Міністерство освіти і науки України
Національний університет водного господарства та природокористування
Навчально-науковий інститут автоматики, кібернетики та обчислювальної техніки
Кафедра комп'ютерних технологій та економічної кібернетики

04-05-76М

Методичні вказівки

до виконання практичних робіт з навчальної дисципліни
«Аналітика цифрових даних в публічному управлінні»

для здобувачів вищої освіти другого (магістерського) рівня
за освітньо-професійними програмами
«Місцеве самоврядування» та «Державна служба»
спеціальності 281 Публічне управління та адміністрування
денної та заочної форм навчання

Рекомендовано
науково-методичною радою
з якості ННІЕМ
Протокол № 6 від 30.11.2023 р.

Рівне – 2023

Методичні вказівки до виконання практичних робіт з навчальної дисципліни «Аналітика цифрових даних в публічному управлінні» для здобувачів вищої освіти другого (магістерського) рівня за освітньо-професійними програмами «Місцеве самоврядування» та «Державна служба» спеціальності 281 Публічне управління та адміністрування денної та заочної форм навчання. [Електронне видання] / Грицюк П. М., Василів В. Б. – Рівне : НУВГП, 2023. – 32 с.

Укладачі : Грицюк П. М. д.е.н. професор, завідувач кафедри комп'ютерних технологій та економічної кібернетики;
Василів В. Б., к.т.н., доцент кафедри комп'ютерних технологій та економічної кібернетики.

Відповідальний за випуск: Грицюк П. М., завідувач кафедри комп'ютерних технологій та економічної кібернетики д.е.н. професор.

Керівники (гаранти):
ОП «Державна служба» Тихончук Л. Х. д.н.держ.упр.
ОП «Місцеве самоврядування» Мартинюк Г. Ф. к.п.н., доцент

Схвалено на засіданні кафедри комп'ютерних технологій та економічної кібернетики протокол № 5 від 1 листопада 2023 р.

Зміст

Вступ

Дисципліна «Аналітика цифрових даних в публічному управлінні» є вибірковою компонентою циклу професійної підготовки магістрів спеціальності 281 Публічне управління та адміністрування за освітньо-професійними програмами «Місцеве самоврядування» та «Державна служба». Вивчення даної дисципліни забезпечує для здобувачів вищої освіти : оволодіння сучасними методами аналізу статистичних даних; отримання навичок комп'ютер-ного аналізу даних; здатність аналізувати результати досліджень та вміння формулювати висновки та рекомендації.

Аналіз даних — це процес, який дозволяє нам зрозуміти інформацію, представлену у вигляді чисел або інших даних. Він допомагає нам виявити закономірності, тенденції та взаємозв'язки в даних, що може бути корисно для прийняття рішень.

Аналіз даних зазвичай складається з трьох етапів:

Збір даних. На цьому етапі ми збираємо дані, які нам потрібні для аналізу. Дані можуть бути зібрані з різних джерел, таких як опитування, анкетування, експерименти або статистичні звіти.

Обробка даних. На цьому етапі ми готуємо дані для аналізу. Це може включати такі операції, як очищення даних, форматування даних та визначення типів даних.

Аналіз даних. На цьому етапі ми застосовуємо різні статистичні методи для виявлення закономірностей, тенденцій та взаємозв'язків у даних.

Результати аналізу даних можна представити в різних формах, таких як таблиці, графіки, діаграми або звіти.

Аналіз даних можна використовувати для вирішення таких завдань, як:

Прогнозування. Аналіз даних може допомогти нам прогнозувати майбутні події, такі як продажі, ціни або погода.

Діагностика. Аналіз даних може допомогти нам діагностувати проблеми, такі як відхилення від норми або наявність шахрайства.

Управління. Аналіз даних може допомогти нам приймати більш ефективні управлінські рішення.

Аналіз даних — це потужний інструмент, який може бути використаний для вирішення різних завдань. Однак важливо використовувати цей інструмент відповідально і виважено.

Методичні вказівки містять набір практичних робіт для модуля 1, виконання яких дозволить поглибити розуміння різних методів аналізу даних: базовий статистичний аналіз вибірки, кореляційний та регресійний аналіз, дисперсійний аналіз, факторний аналіз, кластерний аналіз. Реалізація наведених методів ілюструється використанням програмного забезпечення MS EXCEL та STATISTICA.

1 Тема: Графічний аналіз даних в MS Excel

1. Точкова діаграма (діаграма X – Y)

А) За даними наступної таблиці побудувати точкову діаграму в MS Excel (рис. 1). Використати головне меню MS Excel: Вставка, диаграммы, Точечная.

Опади, мм	10	20	30	40	50	60	70	80	90	100
Врожайність, ц/га	25	29	34	40	47	54	58	60	56	48

Б) Відредагувати діапазони осей: OX 0 – 100; OY 20 – 65. Використати контекстне меню (права кнопка миші - ПКМ) і вибрати команду *формат оси*.

В) Додати підписи осей. Використати головне меню: *Конструктор, Добавить элемент диаграммы, Названия осей*.

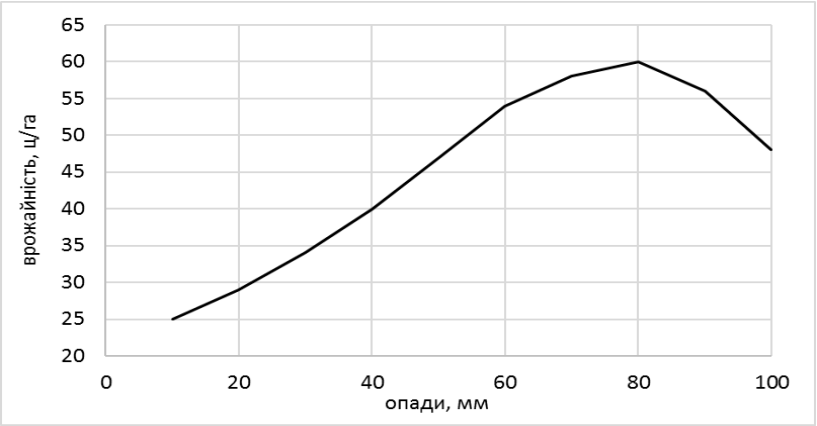


Рис.1. Точковий графік залежності врожайності від кількості опадів

2. Стовпчикова діаграма

А) За даними наступної таблиці побудувати стовпчикову діаграму в MS Excel (рис. 2). Використати головне меню MS Excel: Вставка, диаграммы, Гистограмма.

Рівень освіти	Неповна середня	Середня	Середня спеціальна	Бакалавр	Магістр
Зарплата, грн	3200	4000	5000	8000	11000

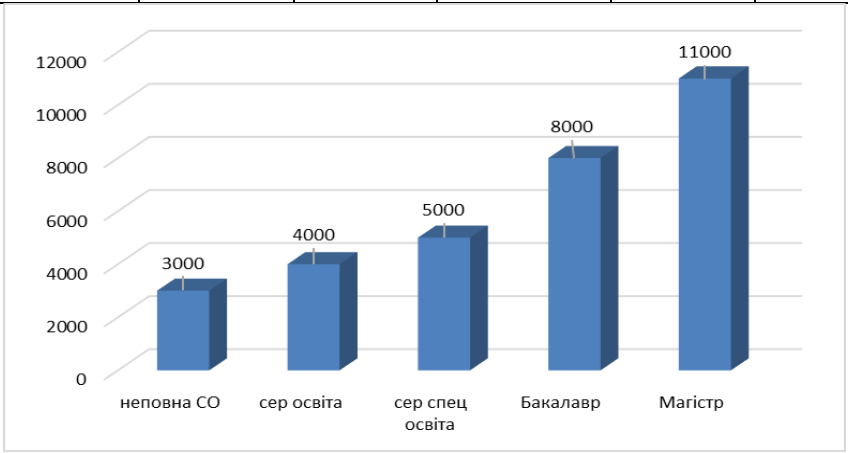


Рис. 2. Стовпчикова діаграма «залежність зарплати від рівня освіти»

Б) Вставити підписи стовпців (Выбрать данные, Пописи горизонтальной оси, Изменить).
В) Додати значення даних над стовпцями. Для цього використати контекстне меню – ПКМ (Добавить подписи данных).

3. **Секторна (кругова) діаграма**

А) За даними наступної таблиці побудувати секторну діаграму в MS Excel (рис. 3). Використати головне меню MS Excel: Вставка, диаграммы, Круговая.

Культура	Пшениця	Жито	Кукурудза	Ячмінь	Просо	Гречка
Посівна	5.9	0.3	4.7	3.5	0.2	0.3

Б) Розрахувати процентний вміст кожної культури у загальній площі посівів.
В) Вставити підписи секторів. Для цього використати контекстне меню – ПКМ (Добавить подписи данных, Добавить выноски данных).

4. **Побудова лінії тренду**

А) За даними попередньої таблиці побудувати точкову діаграму в MS Excel (рис. 4). Використати головне меню MS Excel: Вставка, диаграммы, Точечная.

Б) Відредагувати діапазони осей: OX 2000 – 2020; OY 10 – 50. Використати контекстне меню (права кнопка миші - ПКМ) і вибрати команду *формат оси*.

В) Додати підписи осей. Використати головне меню: *Конструктор,* Добавить *элемент диаграммы, Названия осей*.

Г) Додати лінію тренду. Використати контекстне меню (права кнопка миші - ПКМ) і вибрати команду *Добавить линию тренда, Линейная, Показывать уравнение на диаграмме, Поместить на диаграмму* R^2.

Д) Використовуючи лінію тренду виконати прогноз врожайності на два найближчих роки. Контекстне меню: *Формат линии тренда, Прогноз, Вперед на 2.0.*

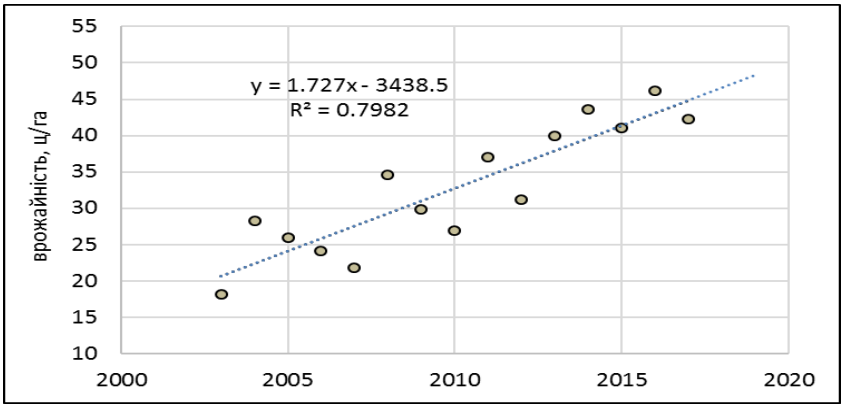


Рис.4. Побудова лінії тренду

5. **Оформлення звіту**

Звіт про лабораторну роботу оформити у MS Word. У звіт включити формулювання кожного завдання і графік, побудований за цим завданням. Графіки вставити з MS Excel у MS Word з використанням контекстного меню та команд «копировать» і «вставить».

2 Тема . Попередня обробка статистичних даних

Теоретичні відомості :

Введемо наступні позначення:

x_i - варіанти випадкових величин; n_i - відповідні частоти; m - кількість варіантів; n - обсяг вибірки; k - крок таблиці (інтервал між сусідніми варіантами). Для варіаційного ряду мода – це варіанта, яка має найбільшу частоту повторень. Для інтервального варіаційного ряду мода описується так :

$$M_0 = x_0 + k\frac{n_i - n_{i-1}}{\left(n_i - n_{i-1}\right) + \left(n_i - n_{i+1}\right)}. \quad (1)$$

Тут x_0 - початок модального інтервалу, тобто інтервалу, що має максимальну частоту; k - довжина модального інтервалу; n_i - частота модального інтервалу; n_{i-1} , n_{i+1} - частоти інтервалів перед і після модального інтервалу відповідно.

Медіана, це значення ознаки, що поділяє варіаційний ряд на дві рівні частини. Для інтервального варіаційного ряду медіана описується так : $M_e = \frac{x_{n/2} + x_{n/2+1}}{2}$, якщо n -парне; $M_e = x_{(n+1)/2}$, якщо n -непарне; тут n - об'єм вибірки.

Емпіричну функцію розподілу знаходять за допомогою накопичених частот:

$$F^*(x) = \sum_{x_i < x} \frac{n_i}{n}. \quad (2)$$

Найпростішою мірою варіації (мінливості) кількісної ознаки є розмах варіації

$$R = x_{\max} - x_{\min}. \quad (3)$$

Найголовнішими числовими характеристиками вибірки є середнє вибіркова та дисперсія. Для їх розрахунків використовують наступні формули:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i; \quad D = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \quad (4)$$

Іншими числовими характеристиками вибірки є : середнє квадратичне відхилення $\sigma = \sqrt{D}$ та коефіцієнт варіації $v = \frac{\sigma}{\overline{x}} \cdot 100\%$.

Скошеність ряду розподілу від центра характеризує асиметрія $A = \frac{\mu_3}{\sigma^3}$.

Тут $\mu_3 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3 n_i}{\sum_{i=1}^{n} n_i}$ - центральний момент третього порядку.

Істотність коефіцієнту асиметрії оцінюють за допомогою середньоквадратичної похибки асиметрії

$$\sigma_A = \sqrt{\frac{6 \cdot (n-1)}{(n+1)(n+3)}}. \qquad (5)$$

Якщо $|A| > 3 \cdot \sigma_A$, то асиметрія розподілу значна і розподіл вважається несиметричним.

Міра ексцесу (гостровершинність) – це крутизна розподілу. Її визначають за співвідношенням $E = \frac{\mu_4}{\sigma^4 - 3}$.

Тут $\mu_4 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4 n_i}{\sum_{i=1}^{n} n_i}$ - центральний момент четвертого порядку.

Істотність міри ексцесу оцінюють за допомогою середньоквадратичної похибки ексцесу

$$\sigma_E = \sqrt{\frac{24n \cdot (n-2) \cdot (n-3)}{(n-1)^2 (n+3)(n+5)}}. \qquad (6)$$

Якщо $|E| > 3 \cdot \sigma_E$, то ексцес розподілу значний.

Завдання: Використовуючи табличні дані сформувати вибірку обсягом 150 чисел. За вибіркою розв'язати наступні задачі:

1) Визначити мінімальне та максимальне значення вибірки. Для вибірки побудувати інтервальний варіаційний ряд, взявши за ширину інтервалу число 4.

2) Обчислити частоти даних n_i для кожного інтервалу.

3) Обчислити відносні частоти (n_i / n) і накопичені частоти F_i^*.

4) Побудувати полігон частот (рис. 5) та гістограму частот (рис.6) варіаційного ряду.

5) Побудувати емпіричну функцію розподілу (рис. 7).

6) Побудувати графік емпіричної функції розподілу.

7) Використовуючи наведені вище формули обчислити характеристики варіаційного ряду:

- Середнє вибіркове $\overline{x}$;
- Дисперсію *D*;
- Середнє квадратичне відхилення σ;
- Моду *Mo*;
- Медіану *Me*;
- Асиметрію *A*;
- Ексцес *E*.

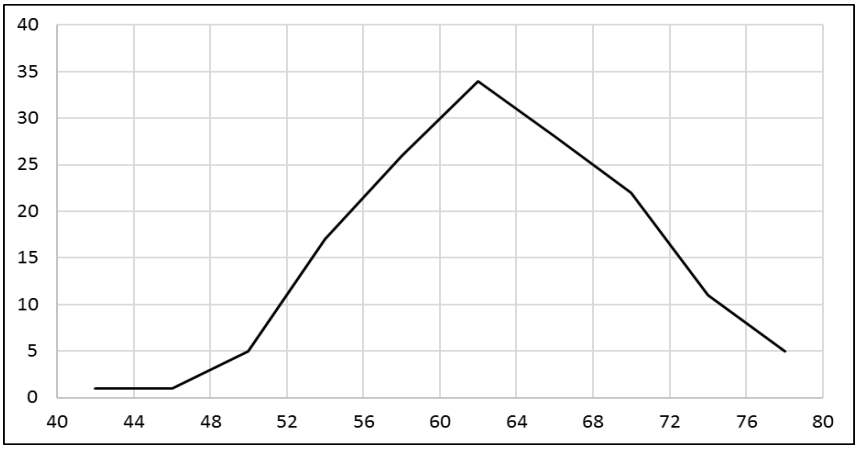


Рис. 5. Полігон частот

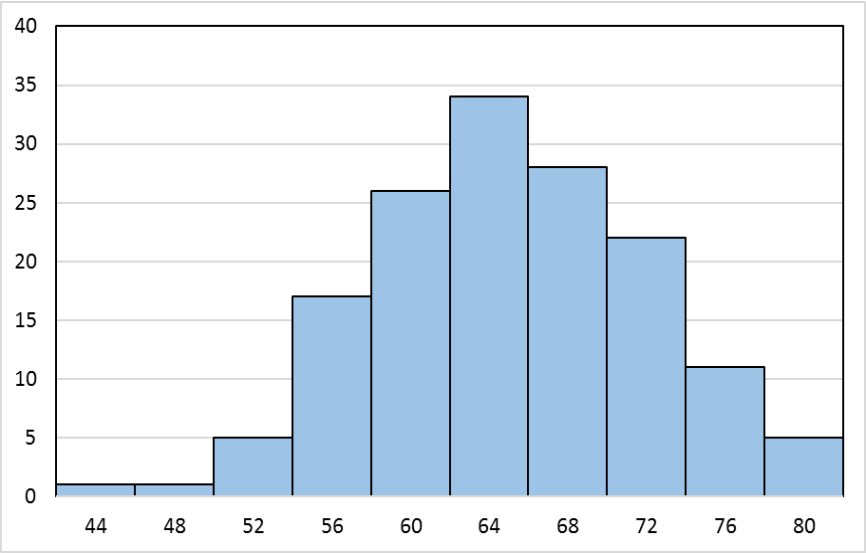


Рис. 6. Гістограма частот

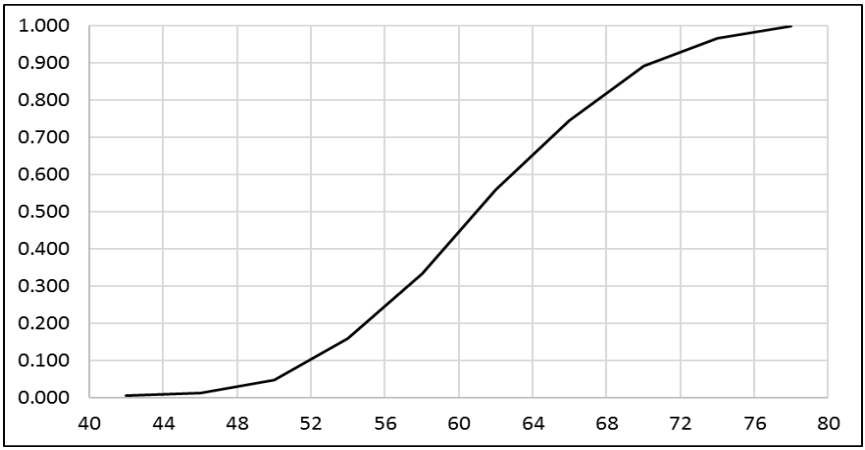


Рис.7. Емпірична функція розподілу

3 Тема . Перевірка статистичних гіпотез

Задача 1. Перевірка показала. що дані про витрати сировини для виробництва продукції за старою та новою технологіями (таблиця) розподілені нормально. Перевірити гіпотезу про рівність дисперсій.

Номер виробу	1	2	3	4	5	6	7	8	9	10	11	12	13
Стара технологія	308	308	307	308	304	307	307	308	307	306			
Нова технологія	308	304	306	306	306	304	304	304	306	304	303	304	303

Розв'язання

Розрахуємо числові статистичні параметри обох вибірок.

Для розрахунку середніх витрат використаємо Excel-функцію СРЗНАЧ(). Отримаємо наступні результати: стара технологія – 307.0, нова технологія – 304.8. Для розрахунку дисперсії середніх витрат використаємо Excel-функцію ДИСП(). Отримаємо наступні результати: стара технологія – 1.556, нова технологія – 2.192.

Спочатку треба перевірити чи істотною є відмінність дисперсій середніх витрат за старою та новою технологіями.

Умови застосування гіпотези. 1. Вибіркові дані незалежні. 2. Вибіркові дані розподілені за нормальним законом.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза: $H_0 : S_1^2 = S_2^2$ (відмінність між дисперсіями неістотна; дисперсії однакові).

Альтернативна гіпотеза: $H_1 : S_1^2 > S_2^2$ (відмінність між дисперсіями істотна; дисперсії неоднакові).

Вибір статистичного критерію. Для перевірки рівності дисперсій застосуємо критерій Фішера.

Визначення критичної області при α = 0.05 (однобічний критерій). Визначаємо кількість ступенів свободи.

$df_1 = n_2 - 1 = 12$ (n_2 – обсяг вибірки з більшою дисперсією);

$df_2 = n_1 - 1 = 9$ (n_1 – обсяг вибірки з меншою дисперсією).

Fкр = 3.07 – F.ОБР.ПХ(0.05; 12; 9).

Розрахунок статистичного критерію Фішера.

$$F = \frac{S_1^2}{S_2^2} = \frac{2.19}{1.56} = 1.40 \cdot$$

Прийняття (відхилення) нульової гіпотези.

Оскільки $F < F_{kp}$ є підстави прийняти нульову гіпотезу і твердити, що відмінність між дисперсіями є неістотною.

Інший спосіб розв'язання попередньої задачі - використання вбудованого пакета аналізу шляхом виклику вікна: *Сервіс — Аналіз даних — Двовибірковий для дисперсій* Особливістю його використання є те, що першим з двох масивів треба ввести діапазон даних з більшою дисперсією (таблиця на попередній сторінці).

Задача 2. За результатами контрольної перевірки якості отриманого магазином маргарину відомими отримали такі дані про вміст консерванту Е205 (%) у 10 пробах: 4.3, 4.2, 3.8, 4.3, 3.7, 3.9, 4.5, 4.4, 4.0, 3.9. З надійністю 95 % перевірити гіпотезу про те, що середній вміст консерванту в усій партії маргарину становить 4.0 %.

Розв'язання

Умови застосування гіпотези. 1. Вибіркові дані незалежні. 2. Нормальний розподіл генеральної сукупності.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза $H_0 : \bar{x} = A$ (вміст консерванту в усій партії маргарину становить 4.0 %).

Альтернативна гіпотеза: $H_1 : \bar{x} \neq A$ (вміст консерванту в усій партії маргарину не становить 4.0 %).

Вибір статистичного критерію. Для перевірки гіпотези про середній вміст консерванту використаємо критерій Стьюдента.

Визначення критичної області при α = 0.05 (двобічний критерій).

Визначаємо кількість ступенів свободи *df* = *n* - 1 = 10 – 1 = 9.

Визначаємо середнє вибіркове значення $\bar{x} = 4.1$

Визначаємо критичне значення статистики Стьюдента *tкр* =2.685 – СТЬЮДЕНТОБР2Х(0.025; 9).

Розрахунок статистичного критерію (статистика Стьюдента)

$$t_p = \frac{(\bar{x} - A)\sqrt{n}}{S} = \frac{(4{,}1 - 4{,}0)\sqrt{10}}{\sqrt{0{,}076}} = 1{,}15 \qquad (10)$$

Прийняття (відхилення) нульової гіпотези.

Розрахунковий статистичний критерій менший за теоретичний, що з імовірністю 95 % дає підстави стверджувати: вміст консерванту у складі маргарину відповідає рівню 4.0 %.

Задача 3. Групу, яка складається з 13 студентів протестували на швидкість набору тексту до і після проведення спеціального тренінгу. Результати тестування (символів за хвилину) є наступними.
До тренінгу: 29; 98; 113; 76; 128; 56; 71; 62; 130; 59; 44; 122; 46.
Після тренінгу: 33; 108; 115; 77; 130; 61; 78; 68; 134; 65; 48; 120; 51.
Чи є статистично істотним збільшення швидкості набору тексту студентами?

Розв'язання

Умови застосування гіпотези. 1. Нормальний розподіл генеральної сукупності. 2. Вибірки взаємопов'язані. 3. Дані незалежні.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза $H_0: \bar{x} \le \bar{y}$ (швидкість набору тексту студентами після проведення тренінгу не змінилася (або зменшилася)) проти альтернативної $H_1: \bar{x} > \bar{y}$ (швидкість набору тексту збільшилася).

Вибір статистичного критерію. Для перевірки гіпотези використаємо критерій Стьюдента для пов'язаних вибірок.

Визначення критичної області при $\alpha = 0.05$ (однобічний критерій).

Кількість ступенів свободи df = n - 1 = 13 - 1 = 12.

Критичне значення статистики Стьюдента $t_{кр}$ = 2.179 - СТЬЮДЕНТ.ОБР.2Х(0.05;12).

Розрахунок статистичного критерію.

Критерій Стьюдента для пов'язаних вибірок розраховується за співвідношенням:

$$t_p = \frac{\sum_{i=1}^{n}(x_i - y_i)\sqrt{n-1}}{\sqrt{n\sum_{i=1}^{n}(x_i - y_i)^2 - \left(\sum_{i=1}^{n}(x_i - y_i)\right)^2}} = \frac{-54 \cdot \sqrt{13-1}}{\sqrt{13 \cdot 332 - (-54)^2}} = -4.999 \cdot \quad (11)$$

Прийняття (відхилення) нульової гіпотези.

Абсолютна величина розрахункового статистичного критерію перевищує величину теоретичного критерію. Це з імовірністю 95% дає підс-

тави стверджувати, що після проведення тренінгу швидкість набору тексту студентами збільшилась.

У пакеті MS Excel розв'язати цю задачу можна за допомогою вбудованого пакета аналізу: Сервис – Анализ данных – Парный двухвыборочный t-тест для средних.

Задача 4. Перевірити гіпотезу про рівність середніх за даними задачі 1. Пояснити отриманий результат.

Розв'язання

Умови застосування гіпотези. 1. Нормальний розподіл. 2. Дані незалежні. 3. Дисперсії однакові. Як бачимо, необхідною попередньою вимогою є перевірка гіпотези про рівність дисперсій. Цю перевірку ми виконали у задачі 1 і отримали позитивний результат.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза: $H_0: \bar{x}_1 \le \bar{x}_2$ (використання старої технології потребує меншої (тієї ж) кількості сировини. що і за нової технології). Альтернативна гіпотеза: $H_1: \bar{x}_1 > \bar{x}_2$ (використання старої технології потребує більшої кількості сировини. ніж за нової технології).

Вибір статистичного критерію.

Для перевірки гіпотези використаємо критерій Стьюдента за умови рівності дисперсій.

Визначення критичної області при $\alpha = 0.05$ (однобічний критерій).

Кількість ступенів свободи df $= n_1 + n_2 - 2 = 10 + 13 - 2 = 21$.

Критичне значення статистики Стьюдента $t_{kp} =$ 2.080 - СТЬЮДЕНТ.ОБР.2Х(0.05;21).

Розрахунок статистичного критерію.

Критерій Стьюдента за умови рівності дисперсій розраховується за співвідношенням

$$t_p = \frac{(\hat{x}_1 - \hat{x}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(S_1^2 \cdot (n_1 - 1) + S_2^2 \cdot (n_2 - 1)\right)}} = \frac{(307 - 304.8)\cdot\sqrt{21}}{\sqrt{\left(\frac{1}{10} + \frac{1}{13}\right)(1.56 \cdot 9) + 2.19 \cdot 12}} = 3.8$$

Прийняття (відхилення) нульової гіпотези.

Розрахунки дають підстави відкинути нульову гіпотезу про рівність

середніх (оскільки $\left|t_p\right| > t_{df,\alpha}$. Отже. запровадження нової технології таки зменшує витрати сировини у середньому на одиницю продукції.

У програмі Excel розв'язати цю задачу можна за допомогою вбудованого пакета аналізу: *Сервис -— Анализ данных — Двухвыборочный t-тест с одинаковыми дисперсиями* .

Задача 5. На підприємстві провели вибіркове обстеження зарплати 36 чоловіків та 40 жінок. Характеристики цих вибірок подано в таблиці.

Показник	Чоловіки	Жінки
Середня зарплата	1230	980
Дисперсія	16	9
Обсяг вибірки	36	40

На попередніх етапах аналізу були підтверджені гіпотези про нормальність обох розподілів, незалежність вибірок та відхилено гіпотезу про рівність дисперсій. За допомогою статистичної гіпотези необхідно визначити чи істотною є відмінність між зарплатнею чоловіків та жінок.

Розв'язання

Перевіримо гіпотезу про рівність середніх за допомогою критерію Стьюдента за умови нерівності дисперсій:

Умови застосування гіпотези. 1. Нормальний розподіл генеральних сукупностей. 2.Вибірки незалежні. 3. Вибіркові дисперсії нерівні.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза: $H_0 : \bar{x}_1 \le \bar{x}_2$ (середня заробітна платня чоловіків менша (дорівнює) середній зарплатні жінок).

Альтернативна гіпотеза: $H_1 : \bar{x}_1 > \bar{x}_2$ (середня заробітна платня чоловіків перевищує середню зарплатню жінок у генеральній сукупності).

Вибір статистичного критерію.

Для перевірки гіпотези використаємо критерій Стьюдента за умови нерівності дисперсій.

Визначення критичної області при $\alpha = 0.05$ (однобічний критерій).

Критерій Стьюдента за умови нерівності дисперсій розраховується за співвідношеннями

$$df = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}{\sqrt{\frac{1}{n_1+1}\cdot\left(\frac{S_1^2}{n_1}\right)^2+\frac{1}{n_2+1}\cdot\left(\frac{S_2^2}{n_2}\right)^2}}-2=\frac{\left(\frac{16}{36}+\frac{9}{36}\right)^2}{\frac{1}{36+1}\cdot\left(\frac{16}{36}\right)^2+\frac{1}{40+1}\cdot\left(\frac{9}{40}\right)^2}-2=66;$$

t_{kp} = 2.00. - СТЬЮДЕНТ.ОБР.2Х(0.05;66).

Розрахунок статистичного критерію.

Критерій Стьюдента за умови нерівності дисперсій розраховується за співвідношенням

$$t_p=\frac{\left(\hat{x}_1-\hat{x}_2\right)}{\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}}=\frac{1230-980}{\sqrt{\left(\frac{16}{36}+\frac{9}{40}\right)}}=305.6\cdot \qquad (12)$$

Прийняття (відхилення) нульової гіпотези.

Розрахунковий статистичний критерій суттєво перевищує критичний, що дає підстави відхилити нульову гіпотезу та прийняти альтернативну. Отже середня зарплатня чоловіків на підприємстві суттєво перевищує середню зарплатню жінок.

У програмі Excel таку гіпотезу перевірити можливо за допомогою вбудованого пакета аналізу: *Сервис - Анализ данных - Двухвыборочный і-тест* с *разными дисперсиями.*

Задача 6. За даними двох вибірок перевірити гіпотезу про рівність середніх вибіркових без припущення щодо дисперсій.

Перша вибірка :5.6; 5.9; 4.9; 5.9; 6.6; 4.5; 3.1; 3.9; 4.9; 5.7; 4.9; 5.7.

Друга вибірка: 3.6; 4.0; 5.2; 4.0; 4.4; 4.7; 5.9; 3.4; 4.0; 5.5; 5.0; 4.2.

Розв'язання

Умови застосування гіпотези. 1. Однакові обсяги вибірок. 2. Нормальний розподіл генеральних сукупностей. 3. Дані незалежні.

Формулювання нульової та альтернативної гіпотез.

Нульова гіпотеза: $H_0: \bar{x}_1 = \bar{x}_2$ (середні вибіркові рівні).

Альтернативна гіпотеза: $H_1: \bar{x}_1 \neq \bar{x}_2$ (середні вибіркові нерівні).

Вибір статистичного критерію.

Для перевірки гіпотези використаємо критерій Стьюдента без припущень щодо дисперсій.

Визначення критичної області при $a = 0.05$ (двобічний критерій). Кількість ступенів свободи $df = n - 1 = 12 - 1 = 11$. Критичне значення статистики Стьюдента $t = 2.593$ - СТЬЮДЕНТ.ОБР.2Х(0.025;11).

Розрахунок статистичного критерію.

Критерій Стьюдента без припущення щодо дисперсій має вигляд

$$t_p = \frac{(\hat{x}_1 - \hat{x}_2)\sqrt{n}}{\sqrt{\sum_{i=1}^{n}\left(x_{1i} - x_{2i} - (\hat{x}_1 - \hat{x}_2)\right)^2 \frac{1}{(n-1)}}} = \frac{(5.13 - 4.49)\sqrt{12}}{\sqrt{22.4492 \cdot \frac{1}{11}}} = 1.55. \quad (13)$$

Прийняття (відхилення) нульової гіпотези.

Розрахунковий статистичний критерій менший за теоретичний, тому з імовірністю 95 % маємо підстави стверджувати. що середні значення досліджуваних вибірок є рівними.

Завдання для самостійної роботи

Задача 1. Перевірити гіпотезу про рівність дисперсій для двох нормально розподілених вибірок

Номер студента	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Група 1	79	83	79	86	78	80	77	81	80	85	82			
Група 2	85	83	79	79	85	82	75	83	77	78	75	77	73	79

Задача 2. З контрольної перевірки жирності отриманого магазином кефіру відомими є такі дані у 10-и пробах, %:

2.6	2.5	2.2	2.6	2.1	2.4	2.6	2.6	2.4	2.6

З імовірністю 95 %, перевірити гіпотезу про те, що середня жирність партії кефіру становить 2.5 %.

Задача 3. Групу з 13 студентів протестували з дисципліни «Аналіз даних» до і після проведення консультації. Отримані результати у балах:
До консультації: 29; 88; 73; 76; 68; 56; 71; 62; 90; 79; 44; 82; 66.
Після консультації: 33; 93; 75; 77; 64; 61; 73; 68; 92; 84; 48; 84; 67.
Чи є статистично істотним покращення успішності студентів?

Задача 4. Перевірити гіпотезу про рівність середніх за даними задачі 1. Пояснити отриманий результат.

Задача 5. Значення середнього бала бюджетників (18 осіб) та платників (20 осіб) наведені у таблиці. Попереднім аналізом було підтверджено

гіпотези про нормальність обох розподілів, незалежність вибірок та відхилено гіпотезу про рівність дисперсій. Методом перевірки статистичної гіпотези визначити, чи істотною є відмінність між середнім балом студентів бюджетної і платної форми навчання.

Показник	Бюджетники	Платники
Середній бал	4.20	4.05
Дисперсія	0.56	0.38
Обсяг вибірки	18	20

Задача 6. Маємо дані двох вибірок (середні бали студентів двох груп):

4.6	4.9	3.9	4.9	5	3.5	2.1	2.9	3.9	4.7	3.9	4.7
2.6	3.8	4.2	3.5	3.4	4.7	4.9	3.4	3.1	4.5	4	3

Перевірити гіпотезу про рівність середніх без припущень про дисперсії.

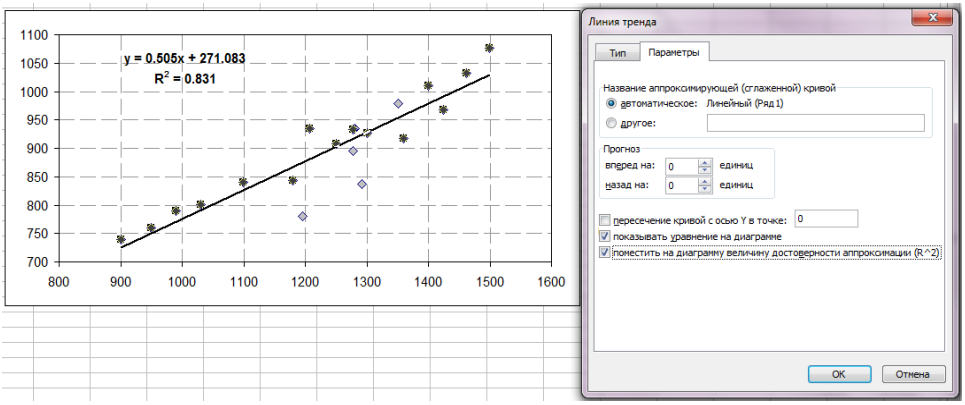
4 Тема . Парна лінійна регресія

Поширеним інструментом структурного аналізу лінійних систем є кореляційний аналіз, який дозволяє ідентифікувати зв'язки між елементами системи. Регресійний аналіз є інструментом моделювання виявлених зв'язків.

Завдання. За табличними даними побудувати рівняння лінійної регресії $y^* = a_0 + a_1 x$ і виконати його дослідження. Зміст параметрів моделі: x - обсяг коштів, які виділяються на технічне обслуговування автомобілів (тис. грн), y - річний прибуток фірми (тис. грн).

Виконання роботи.

1. В середовищі Microsoft Excel розмістити початкові дані у вигляді таблиці, зображеної на наступній сторінці (табл. 1).
2. За даними стовпців x1, y побудувати точковий графік (Диаграмма, Точечная).
3. Додати лінію тренду (права кнопка миші, Добавить линию тренда, Параметры, Показывать уравнение на диаграмме, Поместить на диаграмму R^2 (наступний рисунок).



4. Обчислити коефіцієнт кореляції $r = \sqrt{R^2}$. Оцінити статистичну значущість коефіцієнта кореляції за критерієм $|t_r| > t_{kp}$. Тут $t_{kp} = t_{\alpha/2,n-2}$ де α - рівень значущості, зв'язаний з рівнем надійності $P = 0.95$ співвідношенням $\alpha = 1 - P$; $n = 20$; $t_r = r \cdot \sqrt{n-2} / \sqrt{1-r^2}$.

5. Використовуючи оцінки коефіцієнтів парної лінійної регресії a_0, a_1, отримані з діаграми, розрахувати теоретичні значення параметра $y*$ за формулою $y* = a_0 + a_1 x$.

6. Використовуючи функцію Предсказ, розрахувати теоретичні значення параметра $y**$. Порівняти значення $y*$ і $y**$. Пояснити причину розбіжностей.

7. Заповнивши три останні стовпці таблиці, розрахувати загальну суму квадратів відхилень $SQ_y = \sum (y_i - y_c)^2$; факторну (пояснену) суму квадратів відхилень $SQ_{факт} = \sum (y_i ** - y_c)^2$; залишкову суму квадратів відхилень $SQ_{ост} = \sum (y_i - y_i **)^2$.

8. Перевірити рівність $SQ_y = SQ_{факт} + SQ_{ост}$.

9. Обчислити загальну дисперсію $D = SQ_y / (n-1)$, факторну дисперсію $D_{факт} = SQ_{факт}$ та залишкову дисперсію $D_{ост} = SQ_{ост} / (n-2)$.

10. Визначити параметри лінійної регресії, використовуючи функцію ЛИНЕЙН. Виділяємо 5 рядків і 2 стовпці. Мастер функций, Статистические, ЛИНЕЙН. CTRL+SHIFT+ENTER.

11. Перевірити рівність $F = D_{факт} / D_{ост}$.

12. Перевірити гіпотезу про адекватність лінійної моделі $F > F_{кр}$, якщо $F_{кр} = FРАСПОБР(\alpha, 1, n-2)$. $\alpha = 0.05; n = 20$.

13. Визначити фактичне значення t – критерію Стьюдента для коефіцієнтів регресії: $t_i = a_i / S_{a_i}$. Значення a_i знаходяться в першому рядку таблиці результатів функції ЛИНЕЙН (табл.1 - виділено сірим фоном), значення S_{a_i} - в другому рядку цієї таблиці.

14. Використовуючи функцію СТЬЮДРАСПОБР($\alpha, n-2$), визначити критичне значення параметра Стьюдента $t_{кр}$. $\alpha = 0.05; n = 20$.

5 Тема . Множинна лінійна регресія

Часто на залежну величину впливає відразу декілька чинників. Виникає завдання ідентифікувати ступінь впливу кожного з чинників та відобразити цей вплив шляхом побудови лінійної регресійної моделі. Розглянемо найпростіший варіант – побудова лінійної регресії з двома впливаючими змінними. Спочатку за табличними даними (залежна змінна та впливаючі фактори) необхідно побудувати кореляційну матрицю (табл. 1).

Таблиця 1. Кореляційна матриця

	У	О4	О5	О6	Т4	Т5	Т6
У	1.00						
О4	0.51	1.00					
О5	0.61	0.29	1.00				
О6	-0.18	0.21	0.19	1.00			
Т4	-0.13	-0.18	0.13	-0.19	1.00		
Т5	-0.67	-0.21	-0.72	0.31	-0.20	1.00	
Т6	-0.07	0.17	-0.15	-0.19	-0.18	0.11	1.00

Використовуючи кореляційну матрицю (перший стовпець) потрібно відібрати два фактори для побудови моделі множинної лінійної регресії виду $y^* = a_0 + a_1 x_1 + a_2 x_2$. При цьому потрібно перевірити відсутність кореляції між вибраними факторами за критерієм Стьюдента. Якщо кореляція відсутня, використовуючи функцію ЛИНЕЙН() побудувати модель множинної лінійної регресії та встановити її адекватність.

Для перевірки існування кореляції між двома випадковими факторами використовують коефіцієнт кореляції r_{xy} та t-критерій Стьюдента

$$t < t_{kp}(\alpha, n-2). \qquad (14)$$

Тут n - кількість даних вибірки; α – рівень істотності (значущості), який найчастіше обирають рівним 0.05. Рівень істотності — це та мінімальна ймовірність, починаючи з якої можна вважати подію практично неможливою. Критичне значення $t_{кр}$ розраховують за допомогою Excel-функції СТЬЮДЕНТ.ОБР.2Х(α, n-2) (СТЬЮДРАСПОБР(α, n-2). t-критерій для коефіцієнта кореляції розраховують за формулою

$$t_r = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}. \qquad (15)$$

Якщо критерій Стьюдента (14) виконується, кореляція вважається неістотною. Іншими словами, кореляційний зв'язок між факторами відсутній.

В даній роботі вважається, що попередній кореляційний аналіз проведено і два фактори для побудови моделі регресії вибрано.

Виконання роботи.

Завдання. За табличними даними побудувати рівняння лінійної регресії $y^* = a_0 + a_1 x_1 + a_2 x_2$ і виконати його дослідження. Зміст параметрів моделі: x_1 - обсяг коштів, які виділяються на технічне обслуговування автомобілів (тис. грн.), x_2 - кількість автомобілів у фірмі, яка займається перевезеннями вантажів, y - річний прибуток фірми (тис. грн.).

1. В середовищі Microsoft Excel розмістити початкові дані у вигляді таблиці, зображеної на наступній сторінці (табл. 2).
2. Визначити параметри лінійної регресії, використовуючи функцію ЛИНЕЙН. Виділяємо 5 рядків і 3 стовпці. Мастер функций, Статистические, ЛИНЕЙН. CTRL+SHIFT+ENTER.
3. Використовуючи оцінки коефіцієнтів парної лінійної регресії a_0, a_1, a_2 (перший рядок таблиці результатів функції ЛИНЕЙН) розрахувати теоретичні значення параметра y^* за формулою $y^* = a_0 + a_1 x_1 + a_2 x_2$.
4. Використовуючи функцію Тенденция, розрахувати теоретичні значення параметра y^{**}. Порівняти значення y^* і y^{**}.
5. Заповнивши три останні стовпці таблиці, розрахувати загальну суму квадратів відхилень $SQ_y = \sum (y_i - y_c)^2$; факторну (пояснену) суму квадратів відхилень $SQ_{факт} = \sum (y_i{}^{**} - y_c)^2$; залишкову суму квадратів відхилень $SQ_{ост} = \sum (y_i - y_i{}^{**})^2$.
6. Перевірити рівність $SQ_y = SQ_{факт} + SQ_{ост}$.

7. Обчислити загальну дисперсію $D = SQ_y/(n-1)$, факторну дисперсію $D_{факт} = SQ_{факт}/2$ та залишкову дисперсію $D_{ост} = SQ_{ост}/(n-m-1)$. Тут $n = 20,\ m = 2$.

8. Перевірити рівність $F = D_{факт}/D_{ост}$.

9. Перевірити гіпотезу про адекватність лінійної моделі за критерієм Фішера $F > F_{kp}$, якщо $F_{кр} = FРАСПОБР(\alpha, 1, n-m-1)$. $\alpha = 0.05;$ $n = 20;\ m = 2$.

10. Визначити фактичне значення t – критерію Стьюдента для коефіцієнтів регресії: $t_i = a_i / \sigma_{a_i}$.

11. Використовуючи функцію СТЬЮДРАСПОБР($\alpha, n-m-1$), визначити критичне значення параметра Стьюдента $t_{кр}$. $\alpha = 0.05;\ n = 20;\ m = 2$.

12. Перевірити статистичну значущість коефіцієнтів a_0, a_1 і a_2 за критерієм $t_i > t_{kp}$.

13. Використовуючи побудовану модель регресії розрахувати прогнозне значення y*. Значення впливаючих факторів вибрати як середні значення за три останні періоди.

6 Тема . Однофакторний дисперсійний аналіз

Завдання. За табличними даними (таблиця 1) вияснити, чи впливає контрольований фактор на результативну ознаку і оцінити ступінь такого впливу. Роботу виконуємо у такій послідовності.

1. За допомогою критерію Бартлета перевірити гіпотезу про рівність генеральних дисперсій**.**

1.1. Обчислити незміщені групові дисперсії $D_i = \dfrac{\sum (x_i - \bar{x}_i)^2}{n_i - 1}$,

$i = 1,2,3$.

1.2. Обчислити об'єднану дисперсію $D = \sum_{i=1}^{m} (n_i - 1) D_i \,/\, \sum_{i=1}^{m} n_i - k$.

Тут n_i - обсяги вибірок, k = 3 – кількість вибірок.

1.3. Обчислити параметр q:

$$q = \left[1 + \frac{1}{3(k-1)} \cdot \left(\frac{3}{n-1} - \frac{1}{3(n-1)} \right) \right]^{-1}. \qquad (16)$$

У нашому випадку маємо $k = 3; n = 6$.

1.4. Обчислити критерій Бартлета: $w = q \cdot \left[\sum_{i=1}^{k} (n_i - 1) \cdot \ln \dfrac{D}{D_i} \right]$. (17)

1.5. Визначити критичне значення $w_{кр}(\alpha, k-1)$ – функція *ХИ2ОБР(α,k-1).*

1.6. Зробити висновок про рівність (нерівність) дисперсій. Якщо $w < w_{кр}$ гіпотеза про рівність дисперсій приймається.

2. Якщо гіпотеза про рівність дисперсій підтвердилась, потрібно приступити до перевірки гіпотези про рівність математичних сподівань $H_0 : \bar{x}_A = \bar{x}_B = \bar{x}_C$ за таким алгоритмом.

2.1. Обчислити групові середні $\overline{x_A}, \overline{x_B}, \overline{x_C}$ та середнє вибіркове $\bar{x}$.

2.2. Визначити три незміщені оцінки варіації:

загальну варіацію $S^2_{заг} = \sum_i \sum_j \left(x_{ij} - \overline{x}\right)^2$; (18)

міжгрупову варіацію $S^2_{ф} = \sum_i \left(\overline{x}_i - \overline{x}\right)^2 n_i$; (19)

внутрішньогрупову варіацію $S^2_{зал} = \sum_i \sum_j \left(x_{ij} - \overline{x_i}\right)^2$. (20)

2.3. Перевірити виконання співвідношення $S^2_{заг} = S^2_{ф} + S^2_{зал}$. (21)

2.4. Визначити:

міжгрупову дисперсію $D_{ф} = \frac{S^2_{ф}}{k-1}$; (22)

внутрішньогрупову дисперсію $D_{зал} = \frac{S^2_{зал}}{N-k-1}$; (23)

загальну дисперсію $D_{заг} = \frac{S^2_{заг}}{N-1}$; (24)

2.5. Обчислити ***F*** – критерій Фішера $F = S^2_{ф} \Big/ S^2_{зал}$. (25)

2.6. Визначити критичне значення
$F_{кр} = FРАСПОБР(\alpha, k-1, n-k-1)$.

2.7. Перевірити виконання критерію Фішера $F > F_{кр}$ та зробити висновок про вплив виду добрив на врожайність. Наприклад: *оскільки* $F > F_{кр}$ *гіпотеза про рівність середніх відхиляється, тобто вплив фактора на контрольовану ознаку є суттєвим.*

2.8. Якщо гіпотеза про рівність середніх відхилена, обчислити вибірковий коефіцієнт детермінації $R^2 = \frac{D_{ф}}{D_{заг}}$. Зробити висновок: . . . *% загальної вибіркової варіації . . . зв'язано з впливом фактора* . . .

7 Тема. Кластерний аналіз

Кластерний аналіз передбачає розбиття заданої вибірки об'єктів на підмножини, які називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

Завдання. Виконати кластерний аналіз областей України за деякими показниками їх економічної діяльності.

1. Вхідні табличні дані врожайності зернових культур (24 області, 18 років) завантажити у програму Statistica.
2. **Кластерний аналіз. Побудова ієрархічної класифікації об’єктів.** Виконати команди: Statistics, Multivariate Exploratory Techniques, Cluster Analysis, Joining (Tree clustering), Variables – Select All, Vertical icicle plot (рис. 15).
3. Перемалювати отриману діаграму у зошит. Пояснити отриману класифікацію областей (розбиття на три кластери – рис. 16).

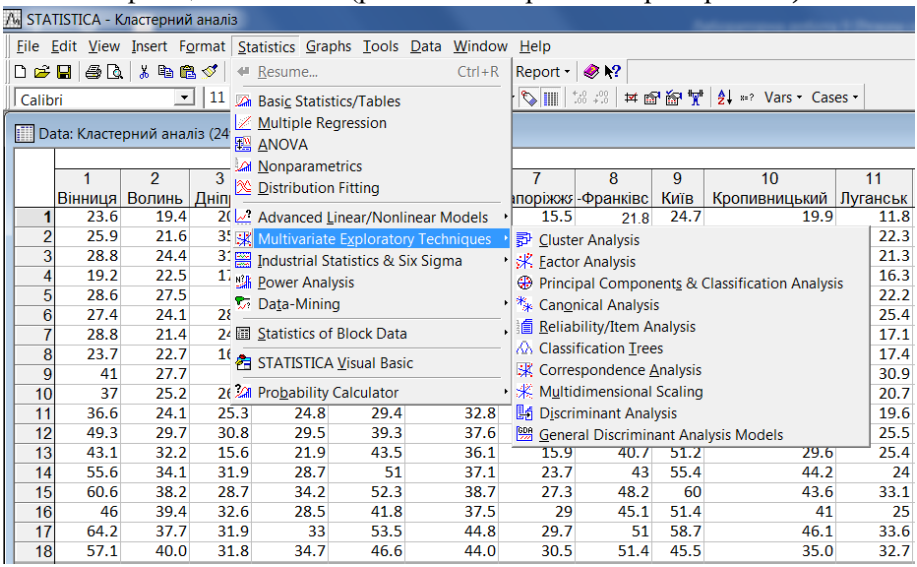


Рис. 15. Кластерний аналіз. Ієрархічна класифікація об’єктів

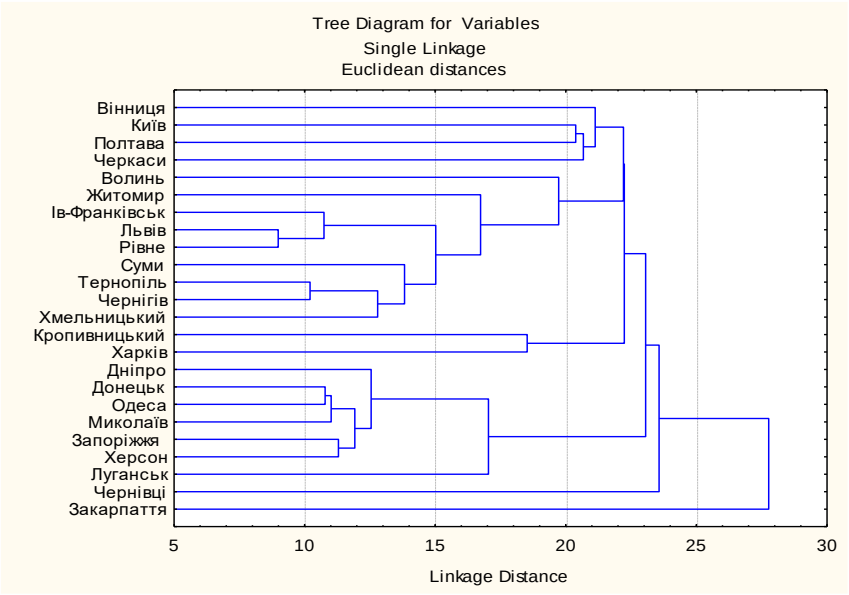


Рис. 16. Кластерний аналіз. Горизонтальна деревоподібна діаграма

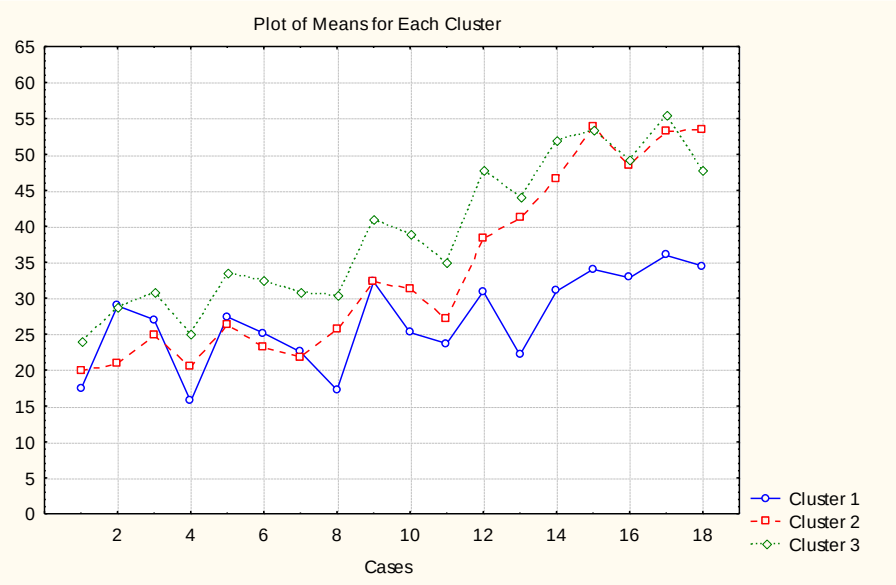


Рис.17. Часова динаміка середніх для кластерів

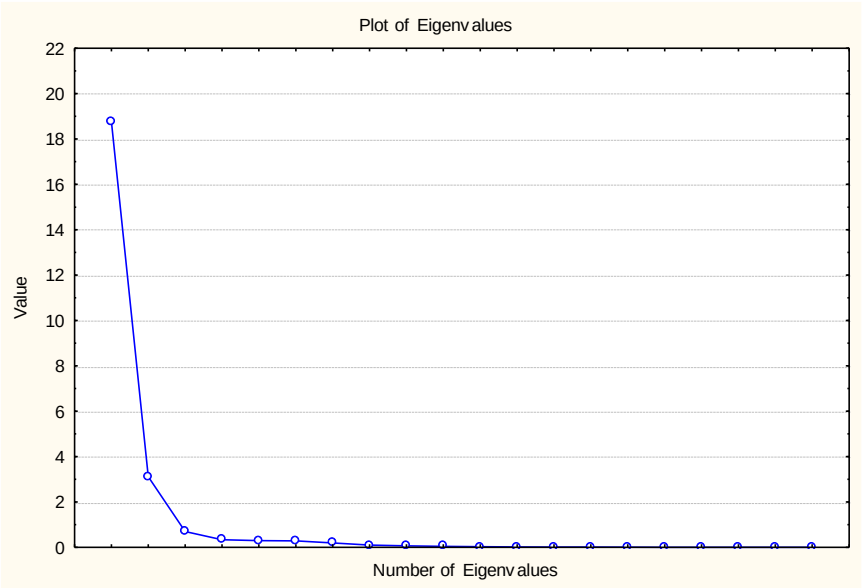


Рис.18. Діаграма власних значень

4. **Кластерний аналіз. Розбиття на три кластери.**
   Виконати команди: Statistics, Multivariate Exploratory Techniques, Cluster Analysis, K-means clustering, Number of clusters – 3, Select Variables for the Analysis – Select All, Graph of means.
5. Прокоментувати отриманий графік (рис. 17). Описати поведінку середніх значень для кожного з трьох кластерів.
6. Виконати команду Advanced, Members of each cluster & distances. Виписати склад кожного з трьох кластерів та відстань кожного елемента кластеру від його центра.
7. **Кластерний аналіз. Розмірність системи.**
   Виконати команди: Statistics, Multivariate Exploratory Techniques, Factor Analysis, Variables – Select All, Principal components, Scree plot. Проаналізувати отриману діаграму (рис. 18) зробити висновок щодо розмірності системи.
8. Виконати команди: Statistics, Multivariate Exploratory Techniques, Factor Analysis, Variables – Select All, Principal components, Loadings,

Factor rotation – Unrotated, Plot oh Loadings 2D. Проаналізувати отриманий графік (рис. 19) та описати отримані кластери.

9. Шляхом редагування осей виділити на попередньому графіку скупчення об’єктів, які накладалися один на одного та збільшити його. Прокоментувати отриманий фрагмент діаграми (рис. 20).
10. Виконати команди: Statistics, Multivariate Exploratory Techniques, Factor Analysis, Variables – Select All, Principal components, Loadings, Factor rotation – Varimax normalized, Plot oh Loadings 2D. Проаналізувати отриманий графік та описати отримані кластери.
11. Шляхом редагування осей виділити на попередньому графіку скупчення об’єктів, які накладалися один на одного та збільшити його. Прокоментувати отриманий фрагмент діаграми.

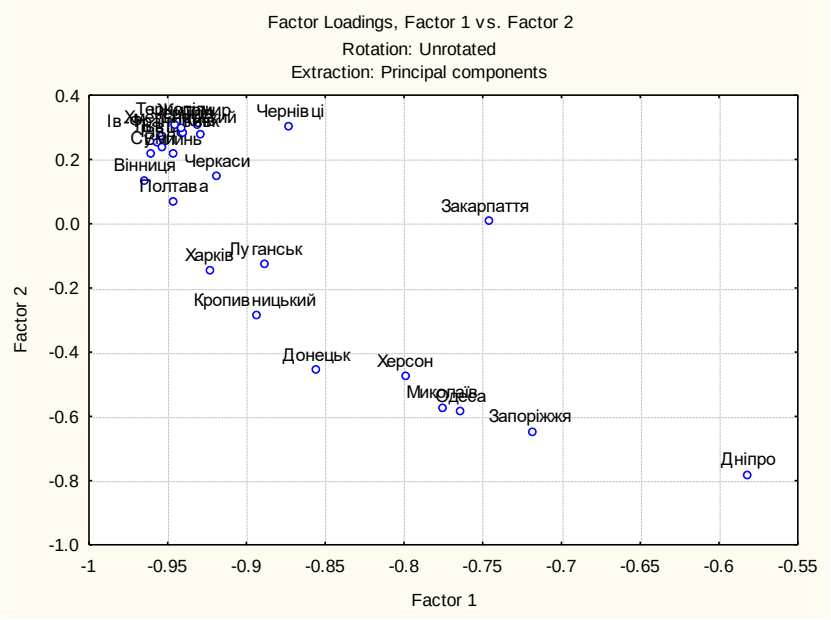


Рис. 19. Кластерний аналіз. Діаграма на осях головних координат

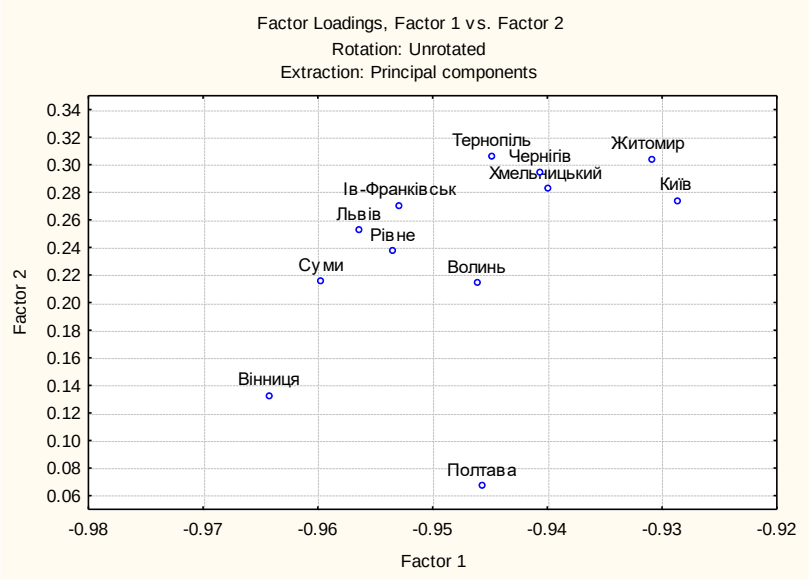


Рис. 20. Кластерний аналіз. Діаграма на осях головних координат

Література

1. Бахрушин В. Є. Методи аналізу даних : навч. посібник. Запоріжжя : КПУ, 2011, 268 с.
2. Грицюк П. М., Остапчук О. П. Аналіз даних : навчальний посібник. Рівне : НУВГП, 2008. 218 с.
3. Гороховатський В. О., Творошенко І. С. Методи інтелектуального аналізу та оброблення даних : навч. посіб. / М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. Харків : ХНУРЕ, 2021. 92 с.
4. Паянок Т. М., Задорожня Т. М. Статистичний аналіз даних : навчальний посібник. Ірпінь : Університет державної фіскальної служби України, 2020. 312 с.