9	9	7	6	8	6	1
10	11	4	9	9	9	8	9	10	
11	10	8	8	8	9	11	11	12	2
12	6	2	8	8	8	9	12	9	3
13	11	9	6	6	5	5	10	4	1
14	4	5	4	4	9	10	10	9	2

За типом поведінки на тренінгу студентів поділили на три групи (двоє студентів участь у тренінгу не брали, для них групу не визначено). Потім було проведено тестування, за результатами якого для кожного з 14 досліджуваних визначено числові значення восьми змінних “лідерство”, “впевненість”, “вимогливість”, “скептицизм”, “поступливість”, “довірливість”, “добросердя” та “чуйність”.

Необхідно з’ясувати, чи можна за результатами тестування визначити тип поведінки студента (групу), оцінити якість такого прогнозу, визначити групову приналежність студентів, які пропустили тренінг.

Приклад 1а. У пакеті SPSS виконаємо для цих даних дискримінантний аналіз, обравши змінну “група” за класифікуючу, а результати тесту (змінні v1, v2, v3, v4, v5, v6, v7 та v8) – за незалежні.

Стан елементів діалогового вікна “Discriminant Analysis” та його підлеглих вікон “Classification” “Statistics” та “Method” наведено на рисунках Рис. 56, Рис. 57, Рис. 58 та Рис. 59 відповідно.

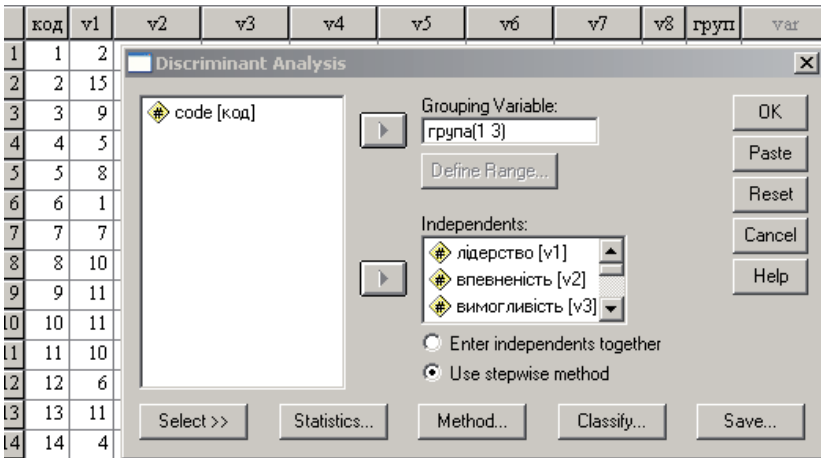


Рис. 56

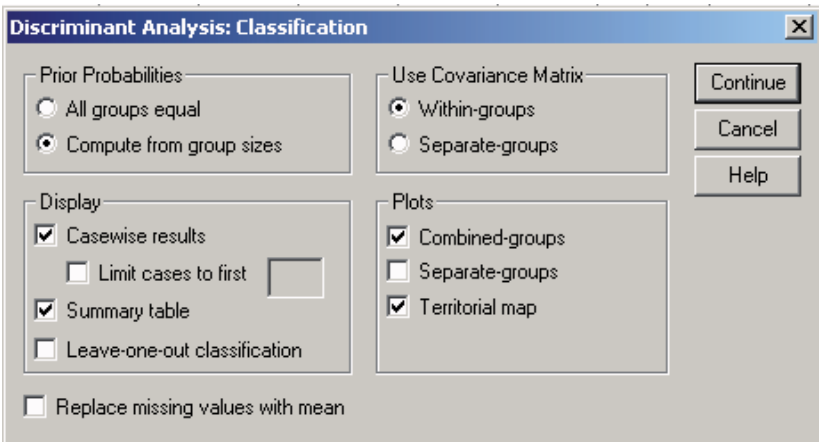


Рис. 57

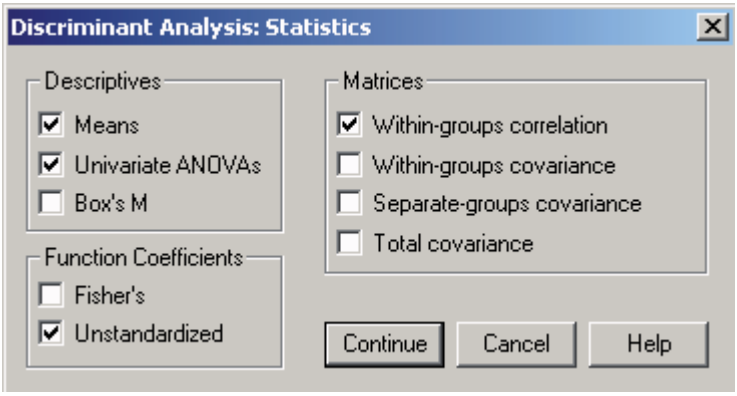


Рис. 58

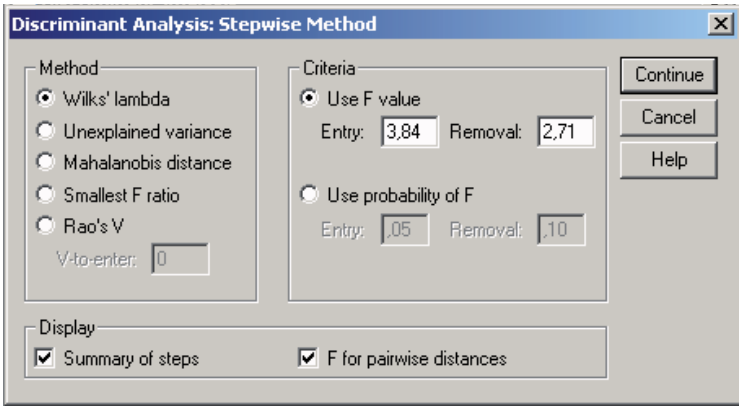


Рис. 59

В результаті аналізу отримаємо:

1. Порівняння групових середніх для кожної змінної за методом однофакторного дисперсійного аналізу (Tests of Equality of Group Means).
2. Об'єднану кореляційну матрицю (Pooled Within-Groups Matrices).
3. Звіт про покрокове включення/виключення змінних до аналізу, з якого видно як на кожному кроці збільшувалася дискримінативна здатність набору даних. Про це свідчить зменшення Λ-Уілкса (Таблиця 47).

Теоретично кроків могло бути 16. Однак при досягненні одним з критеріїв статистичної значущості (толерантністю, F-статистикою вилучення/ включення) граничних значень подальші обрахунки припиняються. У даному випадку це сталося на другому кроці.

Таблиця 47

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	чуйність	,201	1	2	9,000	17,917	2	9,000	,001
2	вимогливість	,077	2	2	9,000	10,435	4	16,000	,000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 16.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

4. Таблиця "Variables in the Analysis" дозволяє прослідкувати зміни (у даному випадку зменшення) значень толерантності, F-вилучення та Λ-Уілкса при включенні змінних до списку.

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	чуйність	1,000	17,917	
2	чуйність	,595	29,413	,641
	вимогливість	,595	6,457	,201

5. Таблиця “Variables Not in the Analysis” дозволяє з’ясувати, чому деякі змінні будуть вилучені з подальшого аналізу. До уваги береться статистика F-включення: змінні з максимальним значенням цієї статистики вилучаються, а у таблиці залишаються змінні, які найменше впливають на розрізнення класів.

Слід звернути увагу на те, що всі включені до подальшого аналізу змінні мають рівень телерантності вищій за 0,1 та рівень F-критерія вищій визначеного порогового (2,71). А от для виключених змінних значення F-критерія менший порогового.

6. Таблиця Wilks' Lambda аналогічно до таблиці Variables Entered/Removed (див. Таблиця 47) показує, що на кожному кроці дискримінативна здатність набору змінних покращується.
7. Таблиця Eigenvalues (власні значення) показує, що перша канонічна функція у майже у 20 разів (95,5% проти 4,5%) інформативніша за другу.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	8,327[a]	95,5	95,5	,945
2	,396[a]	4,5	100,0	,533

a. First 2 canonical discriminant functions were used in the analysis.

8. Таблиця Wilks' Lambda містить значення Λ=0,077 та його статистичну значущість (Sig $\approx$ 0) для усього набору канонічних функцій у першому рядку та Λ=0,716 та його статистичну значущість (Sig = 0,092) для набору канонічних функцій після виключення першої. Видно, що дискримінантна здатність повного набору канонічних функцій надзвичайно висока. Після виключення першої функції дискримінантна здатність набору падає, як падає і

статистична значущість. Це означає, що внесок другої функції незначний і її смисл важко інтерпретувати. Висока статистична значущість (Sig < 0,05) свідчила б про її значний внесок у класифікацію.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,077	21,817	4	,000
2	,716	2,837	1	,092

9. Таблиця стандартизованих коефіцієнтів канонічних функцій (Standardized Canonical Discriminant Function Coefficients) дозволяє визначити співвідношення внесків дискримінантних змінних до кожної з канонічних функцій. Так найбільший внесок до першої функції дає змінна "чуйність", а до другої – "вимогливість".

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
вимогливість	-,959	,872
чуйність	1,283	,186

10. Наступна таблиця містить коефіцієнти канонічних дискримінантних функцій:

Canonical Discriminant Function Coefficients

	Function	
	1	2
вимогливість	-,393	,357
чуйність	,809	,117
(Constant)	-4,328	-3,454

Unstandardized coefficients

11. Таблиця структурних коефіцієнтів канонічних функції (Structure Matrix) показує кореляції канонічних функцій з усіма змінними, що дозволяє проінтерпретувати канонічні функції. Так перша функція має найбільшу, кореляцію зі змінною "чуйність", а друга – зі змінною "вимогливість". Значна кореляція зі змінними, що були виключені з аналізу, дозволяє краще інтерпретувати утворені канонічні функції.

Structure Matrix

	Function	
	1	2
довірливість[a]	,521*	,177
поступливість[a]	,506*	,175
вимогливість	-,14	,990*
чуйність	,672	,740*
добросердя[a]	-,10	,671*
лідерство[a]	-,26	,559*
скептицизм[a]	,274	,519*
впевненість[a]	,185	,291*

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

12. У таблиці значень канонічних функцій для групових центроїдів (Functions at Group Centroids) наведено координати центроїдів усіх груп, що дозволяє інтерпретувати канонічні функції відносно їхньої ролі у розрізненні класів. Так група 1 має найбільше від'ємне значення, група 2 – найбільше додатне значення за за функцією 1, а група 3 – середнє між ними.

Functions at Group Centroids

	Function	
група	1	2
1	-3,420	,195
2	2,308	,403
3	,713	-,931

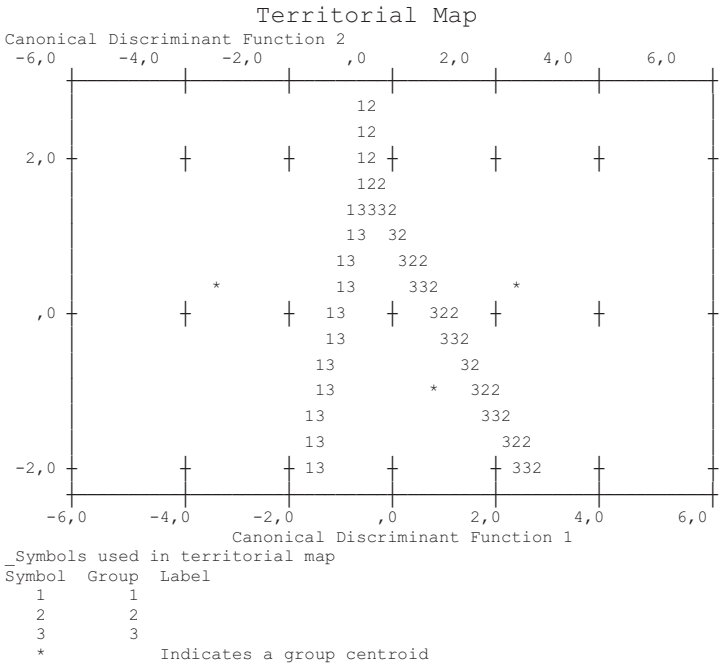
Unstandardized canonical discriminant functions evaluated at group means

13. Таблиця апріорних імовірностей показує долю об'єктів кожного класу у вибірці до аналізу.

Prior Probabilities for Groups

група	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	,333	4	4,000
2	,417	5	5,000
3	,250	3	3,000
Total	1,000	12	12,000

14. На графіку Territorial Map двовимірний простір канонічних змінних поділено на зони визначених класів. Символом * позначено центроїди.



15. У таблиці статистика об’єктів (Casewise Statistics) містяться відомості про приналежність кожного об’єкта до класу: дійсну (Actual) та спрогнозовану за допомогою ДкА (Predicted). Об’єкти, які у результаті ДкА потрапили в інший клас, відзначають “зірочками”. У данному випадку

змінено класову приналежність об'єкта №14, та визначено класову приналежність об'єктів №4 та №10.

Таблиця 48

Casewise Statistics

Case Number	Original										
	Actual Group	Highest Group					Second Highest Group			Discriminant Scores	
		Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1	Function 2
			p	df							
1	3	3	,702	2	,967	,708	2	,033	8,493	,157	-1,563
2	2	2	,653	2	,840	,854	3	,160	3,142	1,430	,690
3	1	1	,482	2	1,000	1,460	3	,000	29,860	-4,628	,227
4	ungrouped	1	,482	2	1,000	1,460	3	,000	29,860	-4,628	,227
5	1	1	,540	2	,999	1,232	3	,001	14,410	-2,593	,935
6	3	3	,260	2	,583	2,690	2	,417	4,384	2,169	-1,686
7	2	2	,433	2	,972	1,676	3	,028	7,764	3,418	-,264
8	2	2	,300	2	,997	2,410	3	,003	12,782	3,072	1,755
9	1	1	,887	2	1,000	,239	3	,000	15,791	-3,009	,461
10	ungrouped	2	,100	2	,512	4,606	3	,484	3,698	,228	,929
11	2	2	,920	2	,957	,168	3	,043	5,352	2,239	,807
12	3	3	,255	2	,760	2,735	2	,220	6,238	-,189	,455
13	1	1	,583	2	1,000	1,080	3	,000	17,327	-3,449	-,844
14	2	3**	,798	2	,654	,451	2	,346	2,748	1,383	-,972

**. Misclassified case

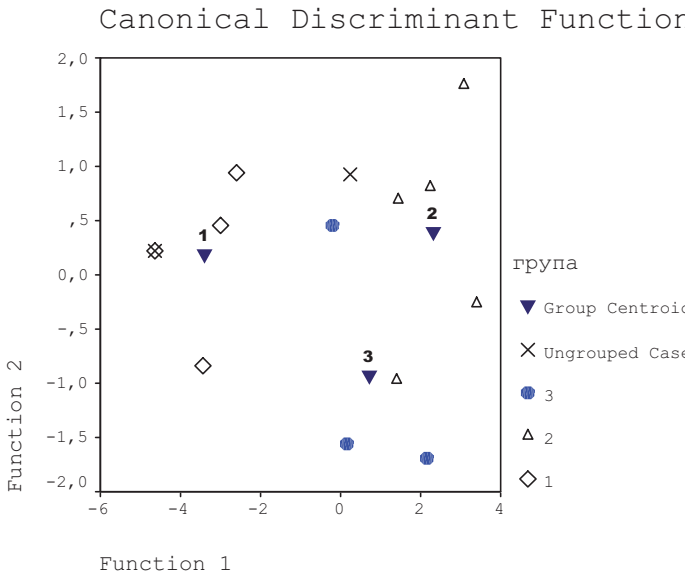
16. Як видно з підсумкової таблиці Classification Results, точність прогнозу становить 11/12=91,7%. Обчислимо τ-статистику. Відповідно до таблиці апріорних імовірностей (див. п.13) Σpini = (0,333*4)+(0,417*5)+(0,250*3) = 4,17. Тоді у числельнику отримаємо 11 – 4,17 = 6,833, а у знаменнику 12 – 4,17=7,833. Відповідно значенням τ-статистики буде 0,872, тобто результат прогнозу на 87,2% краще випадкового результату (буде допущено приблизно 1 помилку замість 8).

Classification Results[a]

			Predicted Group Membership			
		група	1	2	3	Total
Original	Count	1	4	0	0	4
		2	0	4	1	5
		3	0	0	3	3
		Ungrouped cases	1	1	0	2
	%	1	100,0	,0	,0	100,0
		2	,0	80,0	20,0	100,0
		3	,0	,0	100,0	100,0
		Ungrouped cases	50,0	50,0	,0	100,0

a. 91,7% of original grouped cases correctly classified.

17. На графіку канонічних дискримінантних функцій (Canonical Discriminant Functions) зображено групові центроїди та об'єкти в системі координат канонічних функцій. Графік дозволяє візуально оцінити якість класифікації та інтерпретувати канонічні функції.



Приклад 1б (Statistica).

При виконанні завдання у пакеті Statistica (Statistics –> Multivariate Exploratory Techniques –> Discriminant Function Analysis) у покроковому режимі у діалоговому вікні Model Definition на закладці Descriptives отримаємо (у закладці Within) звіти про об'єднану міжгрупову кореляцію (Pooled within-groups covariance & correlation), середні та частоти (Means & number of cases), міжгрупові стандартні відхилення (Within groups standart deviations) (див. п.1-2 прикладу 1а). А також графічне представлення груп (Categorized histogram, Box plot of means, Categorized scatterplot, Categorized normal probability plot), і на закладці All Cases – звіт про кореляцію (Total correlation ...), діаграми розсіювання (Plot of total correlations) та блочні діаграми (Box plot of means).

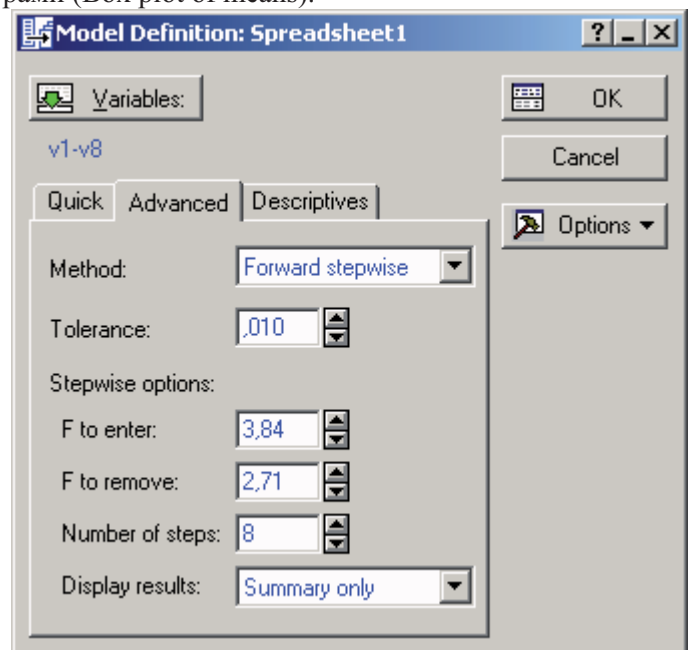


Рис. 60

Потім у діалоговому вікні Model Definition виконаємо настройки, як показано на Рис. 60 та перейдемо до наступного діалогового вікна (OK). Якщо серед вибраних незалежних змінних є змінні з толерантністю меншою вказаного порогового значення (0,01), то аналіз не проводитиметься, а буде виведено повідомлення про помилку.

На закладці Quick діалогового вікна Discriminant Function Analysis Results отримаємо звіти про змінні, які були вкючені та виключені з аналізу (відповідно Таблиця 49 та
Таблиця 50):

Таблиця 49

	Discriminant Function Analysis Summary (liri) Step 2, N of vars in model: 2; Grouping:group (3 grps) Wilks' Lambda: ,07679 approx. F (4,16)=10,435 p< ,0002					
N=12	Wilks' Lambda	Partial Lambda	F-remove (2,8)	p-level	Toler.	1-Toler. (R-Sqr.)
v8	0,641434	0,119714	29,41307	0,000205	0,595175	0,404825
v3	0,200740	0,382528	6,45676	0,021412	0,595175	0,404825

Таблиця 50

	Variables currently not in the model (liri) Df for all F-tests: 2,7					
N=12	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr.)
v1	0,050048	0,651763	1,870053	0,223518	0,619551	0,380449
v2	0,046213	0,601818	2,315717	0,169094	0,881251	0,118749
v4	0,061054	0,795094	0,901993	0,448194	0,655289	0,344711
v5	0,072529	0,944533	0,205534	0,818956	0,713378	0,286622
v6	0,076037	0,990212	0,034597	0,966159	0,696920	0,303080
v7	0,059341	0,772780	1,029105	0,405690	0,539512	0,460488

Отримані результати аналогічні до отриманих у пакеті SPSS (див. п.4-5 прикладу 1а).

На закладці Advanced отримаємо звіт про покрокове виконання аналізу (Stepwize Analysis Summary) аналогічно до п.3 прикладу 1а та звіт про канонічні дискримінантні функції (Perform canonical analysis).

На закладці Advanced у вікні Canonical Analysis отримаємо:

1) аналіз значущості отриманих канонічних функцій (Chi square tests of successive roots) аналогічний до п.7-8 прикладу 1а;
2) стандартизовані коефіцієнти канонічних функцій аналогічно до п.9 прикладу 1а;
3) факторну структуру (Factor structure Matrix), тобто кореляції незалежних змінних з канонічними функціями (аналогічно п.11 прикладу 1а);

4) координати центроїдів груп у системі координат утворених канонічних функцій (Means of canonical variables).

На закладці Canonical scores у вікні Canonical Analysis отримаємо:

1) діаграму розсіювання у системі координат канонічних функцій (Scatterplot of canonical scores);

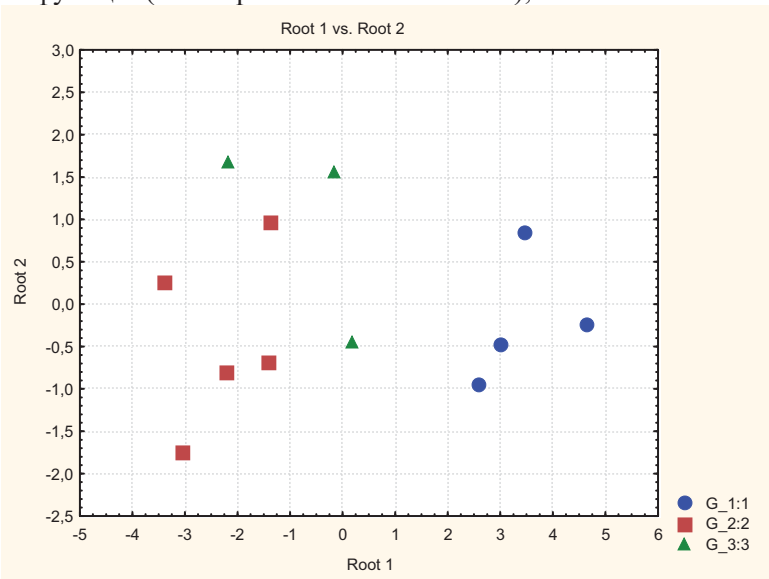


Рис. 61

2) значення канонічних функцій для кожного досліджуваного об'єкта (Canonical scores for each case) (в пакеті SPSS ці значення представляє Таблиця 48);
3) як і в пакеті SPSS, отримані для кожного досліджуваного об'єкта значення канонічних функцій можна зберегти як нові змінні (Save canonical score).

Результати класифікації буде подано на закладці Classification діалогового вікна Discriminant Function Analysis Results Рис. 62:

1) коефіцієнти канонічних дискримінантних функцій (Classification functions) (див. п.10 прикладу 1а);
2) зведені результати класифікації (Classification matrix). У даному випадку, як і в пакеті SPSS (п.16 прикладу 1а), буде отримано точність прогнозу 91,7%.

3) таблиця класифікації досліджуваних об'єктів (Classification of cases) містить розподіл об'єктів по класах у порядку переваги. Аналогічні результати в SPSS містить Таблиця 48.
4) відстані об'єктів від групових центроїдів (Squared Mahalanobis distances) в SPSS також містить Таблиця 48;

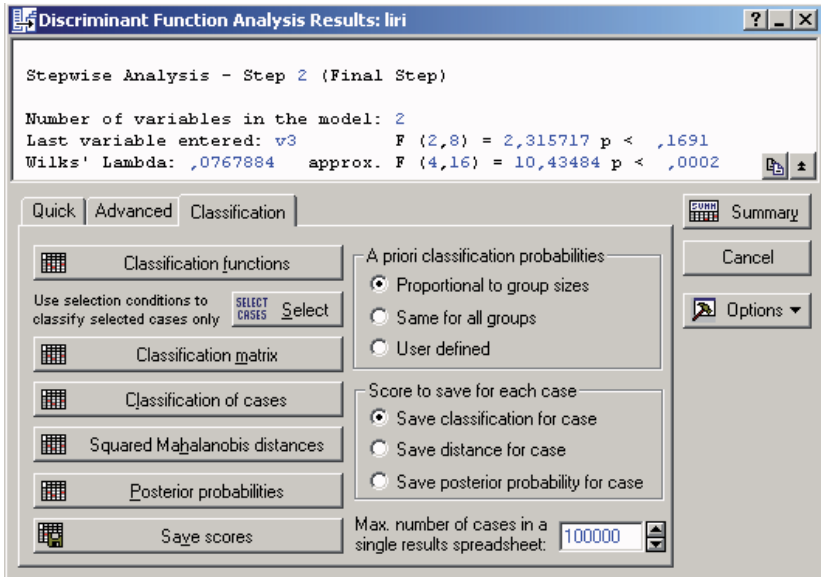


Рис. 62

Group	Classification Matrix (liri) Rows: Observed classifications Columns: Predicted classifications Percent Correct	G_1:1 p=,33333	G_2:2 p=,41667	G_3:3 p=,25000
G_1:1	100,0000	4	0	0
G_2:2	80,0000	0	4	1
G_3:3	100,0000	0	0	3
Total	91,6667	4	4	4

5) у таблиці (Таблиця 51) постеріорних ймовірностей (Posterior probabilities) наведено імовірності приналежності кожного об'єкта до кожного з класів (у таблиці Таблиця 48 з прикладу 1а ці імовірності наведено у порядку переваги) та апріорні імовірності (див. п.13 прикладу 1а). Для того, щоб апріорні імовірності обчислювалися пропорційно

розмірам груп, попередньо у вікні Classification було встановлено перемикач Proportional to group sizes (Рис. 62).

Таблиця 51

Case	Posterior Probabilities (liri in Workbook4) Incorrect classifications are marked with *			
	Observed Classif.	G_1:1 p=,33333	G_2:2 p=,41667	G_3:3 p=,25000
1	G_3:3	0,000652	0,032843	0,966505
2	G_2:2	0,000007	0,839561	0,160432
3	G_1:1	0,999999	0,000000	0,000001
4	---	0,999999	0,000000	0,000001
5	G_1:1	0,998957	0,000012	0,001031
6	G_3:3	0,000000	0,416817	0,583183
7	G_2:2	0,000000	0,972211	0,027789
8	G_2:2	0,000000	0,996654	0,003346
9	G_1:1	0,999684	0,000001	0,000315
10	---	0,004040	0,512066	0,483893
11	G_2:2	0,000000	0,957003	0,042997
12	G_3:3	0,020777	0,219715	0,759508
13	G_1:1	0,999778	0,000000	0,000222

Отже, в обох пакетах отримано однакові результати. Однак для такої малої вибірки результати класифікації і, відповідно, прогнозу будуть нестійкими, в чому легко переконатись, змінивши вхідні дані. Наприклад, віднести об'єкт №4 до групи 1, об'єкт №10 до групи 2, а об'єкт №14 до групи 3, не змінюючи при цьому параметрів включення/виключення.

Контрольні запитання

1. Опишіть призначення дискримінантного аналізу. Чому дискримінантний аналіз називають методом "класифікації з навчанням"?
2. Що таке "центроїд"? В чому полягає геометрична інтерпретація правила інтерпретації?
3. Як будуються канонічні дискримінантні функції? Скільки їх можна побудувати?
4. Критеріями чого є "власне значення", λ-Уілкса та χ^2? Якими повинні бути їхні значення?

5. Що показують структурні коефіцієнти канонічних функцій? Що показує кореляція визначених канонічних функцій з незалежними змінними?
6. Що таке “толерантність”? Які значення вона може мати? При яких її значеннях змінну слід вилучити з аналізу? Чому (як проінтерпретувати вилучення)?
7. Показником чого є статистика F-вилучення?
8. Які основні результати дискримінантного аналізу?
9. Як можна оцінити точність прогнозу, отриманого в результаті дискримінантного аналізу?
10. Як можна графічно представити результати дискримінантного аналізу?

Тема 8: “Кластерний аналіз”

Мета:

Студенти повинні знати:
- призначення та види кластерного аналізу;
- основні алгоритми кластерного аналізу (у загальних рисах);
- способи визначення міри схожості між об’єктами;
- методи визначення міри схожості між кластерами;
- сфери застосування кластерного аналізу;
- застереження при застосуванні кластерного аналізу.

Студенти повинні уміти:
- подавати експериментальні дані у вигляді, зручному для обчислень;
- за допомогою статистичних пакетів виконувати ієрархічний кластерний аналіз, отримувати результати у вигляді дендрограми, інтерпретувати отримані результати;
- за допомогою статистичних пакетів виконувати кластерний аналіз методом к-середніх, оцінювати якість кластеризації.

Теоретичні відомомсті.

Призначення кластерного аналізу

Кластерним аналізом (від англ. *cluster* – гроно, кущ, рій, скупчення, жмуток) називають сукупність методів, призначених для виявлення структури даних. Задачами кластерного аналізу є:
- розробка типології або класифікації;
- дослідження корисних концептуальних схем групування об’єктів;
- представлення гіпотез на основі дослідження даних;
- перевірка гіпотез про існування виділених деяким способом типів досліджуваних даних.

Кластеризація є логічним продовженням класифікації, але її відносять до методів “навчання без вчителя”, оскільки при формуванні кластерів не використовують “навчаючу множину” (множину об’єктів, кластерна приналежність яких заздалегідь відома).

У наукових дослідженнях кластеризація при правильному застосуванні дозволяє навіть відкривати нові перспективні напрямки. Яскравим прикладом є періодична таблиця елементів: безперечною заслугою Д.Менделєєва є те, що у 1869 р. він поділив 60 відомих на той час елементів на кластери або періоди за схожими характеристиками. Вивчення причин об’єднання елементів у явно виражені кластери визначило пріоритети

наукових досліджень на роки вперед. І лише через 50 років засобами квантової фізики вдалося науково обгрунтувати такий поділ [7].

Однак кластеризація є і поки залишається описовою процедурою: на її основі не можна робити статистичних висновків. Правильність кластеризації оцінюють лише непрямими методами:

- шляхом встановлення контрольних точок;
- перевіркою стабільності кластеризації шляхом введення в модель нових змінних;
- порівнянням результатів застосування різних методів кластеризації (створення схожих кластерів при використанні різних методів вказує на правильність кластеризації).

Ієрархічні методи кластерного аналізу

На сьогоднішній день розроблено більше ста різноманітних алгоритмів кластеризації, які поділяються на дві групи – ієрархічні та неієрархічні (ітераційні) алгоритми.

При застосуванні ієрархічного агломеративного (об'єднувального) кластерного аналізу (AGglomerative NESting, AGNES) на першому кроці кожен об'єкт розглядається як окремий кластер. На наступних кроках найбільш схожі між собою об'єкти об'єднуються у кластери. Процес продовжується доти, поки всі об'єкти не об'єднаються в один кластер.

При застосуванні ієрархічного дивізимного (подільного) кластерного аналізу (Divisive ANAlysis, DIANA) на першому кроці уся сукупність досліджуваних об'єктів розглядається як один кластер, який на наступних кроках поділяється на менші.

Ієрархічні методи є найбільш наочними, однак застосовуються як правило на невеликих наборах даних. Результати ієрархічного кластерного аналізу зручно представляти у вигляді дендрограм (Рис. 65). Кожен рівень дендрограми відповідає одному кроку кластеризації.

Усі методи кластерного аналізу базуються на двох основних припущеннях:

1) досліджувані ознаки в принципі допускають поділ досліджуваної сукупності об'єктів на кластери;

2) шкали вимірювання усіх досліджуваних ознак співрозмірні, тобто представлені у нормалізованому

(стандартизованому) масштабі, але з урахуванням важливості (ваги) тої чи іншої ознаки.

Найчастіше для стандартизації усі дані ділять на стандартне квадратичне відхилення досліджуваної ознаки, а для врахування ваги – множать на ваговий коефіцієнт або експертну оцінку її важливості.

Формальна постановка задачі кластеризації полягає у тому, що є набір даних з такими властивостями:

– кожен екземпляр виражений числовим значенням;
– клас для кожного екземпляра невідомий.

Необхідно знайти:

– спосіб порівняння даних між собою;
– спосіб кластеризації (об'єднання/поділ на кластери);
– розподіл даних по кластерах.

Міри схожості об'єктів

Критерієм схожості об'єктів кластеризації є "відстань" між ними у просторі досліджуваних змінних.

Найбільш популярною мірою схожості є коефіцієнт кореляції Пірсона: $r_{ij} = \dfrac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}$.

Основним недоліком такої міри є чутливість до форми за рахунок зниження чутливості до величини різниці між змінними.

Іншими мірами схожості є, наприклад:

1) евклідова відстань $d_{2ij} = \sqrt{\sum_{t=1}^{m}(x_{it} - x_{jt})^2}$, де t – розмірність простору. У просторі двох змінних $d_{2ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Тут d_{ij} – відстань між і-тим та j-тим об'єктами, X та Y – досліджувані ознаки.
2) відстань за Хеммінгом (або Манхеттенська або відстань міських кварталів): $d_{Hij} = \sum_{t=1}^{m}\left|x_{it} - x_{jt}\right|$, – на неї менше впливають окремі "викиди" (великі різниці), оскільки їх не підносять до степеня;

3) відстань Чебишова: $d_{\infty ij} = \max\limits_{1 \le t \le m} \left| x_{it} - x_{jt} \right|$, – дозволяє розрізнити об'єкти, які відрізняються лише однією координатою;
4) відстань Махаланобіса: $d_M(x_i, x_j) = (x_i - x_j)\, S^{-1}\, (x_i - x_j)^t$, – дає хороший результат при застосуванні на конкретному класі і погано працює на усій множині вхідних даних;
5) пікова відстань: $d_{Lij} = \frac{1}{m} \sum_{t=1}^{m} \frac{\left| x_{it} - x_{jt} \right|}{x_{it} + x_{jt}}$, – відстань в ортогональному просторі, тобто необхідною умовою її застосування є незалежність змінних.

Коректно обрати міру відстані можна лише при врахуванні характеристик вхідних даних.

Для бінарних змінних відстані обчислюють за коефіцієнтами асоціативності [24]. Позначивши літерами *a*, *b*, *c*, *d* комірки у 2x2-таблиці асоціативності, обчислюють простий коефіцієнт $S = \frac{a+d}{a+b+c+d}$, або коефіцієнт Жаккара $J = \frac{a}{a+b+c}$.

1 – наявність ознаки, 0 – відсутність ознаки	1	0
1	*a*	*b*
0	*c*	*d*

Коефіцієнт Жаккара виявився частковим випадком (коли усі дані двійкові) для коефіцієнта Гауера: $s_{ij} = \sum_{k=1}^{p} S_{ijk} / \sum_{k=1}^{p} W_{ijk}$, де W_{ijk} – вагова змінна, яка дорівнює 1, якщо порівняння об'єктів за ознакою *k* слід враховувати, та 0 – у протилежному випадку; S_{ijk} – "внесок" у схожість об'єктів *i* та *j* ознаки *k*.

У сферах, де доводиться використовувати бінарні дані, користуються також імовірнісними коефіцієнтами схожості – при створенні кластерів обчислюється "інформаційний виграш" об'єднання.

Міри схожості кластерів

Об'єднання у кластери (визначення міри схожості між двома кластерами, а не об'єктами) здійснюється за правилами об'єднання кластерів. Такими є, наприклад,
1) *метод найближчого сусіда* (*поодинокого зв'язку*) – відстань між кластерами визначається відстанню між двома найбільш близькими об'єктами у різних кластерах;

2) *метод найвіддаленішого сусіда* (*повного зв'язку*) – відстань між кластерами визначається найбільшою відстанню між двома довільними об'єктами у різних класах;
3) *метод незваженого попарного середнього* – відстань між кластерами визначається за середнім між усіма парами об'єктів в них;
4) *метод зваженого попарного середнього* – до середнього між усіма парами об'єктів в кластерах застосовується ваговий коефіцієнт (кількість об'єктів у кластері);
5) *незважений центроїдний метод* – за відстань між кластерами береться відстань між їхніми центрами ваги;
6) *метод Варда* – для оцінки відстаней використовує дисперсійний аналіз: за відстань між кластерами беруть приріст суми квадратів відстаней об'єктів до центрів кластерів, що отримується в результаті об'єднання.

Позначення методів кластеризації у статпакетах

SPSS	***Statistica***	
Ієрархічний кластерний аналіз		
Cluster Method	***Amalgamation (linkage) rule***	***Методи об'єднання***
Between-group linkage	Weighted pair-group average	Міжгрупове зв'язування
Within-group linkage	Unweighted pair-group average	Внутрігрупове зв'язування
Nearest neightbor	Single linkage	Поодиноке зв'язування
Furthest neighbor	Complete linkage	Повне зв'язування
Centroid clustering	Unweighted pair-group centroid	Центроїдний метод
Median clustering	Weighted pair-group centroid (median)	Медіанний метод
Ward's method	Ward's method	Метод Варда
Measure	***Distance measure***	***Міри відстані***
Interval	***Для числових даних***	
Squared Euclidian distance	Squared Euclidian distance	Нормалізована Евклідова відстань
Euclidian distance	Euclidian distance	Евклідова відстань
Block	City-block (Manhattan) distances	Відстань міських кварталів

Chebyshev	Chebyshev distance metric	Відстань Чебишова
Minkowski	Power: SUM(ABS(x-y)**p)**1/r	Метрика Мінковського
	Percent disagreement	Відсоток відмінностей
Pearson correlation	1-Pearson r	1 – коефіцієнт кореляції за Пірсоном
Cosine		Косинусна відстань (1 – косинус кута між об'єктами)
Count	***Для номінативних даних***	
Chi-square measure		Міра Хі-квадрат
Phi- square measure		Міра фі-квадрат
Binary[17]	***Для дихотомічних (бінарних) даних***	
Jaccard		Міра Жаккара

Ітеративні методи кластерного аналізу

При великій кількості спостережень застосовують ітеративні методи кластерного аналіза: формування нових кластерів припиняють при досягненні певної їх кількості.

Найбільш популярним серед ітераційних методів є алгоритм k-середніх. В результаті його застосування буде побудовано k кластерів (потрібно попередньо мати гіпотезу про імовірну кількість кластерів) максимально віддалених один від одного.

На першому кроці алгоритму випадковим чином обираються k точок – центрів майбутніх кластерів. Усі об'єкти розподіляються по кластерах.

На наступних кроках ітераційного процесу центри кластерів переобчислюються з урахуванням віднесених до кластеру об'єктів, і об'єкти знову перерозподіляються між кластерами. Процес припиняється, коли кластери стабілізувалися, тобто

[17] Вказано лише одну міру для дихотомічних даних з переліку, який має пакет SPSS.

після перерозподілу об'єкти залишаються у тих кластерах, де і були, або виконано максимальну кількість ітерацій.

Показником якості кластеризації є достовірна різниця середніх, обчислених для різних кластерів.

Метод k-середніх є простим, зрозумілим, прозорим, але надзвичайно чутливим до викидів (екстремальних даних) та повільним для великих баз даних.

Узагальненням ітеративних методів є методи нечітких середніх, у яких кожен кластер є нечіткою множиною, і кожен елемент належить різним кластерам з різним степенем приналежності. Серед нечітких методів можна назвати метод Fuzzy C-Means та метод кластеризації за Гюстафсоном-Кесселем [7]. Огляд деяких нових алгоритмів кластеризації та посилання на літературу про них можна знайти в [6].

Перспективним розвитком методів кластеризації є застосування теорії графів для удосконалення обчислювальних процедур та створення нуль-гіпотези для перевірки кількості кластерів у матриці схожості [24].

Завдання 1: Ієрархічний кластерний аналіз

1. Отримати дані з файлу[18].
2. Виконати ієрархічний кластерний аналіз у пакеті SPSS: Analyze –> Classify –> Hierarchical Cluster ... У діалоговому вікні кластерного аналізу вказати змінні (Variable(s)), для яких виконуватиметься кластеризація.
3. Встановити параметри кластеризації: у діалоговому вікні Plots... включити прапорець Dendrogram; у діалоговому вікні Method вибрати метод об'єднання у кластери (Cluster Method) – за замовченням встановлено метод Between-group linkage, – та міру відстані між об'єктами (Measure) – за замовченням встановлено Squared Euclidian distance. Зберегти результати кластеризації у вигляді

[18] Для виконання завдання рекомендується попередньо зібрати відповідний статистичний матеріал (див. Індивідуальне завдання 3, с. 184). Якщо таких матеріалів немає, можна використати данні з файлів, що додаються до статистичного пакета SPSS, або вибірку C, вважаючи кожен стовпець окремою змінною.

окремої змінної, вказавши у діалоговому вікні Save необхідну кількість кластерів (Single solution).

4. Виконати кластерний аналіз з іншими параметрами: мірою відстані та методом об'єднання кластерів (див. с.173).
5. Зробити висновки, проінтерпретувати отримані результати, оцінити (візуально) достовірність отриманої кластеризації.
6. Виконати ієрархічний кластерний аналіз у пакеті Statistica: Statistics –> Multivariate Exploratory Technics –> Cluster Analysis –> Joining (Tree Clustering). У діалоговому вікні встановити параметри кластеризації: змінні (Variables), метод об'єднання (Amalgamation rule), міру відстані (Distance measure), вид подання вхідних даних[19] (Input file) та для яких об'єктів здійснювати кластеризацію: для досліджуваних об'єктів (Cases (rows)) чи для змінних[20] (Variables (columns)).
7. У наступному діалоговому вікні на закладці Advanced переглянути результати кластеризації.
7. Порівняти процедури проведення та результати кластерного аналізу, виконаного засобами двох пакетів.

Завдання 2: Ітеративний кластерний аналіз

1. Для даних з попереднього завдання визначити кількість кластерів (наприклад, 2-4).
2. Виконати k-кластерний аналіз засобами пакета SPSS: Analyze –> Classify –> K-Means Cluster Analysis. У діалоговому вікні встановити параметри кластеризації: змінні (Variables) та кількість кластерів (Number of clusters. У допоміжному діалоговому вікні Iterate встановити кількість ітерацій (Maximum Iterations) та включити прапорець Use running means – коригувати середні. Включити всі прапорці у допоміжних діалогових вікнах Save та Options.
3. За даними таблиці Final Cluster Centers побудувати лінійний графік середніх для кожного кластера.
4. Виконати k-кластерний аналіз засобами пакета Statistica (K-means clustering), встановивши у відповідному

[19] Дані можуть бути представлені у вигляді рядків (Raw data) або матриці відстаней (Distance Matrix).

[20] Кластеризація змінних можлива лише для даних, поданих у вигляді рядків (Raw data).

діалоговому вікні параметри кластеризації: змінні (Variables), об'єкти кластеризації[21] (Cluster), кількість кластерів (Number of clusters), кількість ітерацій (Number of iterations) та спосіб визначення початкових центрів кластерів[22] (Initial cluster centers).

5. Порівняти процедури проведення та результати кластерного аналізу, виконаного засобами двох пакетів. Зробити висновки.

Приклади виконання

Приклад 1: ієрархічний (агломеративний) кластерний аналіз.

Виконаємо у пакеті SPSS кластерний аналіз для даних, поданих у таблиці Таблиця 46 (с.153). Враховуючи, що змінні у даному прикладі інтуїтивно досить легко проінтерпретувати, виконаємо кластеризацію змінних. Для цього в основному діалоговому вікні кластерного аналізу перемикач Cluster встановлено у положення Variables (Рис. 63).

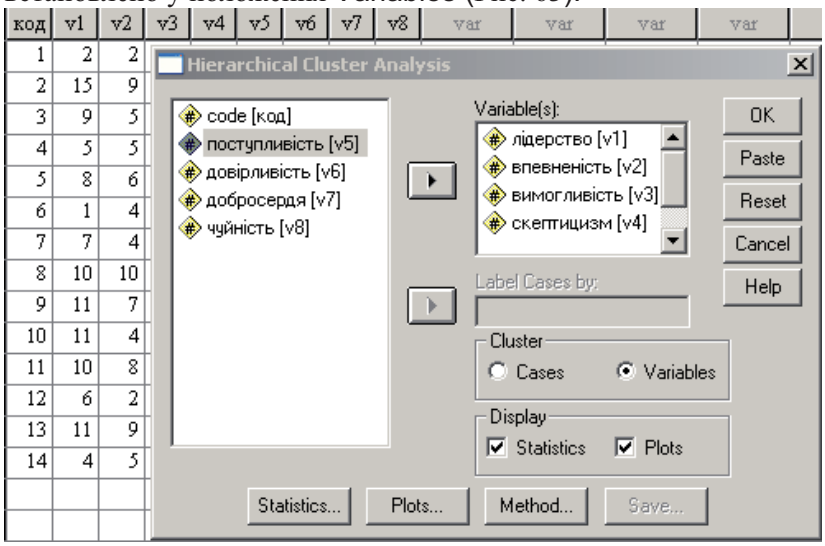


Рис. 63

[21] Див. зноску 20 на с.176.

[22] За замовченням встановлено "Sort distances and take observations at constant intervals".

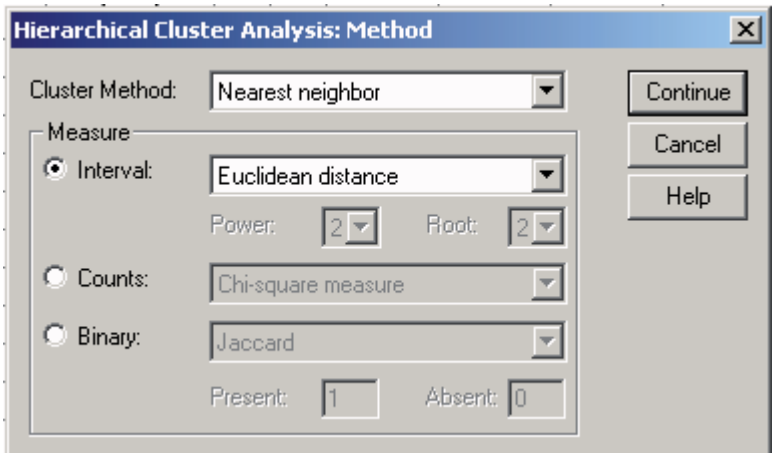


Рис. 64

У допоміжному діалоговому вікні Statistica можна залишити налаштування, виконані за замовченням, а у діалоговому вікні Plots слід включити прапорець Dendrogram. У допоміжному діалоговому вікні Method необхідно вказати методи визначення відстані між об'єктами та об'єднання кластерів (Рис. 64).

Графічно результат кластеризації буде представлено дендрограмою (Рис. 65), на якій добре видно, що усі змінні розподілилися між двома кластерами.

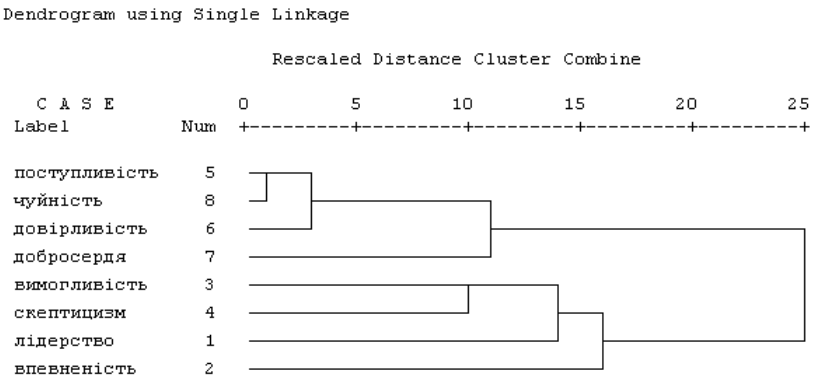


Рис. 65

Крім дендрограми звіт міститиме таблиці послідовності злиття Agglomaration Schedule та Vertical Icicle.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	5	8	7,616	0	0	2
2	5	6	8,185	1	0	4
3	3	4	10,000	0	0	5
4	5	7	10,149	2	0	7
5	1	3	10,770	0	3	6
6	1	2	11,402	5	0	7
7	1	5	13,675	6	4	0

Vertical Icicle

Number of clusters	Case														
	добросердя		довірливість		чуйність		поступливість		впевненість		скептицизм		вимогливість		лідерство
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X		X	X	X	X	X	X	X
3	X	X	X	X	X	X	X		X		X	X	X	X	X
4	X	X	X	X	X	X	X		X		X	X	X		X
5	X		X	X	X	X	X		X		X	X	X		X
6	X		X	X	X	X	X		X		X		X		X
7	X		X		X	X	X		X		X		X		X

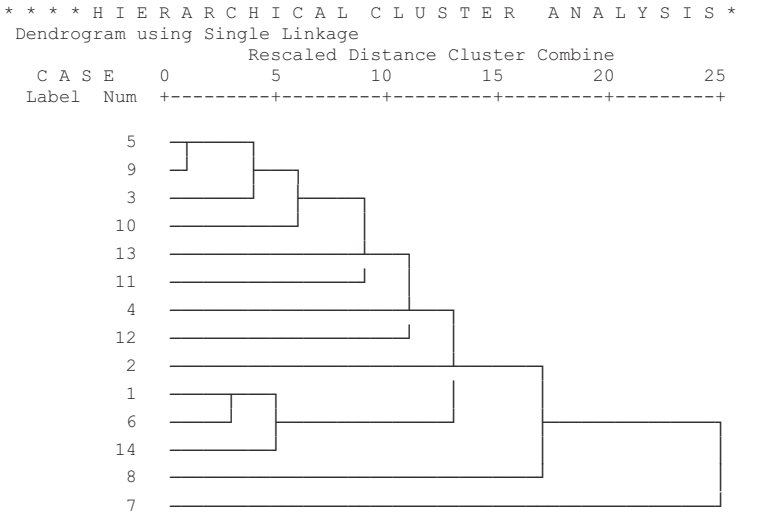


Рис. 66

Якщо виконати кластеризацію даних по рядках (Cases), то у кластери буде об’єднано досліджувані об’єкти. На Рис. 66

видно, що у даному випадку метод найближчого сусіда не дозволяє виділити окремі кластери – проявляється основний недолік даного методу: “ланцюговий ефект”, тобто утворення великого довгастого кластера, до якого один за одним додаються нові об’єкти. У такому разі кластеризацію слід повторити з іншими методами об’єднання.

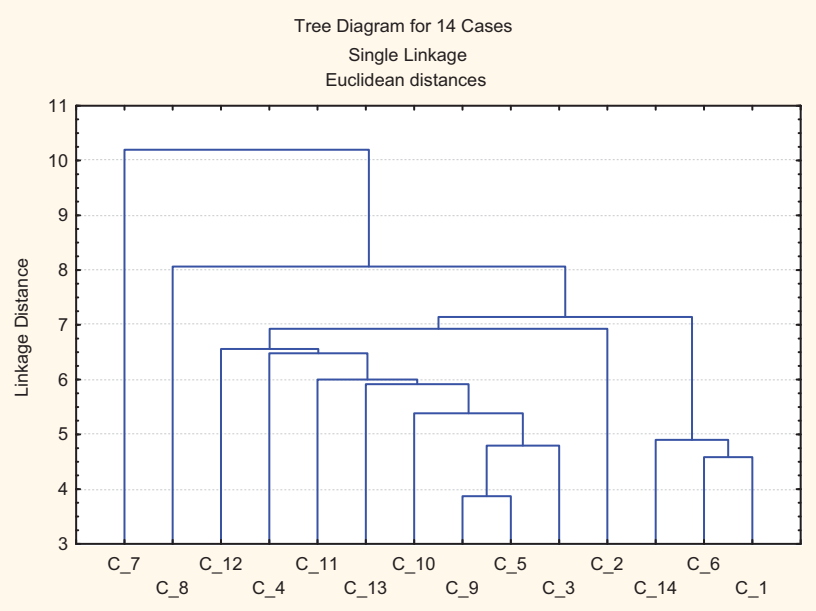


Рис. 67

Таблиця 52

linkage distance	Amalgamation Schedule (liri) Single Linkage Euclidean distances													
	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11	Obj. No. 12	Obj. No. 13	Obj. No. 14
3,872983	C_5	C_9												
4,582576	C_1	C_6												
4,795832	C_3	C_5	C_9											
4,898980	C_1	C_6	C_14											
5,385165	C_3	C_5	C_9	C_10										
5,916080	C_3	C_5	C_9	C_10	C_13									
6,000000	C_3	C_5	C_9	C_10	C_13	C_11								
6,480741	C_3	C_5	C_9	C_10	C_13	C_11	C_4							
6,557438	C_3	C_5	C_9	C_10	C_13	C_11	C_4	C_12						
6,928203	C_2	C_3	C_5	C_9	C_10	C_13	C_11	C_4	C_12					
7,141428	C_1	C_6	C_14	C_2	C_3	C_5	C_9	C_10	C_13	C_11	C_4	C_12		
8,062258	C_1	C_6	C_14	C_2	C_3	C_5	C_9	C_10	C_13	C_11	C_4	C_12	C_8	
10,19804	C_1	C_6	C_14	C_2	C_3	C_5	C_9	C_10	C_13	C_11	C_4	C_12	C_8	C_7

Таблиця 53

Case	Euclidean distances (liri)													
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14
C_1	0,0	17,9	13,9	12,2	13,0	4,6	16,1	18,6	14,3	13,1	13,6	9,5	14,5	4,9
C_2	17,9	0,0	11,6	14,9	9,9	18,0	15,7	11,4	8,6	7,9	6,9	11,8	9,7	13,9
C_3	13,9	11,6	0,0	8,8	5,5	14,4	17,3	17,4	4,8	8,2	12,0	9,5	7,2	12,3
C_4	12,2	14,9	8,8	0,0	6,5	13,4	17,5	18,7	8,3	10,5	12,2	8,8	9,7	10,8
C_5	13,0	9,9	5,5	6,5	0,0	13,8	14,2	13,9	3,9	5,4	8,1	6,8	8,2	10,3
C_6	4,6	18,0	14,4	13,4	13,8	0,0	18,0	19,4	14,9	14,2	14,2	11,0	14,4	5,6
C_7	16,1	15,7	17,3	17,5	14,2	18,0	0,0	10,2	16,0	12,4	11,4	12,1	19,3	14,0
C_8	18,6	11,4	17,4	18,7	13,9	19,4	10,2	0,0	15,0	11,9	8,1	13,5	17,7	14,8
C_9	14,3	8,6	4,8	8,3	3,9	14,9	16,0	15,0	0,0	5,8	8,8	9,3	5,9	11,7
C_10	13,1	7,9	8,2	10,5	5,4	14,2	12,4	11,9	5,8	0,0	6,0	6,6	10,2	10,3
C_11	13,6	6,9	12,0	12,2	8,1	14,2	11,4	8,1	8,8	6,0	0,0	8,2	11,3	9,4
C_12	9,5	11,8	9,5	8,8	6,8	11,0	12,1	13,5	9,3	6,6	8,2	0,0	11,7	7,1
C_13	14,5	9,7	7,2	9,7	8,2	14,4	19,3	17,7	5,9	10,2	11,3	11,7	0,0	11,8
C_14	4,9	13,9	12,3	10,8	10,3	5,6	14,0	14,8	11,7	10,3	9,4	7,1	11,8	0,0

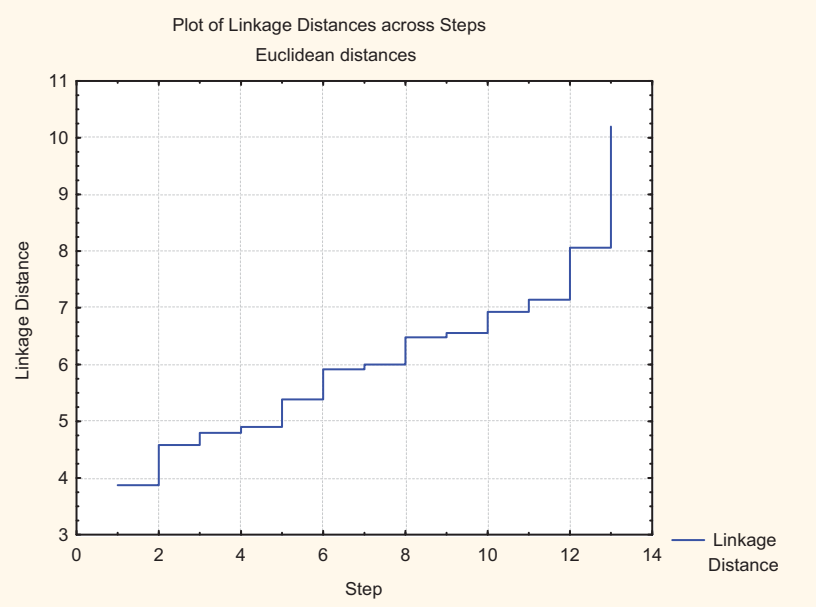


Рис. 68

Виконання ієрархічного кластерного аналізу у пакеті Statistica дозволяє отримати дендрограму (горизонтальну та вертикальну (Рис. 67) та таблицю злиття (Amalgamation Schedule (Таблиця 52)), як і у пакеті SPSS, а також додатково матрицю

відстаней (Distance matrix (Таблиця 53)), середнє та стандартне квадратичне відхилення для кожного об'єкта (Descriptive statistics), графік послідовності злиття (Graph of amalgamation schedule (Рис. 68)).

Контрольні запитання

1. Опишіть призначення кластерного аналізу. Чому кластерний аналіз називають методом "класифікації без навчання"?
2. Назвіть основні групи методів кластерного аналізу та етапи кластеризації.
3. Як графічно представляють результати кластерного аналізу?
4. Які основні кроки алгоритмів агломеративних методів кластеризаццїї?
5. Які основні кроки алгоритмів дивізимних методів кластеризації?
6. Які основні кроки алгоритмів методу k-середніх?
7. Які є способи обчислення відстані між об'єктами?
8. Які є критерії об'єднання у кластери?
9. Як визначається якість кластеризації?

Індивідуальні завдання

Індивідуальне завдання 1

Зібрати статистичний матеріал (20-40 вимірювань):

Варіант 1: оцінки студентів однієї групи з різних дисциплін (набір 1) та оцінки студентів різних груп з однієї дисципліни (набір 2).

Варіант 2: виміряти частоту пульса однорідної за віком та статтю групи людей у спокійному стані та після фізичного навантаження (10 присідань) – набір 1; виміряти чатоту пульса у спокійному стані у групах людей різних за віком (дві вікові групи) або статтю – набір 2.

Варіант 3: провести психологічне тестування однієї групи досліджуваних за двома методиками (набір 1); провести тестування за однією методикою для двох груп досліджуваних, які відрізняються за однією ознакою, наприклад, віком, статтю, рівнем освіти тощо.

Варіанрт 4: запропонувати інший спосіб збору даних та отримати набір 1 – повторні вимірювання на одній вибірці досліджуваних, набір 2 – вимірювання на двох незв'язаних вибірках.

Для набору 1 побудувати статистичні гіпотези, провести статистичний аналіз за допомогою відповідних параметричних та непараметричних методів, зробити висновки.

Побудувати стастистичні гіпотези, виконати статистичний аналіз відповідними методами для набору 2, зробити висновки.

Індивідуальне завдання 2

Знайдіть дослідницьку статтю у будь-якому журналі із суспільних або природничих наук. Виберіть статтю з цікавої для Вас теми. Дайте відповіді на запитання:

1) яка частина статті присвячена статистиці як такій (окремо від гіпотез, ідей, обговорення)?
2) чи проводилося дослідження на вибірці з деякої генеральної сукупності? Якою була генеральна сукупність? Який розмір вибірки? Яким чином відбиралися елементи (суб'єкти) або спостереження? Яким чином забезпечувалася репрезентативність вибірки? Чи можна узагальнити отримані результати на генеральну сукупність?

3) які змінні було використано? Які з них залежні, а які незалежні? Для кожної змінної визначіть шкалу вимірювання. З'ясуйте дискретність або неперервність змінних.
4) що можна сказати про вид розподілу досліджуваних змінних? Які дескриптивні характеристики використані у дослідженні?
5) які статистичні методи було використано? Спробуйте простежити за статистичним аналізом. Проаналізуйте, на які статистичні "підзадачі" розбито дослідження, чому дослідником обрані ті чи інші статистичні методи. Які ще методи можна було б застосувати у даній ситуації?
6) спробуйте з'ясувати, яка загальна гіпотеза дослідження? Які статистичні гіпотези використані для її перевірки? Які статистичні висновки отримано? Як їх інтерпретувати у термінах загального дослідження?

Зробіть висновки про коректність застосування використаних у дослідженні статистичних методів.

Індивідуальне завдання 3

З довідкових джерел вибрати економічні, географічні та інші числові показники для декількох країн.

За зібраними даними виконати ієрархічний кластерний аналіз:

а) для змінних – визначити групи найбільш близьких за смислом показників;

б) для елементів вибірки (країн) – визначити найбільш близькі за обраними показниками країни.

За результатами попереднього аналізу визначити задовільну кількість кластерів та виконати *k*-кластерний аналіз. Оцінити ефективність визначеної кількості кластерів.

Виконати дискримінантний аналіз, обравши за дискримінуючу змінну визначену у попередньому пункті приналежність до кластера. Зробити висновки про можливість та якість прогнозування приналежності країни до кластера.

Виконати факторний аналіз. Згідно критеріїв визначити задовільну кількість факторів. Дати їм якісну інтерпретацію.

Зробити висновки про особливості застосування використаних процедур аналізу даних, якість виконаного аналізу та особливості представлення даних та результатів.

Список використаних джерел

Підручники з математичної статистики та аналізу даних

1. Жалдак М.І., Кузьміна Н.М., Михалін Г.О. Теорія ймовірностей і математична статистика: Підручник для студентів фізико-математичних спеціальностей педагогічних університетів. – Вид. 2. – Полтава: “Довкілля-К”, 2009. – 500 с.
2. Жалдак М.І., Михалін Г.О. Елементи стохастики з комп’ютерною підтримкою. Посібник для вчителів. – Київ, 2006. – 120 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика. – Изд. 4-е. Учеб. пос. для вузов. – М.: “Высш. шк.”, 1972. – 368 с.
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. Учеб. пособие для вузов. – Узд. 2-е. – М.: “Высш. шк..”, 1975. –333 с.
5. Лакин Г.Ф. Биометрия: Учеб. пособие для биологич. спец. вузов. – 3-е изд., перераб. и доп. – М.: Высш. школа, 1980. – 293 с.
6. Чубукова И.А. Data Mining. http://www.intuit.ru/
7. Барсегян А.А. Технологии анализа данніх: Data Mining, Visual Mining, Text Mining, OLAP /А.А.Барсегян, М.С.Куприянов, В.В.Степаненко, И.И.Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.

Посібники з комп’ютерних статистичних пакетів

8. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере /Под ред. В.Э.Фигурнова. – 3-е изд., перераб. и доп. – М.: ИНФРА-М, 2003. – 544 с.
9. Наследов А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. – СПб.: Питер, 2005. – 416 с.
10. Бююль А., Цёфель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. – СПб.: ООО “ДиаСофтЮП”, 2002. – 608 с.

11. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. – 2-е изд. – СПб.: Питер, 2003. – 688 с.
12. Боровиков В. П. Популярное введение в программу STATISTICA. – М.: КомпьютерПресс, 1998. – 267 с.
13. Мамчич Т.І., Оленко А.Я., Осипчук М.М., Шпортюк В.Г. Статистичний аналіз даних з пакетом Statistica. Навчально-методичний посібник. – Дрогобич: Видавнича фірма "Відродження". – 2006. – 208 с.
14. Гуржій А.М., Дудар З.В., Левикін В.М., Шамша Б.В. Математичне забезпечення інформаційно-керуючих систем: Підручник для студентів вищих навчальних закладів. – Харків: ТОВ "Компанія СМІТ", 2006. – 448 с.
15. Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. Статистические функции MS Excel в экономико-статистических расчетах: Учеб. пособие для вузов /Под ред. проф. В.С.мхитаряна. – М.: ЮНИТИ-ДАНА, 2003. – 231 с.
16. Левин Д., Стефан Д., Кребиль Т., Беренсон М. Статистика для менеджеров с использованием Microsoft Excel. – 4-е изд. – М.: Изд. дом "Вильямс", 2004. – 1312 с.

Статиститичні методи у психології та соціології

17. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. – М.: 1976.
18. Суходольский Г.В. Основы математической статистики для психологов. – Л.: Изд-во ЛГУ, 1972. – 430 с.
19. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: ООО "Речь", 2001. – 350 с.
20. Артемьева Е.Ю., Мартынов Е.М. Вероятностные методы в психологии. – М., 1975.
21. Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. – СПб.: Речь, 2006. – 400 с.
22. Хили Дж. Социологические и маркетинговые исследования. 6-е изд. /Пер. с англ. Под общей ред. к. ф.-м.н. А.А.Руденко. – Киев: ООО "ДиаСофтЮП"; СПб.: Питер, 2005. – 638 с.

Окремі питання статистичного аналізу даних

23. Гусев А.Н. Дисперсионный анализ в экспериментальной психологии: Учеб. пособие. – М.: Учебно-методический коллектор “Психология”, 2000. – 136 с.
24. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.
25. Пустыльник Е.И. Статистические методы анализа и обработки наблюдений. – М.: Наука, 1968. – 288 с.

Додаткова література (вправи та завдання)

26. Колде Я.К. практикум по теории вероятностей и математической статистике: Учеб. пособие для техникумов. – М.: Высшая школа, 1991. – 157 с.
27. Турчин В.М. Математична статистика в прикладах і задачах: Навч. Посібник: У 2 ч. – Дніпропетровськ: ДДУ, 1998. – Ч.2. – 228 с.

Електронні ресурси

28. Р 50.1.033–2001. РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I:. Критерии типа хи-квадрат. Издание официальное. ГОССТАНДАРТ РОССИИ. Москва – 2002. <http://ami.nstu.ru/~headrd/seminar/xi_square/start1.htm>
29. Р 50.1.037–2002. РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II: Непараметрические критерии. Издание официальное. ГОССТАНДАРТ РОССИИ. Москва – 2002. <http://ami.nstu.ru/~headrd/seminar/nonparametric/index.htm>
30. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. <http://www.machinelearning.ru/wiki/index.php>
31. SPSS (Statistical Package for Social Science). <http://www.spss.com.ua/>
32. STATISTICA <http://www.statsoft.com/>

Глосарій позначень статистичних функцій:

MS Excel	*OpenOffice.org Calc*	*SPSS*	*Матем.*	*Опис*
До теми 1				
СУММ	SUM	Sum	Σ	Сума
СЧЕТ	COUNT	Number of Cases	n	Кількість випадків (категорій)
МИН	MIN	Minimum		Мінімум
МАКС	MAX	Maximum		Максимум
		Range	R	Діапазон
		Interquartile Range		Міжквартильний розмах – від 25% до 75% вибірки
МЕДИАНА	MEDIAN	Median	M_e	Медіана
МОДА	MODE		M_o	Мода
		Grouped Median		Згрупована медіана
СРЗНАЧ	AVERAGE	Mean	$\overline{x}$	Середнє
		Std. Error of Mean	S_x	Стандартна похибка обчислення середнього
ДОВЕРИТ	CONFIDENCE	Confidence Interval for Mean		Довірчий інтервал для середнього

MS Excel	***OpenOffice.org Calc***	***SPSS***	***Матем.***	***Опис***
СРГАРМ	HARMEAN	Harmonic Mean	$\overline{x}_h$	Середнє гармонійне
СРГЕОМ	GEOMEAN	Geometric Mean	$\overline{x}_g$	Середнє геометричне
УРЕЗСРЕДНЕЕ	TRIMMEAN	Trimmed Mean		Середнє арифметичне без крайніх екстремальних значень (урізане на вказаний відсток від обсягу вибірки)
ДИСП	VAR	Variance	S^2	Дисперсія
СТАНДОТКЛОН	STDEV	StandartDeviation	S	Стандартне квадратичне відхилення
ЭКСЦЕСС	KURT	Kurtosis	E	Ексцес
		Std. Error of Kurtosis	m_e	Стандартна похибка обчислення ексцеса
СКОС	SKEW	Skewness	A	Асиметрія
		Std. Error of Skewness	m_a	Стандартна похибка обчислення асиметрії
		Cumulative percent		Відносна накопичена частота

MS Excel	*OpenOffice.org Calc*	*SPSS*	*Матем.*	*Опис*
		Percent		Відносна частота (обчислюється за кількістю записів у файлі
		Valid percent		Відносна частота (обчислюється за кількістю правильних або валідних, не пропущених даних)
ЧАСТОТА	FREQUENCY[23]			Побудова інтервального статистичного ряду розподілу частот
РАНГ	RANK			Визначення позиції (рангу) числа у масиві значень
ПРОЦЕНТРАНГ	PERCENTRANK			Оцінка відносного положення деякого значення у наборі даних

[23] Порядок роботи з деякими функціями в OpenOffice.org Calc відрізняється від порядку роботи з відповідними функціями MS Excel.

MS Excel	*OpenOffice.org Calc*	*SPSS*	*Матем.*	*Опис*
ПЕРСЕНТИЛЬ	PERCENTILE			Виявлення значення з набору даних за його відносним положенням
		Between Groups		Між групами
		Within Groups		Всередині груп
		Total		Загальний (сумарний)
До теми 2				
ТТЕСТ	TTEST			Повертає імовірність, що відповідає критерію Стьюдента
СТЬЮДРАСПОБР	TINV			Обчислює обернений розподіл Стьюдента
СТЬЮДРАСП	TDIST			Обчислює значення t-розподілу Стьюдента
ФТЕСТ	FTEST			Повертає результат F-теста
FРАСП	FDIST0			Обчислює значення F-розподілу
FРАСПОБР	FINV			Обчислює обернений F-розподіл

MS Excel	*OpenOffice.org Calc*	*SPSS*	*Матем.*	*Опис*
До теми 4				
ПИРСОН	PEARSON	Pearson Correlation		Коефіцієнт кореляції Пірсона
КОРРЕЛ	CORREL			Коефіцієнт кореляції Пірсона
		Spearman’s rho		Коефіцієнт кореляцій Спірмена
		Partitial Correlation Coefficient		Часткова кореляція
		Eta, Eta Squared		ета, ета-квадрат
ФИШЕР	FISHER			перетворення Фішера
ЛИНЕЙН	LINEST			
До теми 5				
НОРМАЛИЗАЦИЯ	STANDARDIZE			
НОРМРАСП	NORMDIST			
НОРМСТРАСП	NORMSDIST			
НОРМСТОБР	NORMSINV			
ЛОГНОРМРАСП	LOGNORMDIST			

MS Excel	*OpenOffice.org Calc*	*SPSS*	*Матем.*	*Опис*
ВЕЙБУЛЛ	WEIBULL			
ЭКСПРАСП	EXPONDIST			
ГАММАРАСП	GAMMADIST			
БЕТАРАСП	ETADIST			
БИНОМРАСП	BINOMDIST			
ГИПЕРГЕОМЕТ	HYPGEOMDIST			
ПУАССОН	POISSON			
ХИ2ТЕСТ	CHITEST			
ХИ2ОБР	CHIINV			
СЛЧИС	RAND			
СЛУЧМЕЖДУ	RANDBETWEEN			
ОКРУГЛ	ROUND			

Додаток А. Вибірки

Варіант 0

Вибірка А0

2	4	2	4	3	3	3	2	0	6	1	2	3	2	2	4	3	3	5	1
0	2	4	3	2	2	3	3	1	3	3	3	1	1	2	3	1	4	3	1
7	4	3	4	2	3	2	3	3	1	4	3	1	4	5	3	4	2	4	5
3	6	4	1	3	2	4	1	3	1	0	0	4	6	4	7	4	1	3	

N=79 Початок першого інтервалу: 0 Довжина інтервалу: 1

Вибірка В0

65	71	67	73	68	68	72	68	67	70	78	74	79	65	72
65	71	70	69	69	76	71	63	77	75	70	74	65	71	68
74	69	69	66	71	69	73	74	80	69	73	76	69	69	67
67	74	68	74	60	70	66	70	68	64	75	78	71	70	69
73	75	74	72	80	72	69	69	71	70	73	65	66	67	69
71	70	72	76	72	73	64	74	71	76	68	69	75	76	73
74	78	66	75	72	69	68	63	70	70	78	76	73	73	67
71	66	66	72	69	71	71	68	72	69	73	73	66	72	73
70	69	74	72	69	74	70	74	72	76	71	66	62	69	74
76	74	69	64	75	71	76	68	68	78	71	71	68	67	74
68	81	72	68	72	71	71	71	69	61	74	66	70	72	65
67	73	78	73	71	75	73	71	72	68	67	69	69	77	63
71	74	67	68	69	74	69	67	74	66	74	74	69	75	70
73	63	77	74	75										

N= 200 Початок першого інтервалу: 59 Довжина інтервалу: 2

Вибірка С0

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
73	-291	577	57	-219	454	61	-241	480	68	-264	538
69	-270	548	71	-281	566	62	-243	486	62	-240	495
72	-279	575	66	-262	522	63	-245	500	70	-277	554
72	-282	573	76	-302	599	71	-282	560	70	-279	552
65	-254	519	70	-275	554	65	-252	517	65	-253	511
67	-264	530	68	-267	542	70	-276	558	70	-275	555
56	-216	443	74	-290	588	70	-276	559	63	-248	496
70	-276	555	68	-266	537	63	-246	496	63	-243	495
63	-248	502	69	-270	550	73	-284	580	67	-264	530
64	-253	506	71	-283	559	68	-271	542	68	-267	534
70	-276	554	60	-237	479	59	-227	462	55	-213	437
67	-262	535	56	-222	446	64	-256	504	56	-218	446
60	-234	478	71	-281	565	79	-309	627	58	-223	460
63	-243	495	68	-269	538	77	-300	607	70	-278	551
80	-313	635	66	-257	520	78	-310	618	59	-236	465

71	-278	564	60	-235	478	66	-255	521	68	-263	543
74	-292	583	70	-275	554	63	-252	497	69	-268	550
68	-271	534	69	-276	542	69	-274	546	63	-243	497
65	-256	518	72	-282	566	74	-291	582	70	-271	558
73	-291	574	70	-277	558						

	X	Y	Z
Початок першого інтервалу	53	-321	420
Довжина інтервалу		17	34

Вибірка D0

N0	F1	F2	F3	F4
1	-19	-31	-35	-31
2	-28	-33	-32	-27
3	-39	-35	-26	-28
4	-36	-25	-35	-35
5	-44	-28	-30	-40
6	-39	-31	-17	-31

Вибірка E0

N0	F1	F2	F3	F4	F5	F6	F7
1	75	104	96	92	76	92	89
2	86	89	88	89	89	87	85
3		92	105		90	88	93
4		90	90		77	82	
5		81	91		75	90	
6						86	

Варіант 1

Вибірка A1

0	4	2	0	5	1	1	3	0	2	2	4	3	2	3	3	0	4	5	1
3	1	5	2	0	2	2	3	2	2	2	6	2	1	3	1	3	1	5	4
5	5	3	2	2	0	2	1	1	3	2	3	5	3	5	2	5	2	1	1
2	3	4	3	2	3	2	4	2											

N=69 Початок першого інтервалу:0 Довжина інтервалу: 1

Вибірка C1

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
67	-201	602	82	-237	730	78	-226	697	90	-269	801
68	-199	602	70	-209	626	75	-218	672	95	-279	847
70	-206	625	83	-243	737	82	-245	729	88	-259	782
76	-221	674	80	-239	716	85	-247	763	86	-251	773
80	-238	718	76	-221	681	68	-201	608	71	-207	635
87	-256	781	81	-238	721	72	-215	643	87	-257	777
75	-222	668	80	-238	710	71	-209	637	77	-227	683
79	-230	702	76	-223	680	72	-212	638	73	-214	647
79	-234	701	70	-207	626	68	-203	611	82	-243	736
73	-217	648	79	-237	708	86	-252	773	74	-214	659
86	-253	773	74	-216	658	85	-251	764	67	-196	599
78	-228	692	77	-228	690	71	-206	632	82	-239	735
79	-230	708	65	-193	579	72	-214	641	72	-210	640
67	-201	596	80	-234	714	76	-227	675	74	-221	664
79	-237	702	79	-229	704						

N=58	X	Y	Z
Початок першого інтервалу	62	-286	557
Довжина інтервалу	6	15	45

Вибірка D1

N0	F1	F2	F3	F4	F5	F6
1	51	56	59	59	52	51
2	57	56		56	58	56
3	55	54		54	53	55
4	52	55		52	55	
5	51			54	55	
6	54			51	52	
7				56	56	

Вибірка E1

N0	F1	F2	F3	F4	F5	F6
1	52	36	43	52	37	45
2	49	42	51	42	45	36
3	45	48	44	40	55	46
4	44	37	47	38	47	36
5	34	37	34	37	39	37

Вибірка B1

135	133	124	132	104	152	134	130	129	120	122	124
117	123	123	129	121	122	125	131	147	124	137	112
126	128	111	129	115	147	131	132	137	119	125	120
129	125	123	127	132	118	133	132	132	134	131	120
135	132	125	132	108	114	121	133	133	135	131	125
114	115	122	131	125	132	120	126	115	117	118	118
132	134	127	127	124	135	128	127	115	144	129	120
137	127	125	116	132	120	117	127	118	109	127	122
120	135	116	118	133	136	125	126	119	126	129	127
129	124	127	132	126	131	127	130	126	124	135	127
124	123	123	130	132	143	122	139	120	134	108	132
121	111	123	140	137	120	125	131	118	120	120	136
129	127	116	138	128	133	122	131	128	140	138	134
120	126	109	137	111	115	117	130	113	126	115	124
125	118	115	128	123	129	128	120	115	134	118	135
134											

N=181 Початок першого інтервалу: 102. Довжина – 4.

Варіант 2

Вибірка A2

3	7	4	6	1	4	2	4	6	5	3	2	9	0	5	6	7	7	3	1	5	5	7	6	6	1	6	7	7
5	5	4	2	6	2	1	5	3	3	1	5	6	4	4	3	4	1	5	5	3	4	3	7	4	5	3	5	4
4	3	6	7	5	2	4	6	4	3	6																		

N=66 Початок першого інтервалу:0. Довжина інтервалу:1

Вибірка B2

95	96	103	89	72	105	85	85	91	101	82	91	91	87	103	101	71	105
80	85	91	87	101	94	98	85	82	94	86	72	97	95	84	79	94	87
89	83	100	86	85	95	95	83	87	92	92	79	92	79	104	79	93	85

93	88	77	92	92	103	85	90	83	86	104	104	104	90	89	84	81	86
85	85	80	95	91	93	70	83	93	95	95	78	96	102	100	86	86	90
111	95	94	84	64	87	85	87	87	81	82	97	81	100	95	90	97	100
101	86	89	80	88	85	93	79	95	90	107	93	88	95	85	78	74	76
96	83	88	91	95	94	88	80	96	93	77	71	78	92	77	66	90	100
88	97	90	86	93	91	98	95	83	84	91	99	75	88	96	89	100	80
109	80	95	87	89	85	87	72	77	90	97	87	100	96	84	82	92	81
95	91	88	91	81	88	78	75	80	97	95	83	87	85	83	98	97	94
99	81	88	87	90	85	74	88	97	82	73	81	84	88	94			

N=213 Початок першого інтервалу:62 Довжина інтервалу:4

Вибірка C2

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
48	99	520	61	128	664	52	112	564	50	105	542
40	83	435	42	88	456	32	67	345	49	99	538
52	106	564	55	117	601	48	97	518	54	113	588
50	107	541	52	110	567	40	82	430	45	99	485
39	79	424	44	94	475	40	81	436	36	75	387
47	100	516	41	82	450	46	92	504	41	91	441
38	80	413	47	97	516	41	90	450	46	101	503
46	96	505	43	91	468	54	110	584	38	83	483
44	97	479	43	87	472	53	110	498	57	120	624
52	107	569	40	86	432	56	112	608	53	107	579
45	92	490	49	104	538	47	97	509	47	99	512
44	90	483	42	89	459	50	102	542	44	93	474
53	108	577	31	71	333	46	101	498	57	120	624
52	107	569	40	86	432	56	112	608	53	107	579
45	96	493	47	97	516	42	93	461	51	108	554
42	86	461	43	89	471	41	84	445	48	100	519
45	98	486	48	101	527	55	112	595	46	92	496
45	97	492	44	93	483	40	80	431	43	89	467

N=72	X	Y	Z
Початок першого інтервалу	28	62	305
Довжина інтервалу	6	11	56

Вибірка D2

N0	F1	F2	F3	F4
1	48	45	41	49
2	33	41	41	46
3	53	49	34	41
4	43	42	38	41
5	38	43	50	47
6	47	41	45	47

Вибірка E2

N0	F1	F2	F3	F4	F5	F6
1	89	92	91	107	103	113
2	106	105	101	99	104	109
3	88	94		76	74	89
4	102	92		101	97	99
5	107	100		101	103	92
6	89			96		100
7	101					

Варіант 3

Вибірка A3

0	0	2	0	1	3	0	1	0	1	2	1	3	0	0	2	1	3	2	2	3	0	2	0	2	0
1	3	3	2	0	2	4	3	2	1	2	2	2	2	3	3	1	1	1	3	2	2	4	2	1	4
2	1	0	1	2	1	4	4	2	3	3	5	5	2	1	2	3	2	3	1	1	0	4	1	1	0
1	0	3	1																						

N=82 Початок першого інтервалу: 0 Довжина інтервалу: 1

Вибірка D3

N0	F1	F2	F3	F4	F5
1	114	106	110	88	99
2	96	107	92	94	100
3	113	99	113	100	114
4	98	95	110	99	95

Вибірка E3

N0	F1	F2	F3	F4	F5	F6	F7
1	59	60	53	51	60	59	46
2		56	69	64		59	61
3		52	67	56		49	65
4			59	54		62	57
5			61	66		60	47
6			62	67		69	

Вибірка B3

-29	-22	-16	-20	-16	-18	-28	-20	-32	-22	-23	-26	-10	-25	-25
-29	-29	-19	-12	-26	-18	-20	-9	-24	-20	-19	-26	-23	-11	-26
-30	-23	-30	-18	-20	-13	-17	-24	-28	-26	-21	-21	-26	-24	-36
-23	-24	-25	-20	-23	-17	-11	-22	-19	-19	-25	-29	-23	-16	-25
-15	-18	-17	-19	-21	-12	-24	-30	-33	-22	-15	-18	-26	-22	-19
-25	-23	-21	-22	-22	-25	-16	-25	-19	-17	-30	-13	-25	-19	-24
-17	-24	-16	-23	-15	-22	-22	-19	-20	-19	-33	-14	-17	-21	-16
-24	-13	-20	-19	-17	-13	-27	-25	-25	-19	-22	-22	-22	-23	-9
-11	-22	-24	-18	-19	-18	-31	-16	-18	-24	-14	-23	-26	-25	-19
-23	-24	-21	-26	-25	-18	-16	-30	-16	-24	-13	-14	-18	-22	-22
-28	-18	-21	-27	-31	-23	-23	-27	-21	-21	-22	-34	-24	-20	-24
-21	-32	-16	-18	-22	-22	-15	-15	-22	-18	-	-	-	-	-

N=175 Початок першого інтервалу:-37 Довжина інтервалу: 2

Вибірка C 3

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
46	50	411	59	59	524	56	62	500	58	62	513
55	57	491	57	60	511	52	58	464	57	57	510
57	61	508	48	52	422	42	46	373	46	49	406
55	58	491	41	47	367	48	50	423	54	63	476
51	51	454	57	57	507	60	63	539	67	72	602
62	70	552	50	57	449	62	64	554	60	64	538
43	43	377	64	67	567	46	47	413	53	53	467
64	71	567	43	45	381	51	59	450	57	69	507
56	64	496	57	62	507	69	76	620	41	45	362
65	67	580	54	56	478	60	60	531	58	61	513
56	63	502	50	51	444	57	61	511	43	45	384

51	58	458	59	66	529	62	65	551	55	57	488
58	60	516	48	51	427	58	66	518	40	43	358
42	47	374	45	54	404	54	60	483	67	74	601
46	54	405	51	53	450	57	63	506	57	57	511
54	60	485	40	41	358	44	45	392	48	54	425
62	67	554	59	65	530	55	62	494	56	57	500
57	58	512	46	54	404	50	54	445	57	65	506
68	68	610	47	52	417	63	67	566	73	79	652
47	56	421	55	59	486	44	48	391	54	56	485
69	74	613	49	55	437	51	60	456	57	66	508
65	72	578	64	70	572	47	55	416	63	67	557

N=88	X	Y	Z
Початок першого інтервалу	37	38	333
Довжина інтервалу	6	7	50

Варіант 4

Вибірка А4

3	3	1	0	0	3	3	5	3	0	0	4	1	5	1	6	5	4	7	4
3	3	3	0	2	3	1	4	1	2	4	3	4	5	4	0	5	6	6	3
5	4	1	3	3	6	3	1	1	5	2	3	5	3	3	4	1	5	6	1
3	3	3	5	6	1	2	1	3	4										

N=70 Початок першого інтервалу:0 Довжина інтервалу:1

Вибірка В4

58	78	84	62	63	100	55	90	102	70	66	89	54	62	82	103	91	119
71	92	71	93	83	42	110	94	56	96	95	87	84	82	56	78	80	75
88	102	104	88	64	96	92	67	78	95	71	105	94	92	89	109	69	103
50	66	73	76	100	72	86	46	102	95	98	84	85	76	85	84	68	74
82	46	60	94	109	93	79	74	62	97	94	91	70	106	68	81	61	109
81	71	89	78	85	82	93	64	65	190	89	55	96	80	77	96	67	110
78	98	108	68	65	75	82	70	84	73	65	79	93	68	65	50	88	72
99	81	92	76	82	100	75	45	110	81	84	68	66	78	64	92	75	101
77	90	103	66	57	84	100	83	68	69	68	81	75	72	69	75	61	53
83	69	90	99	69	73	84	70	80	117	76	104	78	87	57	57	63	102
78	114	79	70	56	93	73	71	77	98	86	82	100	73				

N=194Початок першого інтервалу:39 Довжина інтервалу:6

Вибірка С4

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
67	207	529	75	225	597	62	195	486	68	208	538
57	171	449	68	207	542	71	220	566	71	220	562
82	246	648	65	201	518	67	205	531	76	231	601
62	191	487	80	241	631	66	207	520	77	234	610
57	172	446	71	219	565	67	201	534	73	225	580

83	255	656	61	183	483	70	215	558	65	198	516
69	213	547	77	239	612	64	197	510	70	217	557
86	261	681	69	209	546	66	206	523	74	225	587
66	203	519	64	201	505	78	241	622	73	224	580
79	242	629	68	208	539	65	197	511	77	237	612
82	248	646	52	159	413	63	196	501	7	236	612
73	221	574	59	183	470	69	216	549	49	154	388
65	195	516	63	197	496	59	180	471	63	189	499
55	172	437	56	176	447	71	214	558	71	222	565
61	186	482	62	192	486	59	185	466	69	214	549
64	197	510	67	204	535	64	196	505	52	160	409
59	182	465									

N=65

	X	Y	Z
Початок першого інтервалу	46	145	363
Довжина інтервалу	7	19	50

Вибірка D4

N0	F1	F2	F3	F4	F5
1	-20	-8	-18	-26	-29
2	-28	-7	-12	-23	-8
3	-4	-19	-18	-16	-21
4	-11	-26	-20	-13	-21
5	-20	-11	-21	-22	-8
6	-26	-11	-19	-28	-28
7	-24	-15	-18	-18	-17

Вибірка E4

N0	F1	F2	F3	F4	F5	F6	F7
1	70	67	67	81	82	66	70
2	74	70	71	69	82	77	67
3	59	66		73		62	68
4				77		73	76
5				71		74	
6				87		54	
7						67	

Варіант 5

Вибірка A5

0	2	0	1	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	1	2	0	0
1	0	1	0	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	1	0	2	0	0	0	1	1	1	1	1	0	2	1	0	1	0

N=81 Початок першого інтервалу:0 Довжина інтервалу:1

Вибірка B5

34	14	-14	10	9	29	27	-1	-4	17	23	13	18	-17	-22	9	18	-8	25
1	8	-9	3	11	6	26	6	8	16	19	22	-8	23	-5	21	33	-2	6
17	-21	-20	-17	16	3	6	25	0	4	5	6	-21	-2	8	-16	-22	20	18
-6	11	3	-2	17	13	8	27	11	9	12	12	-1	25	4	34	-16	20	-8
19	-8	29	0	-13	0	9	26	19	29	9	22	30	13	19	6	-7	3	2
-1	-10	20	-7	21	10	8	-5	-2	9	-10	1	12	8	35	36	-13	32	19
11	15	13	2	-5	-12	11	9	34	9	-2	-20	-4	-2	19	-16	0	14	25
31	31	-11	-7	23	-20	-2	-12	-3	13	-7	15	8	-9	19	7	19	1	18

-8	-12	8	30	-22	18	-9	-20	17	28	26	6	-7	0	-9	-7	13	25	-8
7	11	20	23	12	19	52	18	32	29	33	3	-8	5	-4	-9	5	24	2
4	24	30	21	7	27	12	1	-1	-5	6	-5	21	9	17	4	-14	5	25
19	5	12	-2	-4	2	20	14	14	27	16	-6	8	-2	-3	-2	18	28	-9
-2																		

N=229 Початок першого інтервалу:-25 Довжина інтервалу: 6

Вибірка C5

X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
31	318	-37	23	231	-32	28	284	-34	31	316	-38
28	280	-36	28	288	-30	26	265	-29	25	251	-30
30	305	-32	26	269	-35	29	298	-38	27	278	-31
23	234	-30	26	267	-34	26	264	-30	26	268	-32
25	255	-32	28	284	-29	23	232	-33	27	273	-31
25	258	-33	27	277	-35	27	272	-30	28	280	-38
25	258	-27	25	253	-34	29	292	-39	24	244	-32
27	272	-28	26	268	-30	24	245	-25	24	243	-32
31	317	-40	27	270	-35	27	274	-32	26	269	-28
25	252	-33	28	289	-30	25	256	-35	25	255	-34
28	280	-31	27	278	-35	29	291	-30	22	229	-29
25	258	-31	29	295	-36	29	290	-34	27	272	-35
30	304	-39	24	246	-33	25	257	-33	26	263	-32
28	289	-30	23	238	-24	25	258	-31	24	249	-30
28	288	-34	27	279	-29	30	309	-34	25	252	-29
31	317	-37	27	273	-32	25	257	-39	22	229	-31

N=64	X	Y	Z
Початок першого інтервалу	21	221	-41
Довжина інтервалу	2	16	3

Вибірка D5

N0	F1	F2	F3	F4
1	46	45	49	45
2	49	43	43	46
3	48	43	47	46
4	47	42	43	44
5	45	48	46	43
6	48	44	45	45

Вибірка E5

N0	F1	F2	F3	F4	F5	F6	F7
1	123	125	113	115	122	127	109
2	117	113		108	123	104	
3	115	98		116	131	111	
4	97	103		123	116	103	
5	108	95		124	102		
6				132	98		
7				113	105		

Додаток Б. Варіанти завдань до лабораторних робіт

Завдання до теми 2

Варіанти для парного t-тесту (завдання 2):

Варіант 1

X	25	30	28	50	20	40	32	36	42	38
Y	28	31	26	52	24	36	33	35	45	40

Варіант 2

X	76	71	57	49	70	69	26	65	59
Y	81	85	52	52	70	63	33	83	62

Варіант 3

X	15	20	16	22	24	14	18	20
Y	15	22	14	25	29	16	20	24

Варіант 4

X	0,18	0,12	0,12	0,08	0,08	0,12	0,19	0,32	0,27	0,22
Y	0,16	0,09	0,08	0,05	0,13	0,10	0,14	0,30	0,31	0,24

Варіант 5

X	111	104	107	90	115	107	106	107	95	116	127
Y	113	107	123	122	117	112	105	108	111	114	102

Варіант 6

X	81	80	73	72	72	69	69	65	65	62	62	60	54
Y	70	66	66	63	63	61	60	54	47	43	41	40	39

Варіант 7

X	76,10	76,20	76,00	76,04	76,10	76,08	76,18	76,02
Y	76,20	76,00	76,25	76,02	76,18	76,06	76,04	76,25

Варіант 8

X	7,52	8,18	2,02	4,46	1,95	9,47	6,79	6,45	1,50	9,91
Y	0,75	7,94	4,82	4,80	2,36	7,68	0,23	4,15	3,51	1,70

Варіант 9

X	7,0	7,1	7,3	7,2	7,6	7,7	7,4
Y	7,7	8,2	7,5	8,1	7,5	7,9	7,5

Варіант 10

X	7,7	9,0	9,4	7,4	7,4	10,9	8,0
Y	8,26	7,22	8,43	5,57	6,35	8,00	9,13

Варіант 11

X	31,6	24,2	24,8	29,1	29,9	31,0
Y	31,1	24,0	24,6	28,6	29,1	30,1

Варіант 12

X	7	10	12	12	14	15
Y	11	12	13	15	16	18

Варіант 13

X	420	470	490	530	560	580	580	600
Y	561	580	621	630	640	680	692	700

Варіант 14

X	3770	3817	2450	3463	3500	5544	3112	3150	3118
Y	2991	4593	3529	4274	3103	3949	3491	3559	2916

Варіант 15

X	7,8	1,4	28,0	2,5	18,6	11,0	9,8	11,3
Y	12,7	6,5	25,0	6,7	20,0	18,7	11,2	14,0

Варіант 16

X	10	13	28	12	17	7	15	15	14	16	9	8
Y	5	15	22	11	8	11	13	10	12	7	16	11

Варіант 17

X	37,5	38,2	24,5	34,6	35,0	55,4	31,1	31,5	31,2	30,2
Y	35,6	45,8	29,2	34,9	35,3	45,9	31,0	45,1	41,4	29,9

Завдання до теми 3

Варіанти для однофакторного аналізу (завдання 1, 2):

Крім наведених нижче варіантів завдань із соціологічним сюжетом пропонується також виконати однофакторний аналіз для вибірок D та E (див. Вибірки). Сюжет для них придумати самостійно.

Варіант 1 [22]:

Серед населення деякої місцевості випадкову вибірку з 18 подружніх пар оцінювали за двома шкалами.

а) відповідальність за прийняття рішень:			б) щастя у шлюбі:		
Традиційні	Двокар'єрні	Громадянські	Традиційні	Двокар'єрні	Громадянські
7	8	2	10	12	12
8	5	1	14	12	14
2	4	3	20	12	15
5	4	4	22	14	17
7	5	1	23	15	18
6	5	2	24	20	22

За першою шкалою вимірювали ступінь розподілу влади та відповідальності за прийняття рішень у сім'ї: низькі бали відповідають рівномірному розподілу влади між подружжям;

високі – монополізації з одного боку. За другою шкалою вимірювали щастя у шлюбі: низькі бали – низький рівень незадоволеності. Подружні пари було класифіковано за типом стосунків: традиційні (працює лише чоловік), двокар'єрні (обидва працюють), громадянські (живуть разом, але шлюб офіційно не зареєстровано, незалежно від того, хто працює). Чи достовірно змінюється розподіл влади або щастя у шлюбі залежно від типу стосунків?

Варіант 2 [8, с. 226]: Визначити, чи впливає вага браслета на частоту тремору руки (Гц).

Вага браслета (фунт)	0	1,25	2,5	5	7,5
Досліджуваний \ номер досліду	1	2	3	4	5
д1	3,01	2,85	2,62	2,63	2,58
д2	3,47	3,43	3,15	2,83	2,70
д3	3,35	3,14	3,02	2,71	2,78
д4	3,10	2,86	2,58	2,49	2,36
д5	3,41	3,32	3,08	2,96	2,67
д6	3,07	3,06	2,85	2,50	2,43

Варіант 3 [22]:

Студентів біологічного факультету навчали за різними методиками: першу групу – традиційно "лекції + лабораторні", другу – за методом "тільки лабораторні + демонстрації, без лекцій", третю – за методом "відеозаписи лекцій + демонстрації". Студентів випадковим чином розподілили по групах, а у кінці семестру з кожної групи випадковим чином відібрали оцінки 9-ох студентів для того, щоб з'ясувати, чи існує різниця у знаннях (оцінках) студентів залежно від методу навчання.

Лекції	Демонстрації	Відеозаписи
55	56	50
57	60	52
60	62	60
63	67	61
72	70	63
73	71	69
79	82	71
85	88	80
92	95	82

Варіант 4 [22]: 16 осіб – представників різних релігій опитували за “Шкалою підтримки смертної кари”.

Результати опитування наведено у таблиці. З’ясувати, чи відрізняється ставлення до смертної кари у представників різних релігій.

Протестанти	8	12	13	17	50
Католики	12	20	25	27	84
Іудеї	12	13	18	21	64
Атеїсти	15	16	23	28	82

Варіант 5 [22]:

Оцінювали ефективність роботи трьох соціальних служб (агенцій). Для цього дослідник зібрав дані про кількості днів, необхідних для оформлення документів. Для кожної агенції було вибрано 10 спостережень. З’ясувати, чи існує значима різниця між трьома агенціями у швидкості обробки документів.

Клієнти	Агенція A	Агенція B	Агенція C
1	5	12	9
2	7	10	8
3	8	19	12
4	10	20	15
5	4	12	20
6	9	11	21
7	6	13	20
8	9	14	19
9	6	10	15
10	6	9	11

Варіант 6 [22]:

Чи втрачають громадяни похилого віку інтерес до політики та місцевого життя? Для 4-ох випадкових вибірок респондентів – з 4-ох вікових груп відповідно, – було проведене опитування стосовно обізнаності у питаннях з останнього випуску новин. У таблиці наведено дані про кількість правильних відповідей. Чи значимі відмінності між різними віковими групами?

Школярі (15-18 років)	0	1	1	2	2	2	3	5	5	7	7	8
Молодь (21-30 років)	0	0	2	2	4	4	4	6	7	7	7	10
Люди середнього віку (40-55 років)	2	3	3	4	4	5	6	7	7	8	8	10
Пенсіонери (65 років і старші)	5	6	6	6	7	7	8	10	10	10	10	10

Варіант 7 [22]: Люди якого типу біше за інших залучені до громадських організацій?

15 випадковим респондентам було задано питання: “Членом скількох громадських організацій Ви є?” З’ясувати, які відмінності є значущими.

а) членство залежно від освіти:	б) членство залежно від захоплення переглядом телепередач:

Менше ніж школа	Школа	Коледж	мала або відсутня	Середня	Висока
0	1	0	0	3	4
1	3	3	0	3	4
2	3	4	1	3	4
3	4	4	1	3	4
4	5	4	2	4	5

Варіант 8 [22]:

У місті Таун було запроваджено дві програми, спрямовані на зниження злочинності. Причому в деяких районах реалізовували першу програму (спостереження за сусідами), в інших – другу (пішохідне патрулювання району дільничим поліцейським). А в деяких районах – жодну з програм реалізовано не було.

З'ясувати, чи були вказані програми успішними. У таблиці наведено дані показників зменшення кількості злочинів за рік у відсотках з 18 районів міста (тобто 3 випадкові вибірки для всього міста).

Перша програма	Друга програма	Відсутність будь-яких програм
-10	-21	+3
-2	-15	-10
+1	-8	+14
+2	-10	+8
+7	-5	+5
+10	-1	-2

Варіант 9 [22]:

Чи правда, що сексуально активні підлітки краще поінформовані про СНІД та інші потенційні проблеми, пов'язані із сексом, ніж неактивні?

Для трьох випадкових вибірок підлітків: сексуально неактивних (Група А); активних, які мають лише одного постійного сексуального партнера (Група В); активних, які мають більше одного сексуального партнера (Група С), – було проведене опитування з 15 пунктів на тему загальних знань про секс та здоров'я. З'ясувати, чи існують значущі відмінності у результатах опитування (у балах) між групами підлітків?

Група А	Група В	Група С
10	11	12
12	11	12
8	6	10
10	5	4
8	15	3
5	10	15

Варіант 10 [22]:

Чи істотно змінюється відсоток явки виборців в залежності від типу виборів? Оцініть значимість відмінностей наведених даних. Випадкова вибірка з 12 виборчих дільниць показує явку виборців залежно від типу виборів.

Тільки місцеві вибори	33	78	32	28	10	12	61	28	29	45	44	41
Вибори даного штату	35	56	35	40	45	42	65	62	25	47	52	55
Національні	42	40	52	66	78	62	57	75	72	51	69	59

Варіант 11 [22]:

За випадковою вибіркою респондентів соціологічного опитування, розділеною на три категорії: міські жителі, мешканці передмістя та мешканці села. З'ясувати, чи існує між ними значима відмінність за ознакою "престиж професії".

Міські жителі	32	45	42	47	48	50	51	55	60	65
Мешканці передмістя	40	48	50	55	55	60	65	70	75	75
Мешканці села	30	40	40	45	45	50	52	55	55	60

Варіант 12 [22]:

За випадковою вибіркою респондентів соціологічного опитування, розділеною на три категорії: міські жителі, мешканці передмістя та мешканці села. З'ясувати, чи існує між ними значима відмінність за ознакою "кількість дітей".

Міські жителі	1	1	0	2	1	0	2	2	1	0
Мешканці передмістя	0	1	0	0	2	2	3	2	2	1
Мешканці села	1	4	2	3	3	2	5	0	4	6

Варіант 13 [22]:

За випадковою вибіркою респондентів соціологічного опитування, розділеною на три категорії: міські жителі, мешканці пригороду та мешканці села. З'ясувати, чи існує між ними значима відмінність за ознакою "доход сім'ї".

Міські жителі	5	7	8	11	8	9	8	3	9	10
Мешканці пригороду	6	8	11	12	12	11	11	9	10	12
Мешканці села	5	5	11	10	9	6	10	7	9	8

Варіант 14 [22]:

За випадковою вибіркою респондентів соціологічного опитування, розділеною на три категорії: міські жителі, мешканці передмістя та мешканці села. З'ясувати, чи існує між

ними значима відмінність за ознакою “кількість годин, що витрачається на перегляд телепередач”.

Міські жителі	5	3	12	2	0	2	3	4	5	9
Мешканці передмістя	5	7	10	2	3	0	1	3	4	1
Мешканці села	3	7	5	0	1	8	5	10	3	1

Варіанти для двофакторного аналізу (завдання 3, 4):

Варіант 1 [10, с. 325]:

З’ясувати, що впливає на рівень уваги: стать чи вік.

Чоловіки			Жінки		
До 30 років	Від 31 до 50 років	Старші 50 років	До 30 років	Від 31 до 50 років	Старші 50 років
16	15	13	17	15	12
17	16	14	15	17	10
15	13	13	16	14	10
16	14	10	16	14	9

Варіант 2 [5]:

На двох земельних ділянках випробували врожайність двох сортів ячменю. Визначити, що впливає на врожайність: сортність, грунт, чи взаємодія цих факторів.

	Врожайність			
	Перша ділянка		Друга ділянка	
	Сорт Вінер	Сорт Нутас	Сорт Вінер	Сорт Нутас
Дослід1	27,0	32,6	19,7	23,8
Дослід2	25,6	35,0	17,0	23,0
Дослід3	25,5	33,7	21,1	25,7
Дослід4	27,1	31,9	20,1	22,4
Дослід5	27,0	33,0	19,6	20,9
Дослід6	25,7	33,2	23,4	23,6

Варіант 3 [16, 10.21]:

Досліджували вплив концентрації проявника (А) та тривалість проявки (В) на міцність фотопластинки (чим вища міцність, тим краще).

А \ В	Час проявки= 10 хв.	Час проявки =14 хв.
Концентрація 1	0, 5, 2	1, 4, 3
Концентрація 2	4, 7, 6	6, 7, 8

Варіант 4 [5]:

У таблиці наведено відомості про вплив порідних властивостей та якості бджолиних маток на яйценоскість потомства. З'ясувати, що впливає на яйценоскість дочірніх бджіл.

Породи бджіл	A1			A2			A3		
Матки	B1	B2	B3	B1	B2	B3	B1	B2	B3
Яйценоскість дочок	14	14	15	14	20	16	10	16	12
	10	14	17	17	18	17	12	12	15
	15	15	16	20	18	20	8	10	8
	14	17	13	16	16	17	13	16	12
	16	15	14	15	20	18	18	10	16
	12	16	15	20	19	20	17	15	18

Варіант 5 [16, 10.22]:

Якість приготування спагеті (ризик їх розварити) перевірялася на двох типах спагеті у двох режимах приготування за вагою, оскільки спагеті, які більш інтенсивно набирають вологу, швидко розварюються. Заготовки вагою по 150 г. запускались в каструлю з окропом, а через заданий час виймалися та зважувалися. З'Ясувати, від чого залежить якість приготування спагеті – від типу чи режима приготування?

	Час приготування= 4 хв.	Час приготування= 8 хв.
Американські	265, 270	310, 320
Італійськи	250, 245	300, 305

Варіант 6 [16, 10.23]:

Студенти досліджували залежність часу розчинювання знеболюючих таблеток (в секундах) у склянці води в залежності від торгівельної марки та температури води.

	Equate	Kroger	Alka-Seltzer
Холодна	85,87	75,98	100,11
	78,69	87,66	99,65
	76,42	85,71	100,83
	74,43	86,31	94,16
Тепла	21,53	24,10	23,80
	26,26	25,83	21,29
	24,95	26,32	20,82
	21,52	22,91	23,21

Варіант 7 [16, 10.24]:

Виготовлення мікросхем здійснюється поетапно. В експерименті досліджували вплив способів очищення та травлення на обсяг виробництва.

	Спосіб травлення	
Спосіб очищення	Новий	Стандартний
Новий1	38, 34, 38	34, 19, 28
Новий 2	29, 35, 34	20, 35, 37
Стандартний	31, 23, 38	29, 32, 30

Варіант 8 [16, 10.25]:

Досліджували залежність міцності автомобільних шин від довжини шипів та виду установки.

	Низька установка	Висока установка
Короткі шипи	18,0, 16,5, 26,0, 22,5, 21,5, 21,0, 30,0, 24,5	13,5, 8,5, 11,5, 16,0, 4,5, 4,0, 1,0, 9,0
Довгі шипи	27,5, 19,5, 31,0, 27,0, 17,0, 14,0, 18,0, 17,5	17,5, 11,5, 10,0, 1,0, 14,5, 3,5, 7,5, 6,5

Варіант 9 [16, 10.48]:

У 32 пральні машини завантажували однаковий об’єм однаково забрудненої білизни (по два завантаження на кожну з 16-ти комбінацій факторів). Вимірювали вагу видаленого бруду (у фунтах) у залежності від марки прального порошку (A, B, C, D) та тривалості прання (18, 20, 22 та 24 хв.).

	Тривалість прання (хв)			
Порошок	18	20	22	24
A	0,11 0,09	0,13 0,13	0,17 0,19	0,17 0,18
B	0,12 0,10	0,14 0,15	0,17 0,18	0,19 0,17
C	0,08 0,09	0,16 0,13	0,18 0,17	0,20 0,16
D	0,11 0,13	0,12 0,15	0,16 0,17	0,15 0,17

Варіант 10 [16, 10.49]:

Директор ткацької фабрики хоче порівняти вплив майстерності ткаль (A, B, C, D) та марки верстата (1, 2, 3) на міцність вовняної тканини.

	Верстат		
Ткачиха	1	2	3
A	115, 115, 119	111, 108, 114	109, 110, 107

B	117, 114, 114	105, 102, 106	110, 113, 114
C	109, 110, 106	100, 103, 101	103, 102, 105
D	112, 115, 111	105, 107, 107	108, 111, 110

Варіант 11 [16, 10.52]:

Вивчали залежність часу завантаження комп'ютерів від їхнього типу (MAC або Dell) та виду браузера (Netscape Communicator або Internet Explorer).

	MAC	Dell
Netscape Communicator	142, 132, 125, 136, 127, 138, 147, 143	284, 304, 273, 340, 326, 301, 291, 285
Internet Explorer	198, 210, 199, 202, 196, 213, 207, 201	285, 292, 305, 325, 297, 301, 285, 290

Завдання до теми 5

Варіанти для кореляції якісних ознак (завдання 4):

Варіант 1. [27, 5.15] Колір волосся і колір очей.

У таблиці наведено дані про 147 навмання вибраних студентів, яких було розподілено згідно з кольором їх волосся (білясте, темне) та очей (блакитні, карі).

Чи можна на підставі цих даних зробити висновок про те, що колір очей пов'язаний з кольором волосся?

Колір волосся	Колір очей		Разом
	Блакитні	Карі	
Темне	31	41	72
Біляcте	40	35	75
Разом	71	76	147

Варіант 2. [27, 5.19]

У таблиці наведено дані про 1426 ув'язнених, яких було класифіковано щодо алкогольної залежності (алкоголік, неалкоголік) і характеру злочинів, за які їх засудили (дані Горінга, цитовані К.Пірсоном). Стовпці таблиці впорядковано відповідно до "інтелектуальності" виду злочину, хоча цей зв'язок досить умовний.

	Алкогольна залежність		Разом
Вид злочину	Алкоголіки	Неалкоголіки	
Підпал	50	43	93
Згвалтування	88	62	150
Насильницькі дії	155	110	265
Крадіжка	379	300	679
Виготовлення фальшивих грошей	18	14	32
Шахрайство	63	144	207

Разом	753	673	1426

Чи можна на підставі цих даних зробити висновок про існування зв'язку між алкоголізмом і характером злочину?

Варіант 3. [27, 5.23] Розумові здібності та якість одягу.

У таблиці (Гілбі, Biometrika, 8.94) наведено розподіл 1725 школярів, класифікованих відповідно до: 1) якості їхнього одягу; 2) розумових здібностей. При цьому для характеристики розумових здібностей було використано таку градацію: *A* – розумово відсталий; *B* – млявий і недостатньо розвинений; *C* – недостатньо розвинений; *B* – млявий, але розумний; *E* – досить розумний; *F* – явно здібний; *G* – дуже здібний.

Здібності	Як одягається				Сума
	Дуже добре	Добре	Задовільно	Дуже погано	
A і B	33	41	39	17	130
C	48	100	58	13	219
D	113	202	70	22	407
E	209	255	61	10	535
F	194	138	33	10	375
G	39	15	4	1	59
Сума	636	751	265	73	1725

Чи можна на підставі цих даних дійти висновку, що якість одягу школярів і їхні розумові здібності — незалежні ознаки?

Варіант 4. [27, 5.35] Глухонімота і стать.

Під час перепису населення Англії та Уельсу в 1901 р. було зареєстровано (з точністю до тисяч) 15 729 000 чоловіків і 16 799 000 жінок; з них 3497 чоловіків і 3072 жінки глухонімі від народження.

Перевірити гіпотезу про те, що глухонімота не пов'язана зі статтю.

Варіант 5. [27, 5.27] Кмітливість та якість харчування.

У соціальному огляді (K. Pearson and Moul, 1925) 618 хлопчиків були класифіковані згідно з їхнім рівнем кмітливості та якістю харчування.

Якість	Здібності						
харчування	A	B	C	D	E	F	Разом
Добра	9	27	60	63	24	5	188
Задовільна	5	41	126	120	36	6	34
Погана	5	12	38	32	8	1	96

Разом	19	80	224	215	68	12	618

Результати обстежень наведено у вміщеній нижче таблиці. При цьому використано позначення: A – very capable (дуже здібний), B – capable (здібний), C – intelligent (кмітливий), D – slow intelligent (мало кмітливий), E — dull (млявий), F – very dull (дуже млявий).

Чи існує зв'язок між якістю харчування дітей та їхньою кмітливістю?

Варіант 6. [27, 5.31]

У таблиці наведено дані статті, в якій досліджувався взаємозв'язок між розвитком очей (який визначали за астигматизмом, гостротою зору і т.д.) та розвитком рук (який визначали за піднятою масою).

Розвиток рук	Розвиток очей			Разом
	Лівоокі	Двоокі	Правоокі	
Ліворукі	34	62	28	124
Дворукі	27	28	20	75
Праворукі	57	105	52	214
Разом	118	195	100	413

Чи можна на підставі цих даних зробити висновок про те, що розвиток рук не залежить від розвитку очей?

Варіант 7. [27, 5.39] Тістечка та бактеріальні колонії.

У таблиці наведено дані про колонії бактерій, що містяться в трьох видах тістечок. (Дані з Abrahamson Abraham E., A study of the Control of Sanitary Quality of Custard Filled Bakery Products in Large City. Food Research 17 (1958), 268-277.)

Вид тістечка	Розміри колоній			Разом
	малі	середні	великі	
Еклер	92	37	46	175
Наполеон	53	15	19	87
Горіхове	75	19	12	106
Разом	220	71	77	368

Чи можна стверджувати, що існує залежність між видами тістечок і розмірами бактеріальних колоній, що містяться в них?

Відповідь дати в термінах перевірки статистичних гіпотез.

Статура школяра	Ступінь кмітливості		Разом
	високий	низький	
Атлетична	581	567	1148

Варіант 8. [27, 5.43] Кмітливість і статура.

Неатлетична	209	351	560
Разом	790	918	1708

Нижче наведено дані про ступінь кмітливості школярів, які мають атлетичну та неатлетичну статуру.

Що можна сказати про зв'язок між кмітливістю школярів та їхньою статурою?

Відповідь дати в термінах перевірки статистичних гіпотез.

Варіант 9. [27, 5.47] Вік молодих і рівень доходів.

Проводилось обстеження з метою виявлення наявності зв'язку між віком тих, хто одружується вперше, та рівнем їхніх доходів. Результати обстеження наведено в таблиці.

Рівень доходів	Вік молодих		
	До 18	18-21	Старші 21
Низький	45	25	15
Середній	35	60	25
Високий	10	28	24

Чи свідчать ці дані про наявність зв'язку між віком першого одруження та рівнем доходів молодих?

Варіант 10. [27, 5.50]

Для виявлення зв'язку між дихальною функцією і звичкою до паління результати легеневих проб у групі співробітників установи були співставлені з режимом паління. В одній з таких проб (проба FEV1) вимірювався об'єм (у літрах) повітря, що видихується через 1 с після форсованого видиху.

Легенева проба	Нікотинова залежність		Разом
	Не палять	Палять	
Ненормальна	2	16	18
Нормальна	64	83	147
Разом	66	99	165

Результати обстеження наведено в таблиці.

Чи свідчать ці дані про зв'язок між палінням і дихальною функцією?

Варіант 11. [27, 5.51] Гострота зору.

У таблиці зібрано дані про гостроту зору у 3242 чоловіків віком 30-39 років – службовців Королівських артилерійських заводів Великобританії (1943-1946 рр.) (гострота зору визначається неозброєним оком за ступенями: вищий, другий, третій, нижчий).

Ступінь (праве око)	Ступінь (ліве око)				Сума
	вищий	другий	третій	нижчий	

Вищий	821	112	85	35	1053
Другий	116	494	145	27	782
Третій	72	151	583	87	893
Нижчий	43	34	106	331	514
Сума	1052	791	919	480	3242

Чи можна на підставі цих даних зробити висновок про те, що гострота зору правого і лівого очей не пов'язані між собою?

Варіант 12. [Турчин, 5.55] Колір волосся та колір брів.

У таблиці наведено розподіл кольору волосся та кольору брів у 46542 шведських призовників.

Колір брів	Колір волосся		Разом
	Біляве, руде	Темне	
Біляві, руді	30472	3238	33710
Темні	3364	9468	12832
Разом	33836	12706	46542

Чи свідчать ці дані про зв'язок між кольором волосся та кольором брів?

Варіант13. [27, 5.7] Кмітливість та матеріальні умови.

У таблиці наведено результати обстеження 697 школярів.

	Кмітливість хлопчиків					Разом
Забезпеченість	A	B	C	D	E	
Добра	33	137	125	47	8	350
Погана	21	127	129	61	9	347
Разом	54	264	254	108	17	697

Хлопчиків було впорядковано згідно з IQ і відповідної до умов їхнього життя вдома. При цьому використано позначення: *A* – дуже здібний, *B* — досить розумний, *C* – має середні здібності, D – недостатньо розвинений *E* – розумово відсталий. Чи можна вважати, що умови життя дітей (забезпеченість) впливають на їхню кмітливість?

Зауваження. IQ – Intellectual quality (розумові здібності) – показник розумових здібностей учнів у балах, який використовується в американській педагогічній практиці.