

Міністерство освіти і науки України
Національний університет «Запорізька політехніка»

МЕТОДИЧНІ ВКАЗІВКИ
до виконання лабораторних робіт
з дисципліни
**«Емпіричні методи в
інформаційних технологіях»**

2023

Методичні вказівки до виконання лабораторних робіт з дисципліни “Емпіричні методи в інформаційних технологіях” / Укл. А.О. Олійник, С.Д. Леощенко, М.О. Андреєв, Є.М. Федорченко. – Запоріжжя: НУ «Запорізька політехніка», 2023. – 145 с.

Укладачі: А. О. Олійник, д.т.н., професор
С. Д. Леощенко, доктор філософії, ст. викл.
М. О. Андреєв, асистент
Є. М. Федорченко, старший викладач

Рецензент: В.М. Льовкін, к.т.н., доцент

Відповідальний
за випуск: С. О. Субботін, д.т.н., професор

Затверджено
на засіданні кафедри
програмних засобів

Протокол № 1
від “17” серпня 2023 р.

ЗМІСТ

Вступ	7
1 Лабораторна робота № 1 Статистичний аналіз і первинна обробка даних	8
1.1 Мета роботи	8
1.2 Короткі теоретичні відомості	8
1.3 Приклад обчислення характеристик випадкової величини	12
1.4 Відомості про програмні пакети та мови програмування для опрацювання даних	14
1.4.1 Пакет Statgraphics	14
1.4.1.1 Загальні відомості про склад пакета Statgraphics	14
1.4.2 Мова програмування R та середовище розробки R-Studio	16
1.5 Порядок виконання роботи	23
1.5.1 Приклад статистичного аналізу даних у пакеті Statgraphics	23
1.5.2 Первинна обробка даних з використанням мови R та середовища R-Studio	25
1.5.2.1 Основні функції та команди, необхідні для виконання роботи	25
1.5.2.2 Приклад реалізації у R-Studio	27
1.6 Завдання на лабораторну роботу	32
1.7 Зміст звіту	33
1.8 Контрольні запитання	33
2 Лабораторна робота № 2 Статистична перевірка гіпотез	34
2.1 Мета роботи	34
2.2 Короткі теоретичні відомості	34
2.2.1 Нормальний розподіл	34
2.2.2 Перевірка гіпотези про розподіл випадкової величини	36
2.2.3 Приклади перевірки гіпотез та побудови довірчих інтервалів ...	39
2.3 Порядок виконання роботи	42
2.3.1 Виконання у пакеті Statgraphics	42
2.3.2 Статистична перевірка гіпотез засобами мови R	45
2.3.2.1 Перевірка гіпотез у R-Studio для однієї вибірки	46
2.3.2.2 Перевірка гіпотез у R-Studio для двох вибірок	48
2.4 Завдання на лабораторну роботу	49
2.5 Зміст звіту	50

2.6 Контрольні запитання	50
3 Лабораторна робота № 3 Дисперсійний аналіз	51
3.1 Мета роботи	51
3.2 Короткі теоретичні відомості	51
3.2.1 Постанова завдання	51
3.2.2 Дисперсійний аналіз.....	53
3.2.3 Зв'язок двохфакторного й однофакторного аналізу	56
3.2.4 Таблиця вихідних даних для двохфакторного аналізу	57
3.2.5 Адитивна модель даних двохфакторного експерименту при незалежній дії факторів.....	58
3.2.6 Двохфакторний дисперсійний аналіз.....	59
3.2.7 Порядок виконання роботи.....	61
3.2.7.1 Порядок виконання однофакторного дисперсійного аналізу в пакеті Statgraphics	61
3.2.7.2 Порядок виконання двохфакторного дисперсійного аналізу в мові програмування R та середовища R-Studio	68
3.3 Завдання на лабораторну роботу.....	78
3.4 Зміст звіту.....	79
3.5 Контрольні запитання	79
4 Лабораторна робота № 4 Методи вивчення взаємозв'язків	80
4.1 Мета роботи	80
4.2 Короткі теоретичні відомості	80
4.2.1 Аналіз парних зв'язків: коефіцієнт парної кореляції	80
4.2.2 Вибіркове значення коефіцієнта кореляції	81
4.2.3 Кореляційне відношення.....	82
4.2.4 Перевірка гіпотези про відсутність кореляційного зв'язку	83
4.2.5 Лінійний регресійний аналіз.....	84
4.2.6 Метод найменших квадратів	85
4.2.7 Перевірка гіпотези про коефіцієнт нахилу.....	86
4.2.8 Аналіз рівняння регресії	87
4.2.9 Приклад побудови рівняння регресії	88
4.3 Порядок виконання роботи.....	91
4.3.1 Постанова завдання	91
4.3.2 Розв'язання задачі.....	92
4.3.3 Результати	94
4.3.4 Відповідність моделі	96
4.3.5 Графічні результати.....	96

4.3.6 Кореляційний та регресійний аналіз з використанням мови R ...	97
4.3.6.1 Основні функції та команди, необхідні для виконання роботи	97
4.3.6.2 Розрахунок коефіцієнту кореляції у R-Studio	99
4.3.6.3 Побудова лінійної регресії засобами мови R	100
4.4 Завдання на лабораторну роботу.....	101
4.5 Зміст звіту.....	102
4.6 Контрольні запитання	102
5 Лабораторна робота № 5 Ранговий аналіз.....	104
5.1 Мета роботи	104
5.2 Короткі теоретичні відомості	104
5.2.1 Характеристика ознак	104
5.2.2 Поняття рангової кореляції.....	105
5.2.3 Ранговий коефіцієнт кореляції Спірмена	105
5.2.4 Ранговий коефіцієнт кореляції Кендела	107
5.2.5 Ранговий однофакторний аналіз	108
5.2.6 Критерій Краскела-Уолліса	109
5.2.7 Практичний приклад	111
5.3 Порядок виконання роботи.....	116
5.4 Завдання на лабораторну роботу.....	120
5.5 Зміст звіту.....	121
5.6 Контрольні запитання	121
6 Лабораторна робота №6 Повний факторний експеримент	122
6.1 Мета роботи	122
6.2 Короткі теоретичні відомості	122
6.2.1 Постановка завдання	122
6.2.2 Вибір параметру оптимізації	124
6.2.3 Вибір факторів	124
6.2.4 Вибір координат базової точки	125
6.2.5 Вибір ступенів варіювання	125
6.2.6 Отримання матриці планування. Визначення координат точок для пробних експериментів	125
6.2.7 Проведення експериментів у запланованих точках. Рандомізація дослідів	128
6.2.8 Обчислення коефіцієнтів рівняння регресії	129
6.2.9 Статистична оцінка значущості коефіцієнтів рівняння регресії.....	130
6.2.10 Статистична перевірка адекватності рівняння регресії	131

6.2.11 Використання нормованого рівняння регресії для передбачення цільової функції.....	132
6.2.12 Приклад	133
6.3 Порядок виконання роботи.....	135
6.3.1 Реалізація ПФЕ у середовищі Statgraphics	135
6.3.2 Програмна реалізація повного факторного експерименту засобами мови R	140
6.4 Завдання на лабораторну роботу.....	142
6.5 Зміст звіту.....	142
6.6 Контрольні запитання	142
Література.....	144

ВСТУП

Дане видання призначене для вивчення та практичного освоєння студентами усіх форм навчання основ емпіричних методів в інформаційних технологіях.

Відповідно до графіка студенти перед виконанням лабораторної роботи повинні ознайомитися з конспектом лекцій та рекомендованою літературою.

Для одержання заліку з кожної роботи студент здає викладачу оформлений звіт, а також демонструє на екрані комп'ютера результати виконання лабораторної роботи.

Звіт має містити:

- титульний аркуш;
- тему та мету роботи;
- завдання до роботи;
- лаконічний опис теоретичних відомостей;
- результати виконання лабораторної роботи;
- змістовний аналіз отриманих результатів та висновки.

Звіт виконують на білому папері формату А4 (210 × 297 мм) або подають в електронному вигляді.

Під час співбесіди при захисті лабораторної роботи студент повинний виявити знання про зміст роботи та методи виконання кожного етапу роботи, а також вміти продемонструвати результати роботи на конкретних прикладах. Студент повинний вміти правильно аналізувати отримані результати. Для самоперевірки при підготовці до виконання і захисту роботи студент повинен відповісти на контрольні запитання, наведені наприкінці опису відповідної роботи.

1 ЛАБОРАТОРНА РОБОТА № 1 СТАТИСТИЧНИЙ АНАЛІЗ І ПЕРВИННА ОБРОБКА ДАНИХ

1.1 Мета роботи

Ознайомитися з можливостями пакетів статистичної обробки даних. Навчитися використовувати пакети статистичної обробки даних для первинного аналізу даних.

1.2 Короткі теоретичні відомості

Числові характеристики розподілу ймовірностей допомагають скласти наочне уявлення про цей розподіл. Основними характеристиками розподілу ймовірностей випадкової величини служать моменти та квантілі.

Перший момент випадкової величини X називається математичним очікуванням або середнім значенням. Для дискретної випадкової величини X із значеннями x_1, x_2, \dots що мають ймовірності p_1, p_2, \dots математичне очікування дорівнює:

$$M(X) = \sum x_i p_i \quad (1.1)$$

Для неперервної випадкової величини X із щільністю $\varphi(x)$

$$M(X) = \int_{-\infty}^{\infty} x \varphi(x) dx \quad (1.2)$$

причому інтеграл повинен збігатися абсолютно.

Середнє значення випадкової величини у певному розумінні характеризує центр розподілу ймовірностей. Для характеристики (кількісного опису) розкиду випадкової величини відносно цього центру в теорії ймовірностей використовують другий центральний момент випадкової величини. Його називають дисперсією та, як правило, позначають $\sigma^2 = D(X)$. Дисперсія $D(X)$ випадкової величини – це матема-

тичне сподівання квадрату відхилень значення випадкової величини від її математичного очікування:

$$D(X) = M[(x - M(X))^2]. \quad (1.3)$$

Якщо необхідно, щоб показник розкиду випадкової величини був виражений у тих же одиницях, що й значення випадкової величини, то замість $D(X)$ використовують величину $\sqrt{D(X)}$, яка називається середнім квадратичним відхиленням або стандартним відхиленням випадкової величини X .

Чим більше дисперсія, тим ширше діапазон розсіювання точок, що зображують випадкові числа на числовій вісі.

Центральні моменти не змінюються при додаванні до випадкової величини постійного додатку, тобто вони не залежать від вибору початку відліку на шкалі виміру випадкової величини. Але від обраної одиниці вимірювання залежність залишається. В таких випадках, щоб усунути подібний вплив, моменти тим чи іншим способом нормують, поділивши їх на відповідну ступінь середньоквадратичного відхилення. В результаті отримаємо безрозмірну величину, що не залежить від вибору початку відліку та одиницю вимірювання вихідної випадкової величини. Частіше за все з нормування моментів використовуються асиметрія та ексцес – відповідно третій та четвертий нормовані центральні моменти:

$$\text{асиметрія} = \frac{M[x - M(X)]^3}{(\sqrt{D(X)})^3}, \quad (1.4)$$

$$\text{ексцес} = \frac{M[x - M(X)]^4}{(\sqrt{D(X)})^4}. \quad (1.5)$$

Асиметрія характеризує несиметричність розподілу випадкової величини, а ексцес – ступінь виразності "хвостів" розподілу, тобто частоту появи значень, віддалених від середнього. Якщо у варіаційному ряді переважають варіанти, менші \bar{x} , то емпіричний коефіцієнт асиметрії від'ємний; кажуть, що в цьому випадку має місце лівобічна асиметрія, інакше – правобічна. При лівобічній асиметрії лівий "хвіст" полігону довший за правий. При правобічній, більш довгим є правий

“хвіст”. Криві, у яких ексцес є від’ємним, у порівнянні з нормальною кривою менш круті, мають більш плоску вершину. Криві з додатним ексцесом більш круті в порівнянні з нормальною кривою, мають гостру вершину.

Інколи значення асиметрії та ексцесу використовують для перевірки гіпотези про те, що дані, що спостерігаються, підкоряються заданому закону розподілу. Так, для будь якого нормального закону розподілу асиметрія дорівнює нулю, а ексцес – трьом.

Квантілем p випадкової величини, яка має функцію розподілу $\varphi(x)$ називається розв'язок x_p рівняння $\varphi(x) = p$. Величину x_p називають квантілем рівня p розподілу $\varphi(x)$. Основні квантілі – медіана, яка відповідає значенню $p = 0,5$, верхній та нижній квантілі відповідають значенню $p = 0,75$ та $p = 0,25$ відповідно.

Одним з найпростіших методів статистичного аналізу є побудова ряду розподілу.

Ряд розподілу (варіаційний або статичний ряд) – це таблиця, в якій перераховані та вказані границі i -х інтервалів можливих значень випадкової змінної x та відповідні їм ймовірності P_i появи x в i -х інтервалах. Якщо невідомі ймовірності P_i , то вказують абсолютні емпіричні частоти m'_i , тобто число елементів статичної сукупності для x , що опинилися в i -х інтервалах, або відносні частоти ν_i , що розраховуються за формулою

$$\nu_i = \frac{m'_i}{N}, \quad (1.6)$$

де m'_i – кількість елементів, що потрапили в i -й інтервал;

N – загальна кількість елементів досліджуваної сукупності.

Ширину інтервалів δ_x , як правило, приймають постійною (або необхідно кожний раз враховувати удільну вагу ширини інтервалу) для всіх інтервалів і розраховують за формулою:

$$\delta_x = \frac{x_{\max} - x_{\min}}{k}, \quad (1.7)$$

де x_{\max} – найбільше значення x в даній сукупності; x_{\min} – найменше серед всіх x ; k – кількість інтервалів:

$$k \approx 1 + 3.2 \lg n. \quad (1.8)$$

Число k , що розраховане за формулою (1.8), округлюють до найближчого цілого. Як правило, $k = 6 \dots 12$. При невеликій кількості k інтервалів можна пропустити характерні особливості кривої розподілу, а при дуже великому значенні k і порівняно малому значенню N навіть у середні інтервали потрапляє мало елементів статичної сукупності, і результати розрахунків будуть мати велику похибку.

Границі інтервалів рекомендується вибирати наступним чином. Для статичної сукупності з N елементів необхідно розрахувати середнє значення \bar{x} , потім ширину інтервалів за формулою (1.8). Далі побудувати числову вісь і відмітити на ній середнє значення \bar{x} . По обидві сторони від \bar{x} відкласти спочатку по половині інтервалу $\delta_x/2$, а потім – по цілому інтервалу δ_x , доки крайні інтервали не перекриють x_{\max} та x_{\min} . Подібний спосіб розбиття полегшує подальші розрахунки.

Границі зліва позначаються круглою дужкою, а справа – квадратною.

Квадратною дужкою будемо позначати закрити (тобто включаючи те число, яким позначена границя), а круглою – відкрити границю інтервала (тобто виключаючи позначене число). Це значить, що якщо один із елементів сукупності потрапив на границю, то його слід відносити до лівого інтервала, якщо праві границі інтервалів закриті.

Обчислимо кількість елементів m'_i , що потрапили в кожен i -й інтервал:

$$\sum_{i=1}^k m'_i = N \quad (1.10)$$

або

$$\sum_{i=1}^k \nu_i = 1, \quad (1.11)$$

де i – номер інтервалів.

Гістограма – графічне подання ряду розподілу, де по осі x відкладаються границі інтервалів, а по осі y – частота m'_i .

1.3 Приклад обчислення характеристик випадкової величини

Задана вибірка:

6,75; 6,77; 6,77; 6,73; 6,76; 6,74; 6,70; 6,75; 6,71; 6,72; 6,77; 6,79;
 6,71; 6,78; 6,73; 6,70; 6,73; 6,77; 6,75; 6,74; 6,71; 6,70; 6,78; 6,76; 6,81;
 6,69; 6,80; 6,80; 6,77; 6,68; 6,74; 6,70; 6,70; 6,74; 6,77; 6,83; 6,76; 6,76;
 6,82; 6,77; 6,71; 6,74; 6,70; 6,75; 6,74; 6,75; 6,77; 6,72; 6,74; 6,80; 6,75;
 6,80; 6,72; 6,78; 6,70; 6,75; 6,78; 6,78; 6,76; 6,77; 6,74; 6,74; 6,77; 6,73;
 6,74; 6,77; 6,74; 6,75; 6,74; 6,76; 6,76; 6,74; 6,74; 6,74; 6,76; 6,74;
 6,72; 6,80; 6,76; 6,78; 6,73; 6,70; 6,76; 6,76; 6,77; 6,75; 6,78; 6,72; 6,76;
 6,78; 6,68; 6,75; 6,73; 6,82; 6,73; 6,80; 6,81; 6,71; 6,82; 6,77; 6,80; 6,80;
 6,70; 6,70; 6,82; 6,72; 6,69; 6,73; 6,76; 6,74; 6,77; 6,72; 6,76; 6,78; 6,73;
 6,76; 6,80; 6,76; 6,72; 6,76; 6,76; 6,70; 6,73; 6,75; 6,77; 6,77; 6,70; 6,81;
 6,74; 6,73; 6,77; 6,74; 6,78; 6,69; 6,74; 6,71; 6,76; 6,76; 6,77; 6,70; 6,81;
 6,74; 6,74; 6,77; 6,75; 6,80; 6,74; 6,76; 6,77; 6,77; 6,81; 6,75; 6,78; 6,73;
 6,76; 6,76; 6,76; 6,77; 6,76; 6,80; 6,77; 6,74; 6,77; 6,72; 6,75; 6,76; 6,77;
 6,81; 6,76; 6,76; 6,76; 6,80; 6,74; 6,80; 6,74; 6,73; 6,75; 6,77; 6,74; 6,76;
 6,77; 6,77; 6,75; 6,76; 6,74; 6,82; 6,76; 6,73; 6,74; 6,75; 6,76; 6,72; 6,78;
 6,72; 6,76; 6,77; 6,75; 6,78.

За заданою вибіркою визначити оптимальну кількість інтервалів для розбиття за формулою Стерджеса. Традиційним способом розрахувати: математичне сподівання; дисперсію; середньоквадратичне відхилення; розмах.

Побудувати таблицю, яка містить наступні стовпці: нижня і верхня границя інтервалу; частота попадання в інтервал; відносна частота; накопичувана частота; накопичувана відносна частота.

За отриманою таблицею побудувати гістограму.

Розв'язок

Знайдемо $x_{\max} = 6,83$; $x_{\min} = 6,68$. Оптимальну кількість інтервалів визначимо за формулою (1.8):

$$h = (1 + 3,322 \lg 200) = 1 + 3,322 \lg 200 \approx 8,64 \approx 9.$$

Величину інтервалу визначимо за формулою:

$$h = (6,83 - 6,68) / (1 + 3,322 \lg 200) \approx 0,15 / 8,64 \approx 0,02.$$

За нижню границю інтервалу приймаємо величину $a_1 = 6,68 - 0,01 = 6,67$. Тоді $a_2 = 6,67 + 0,02 = 6,69$; $a_3 = 6,71$ і т.д. Шкала інтервалів та групування результатів спостережень наведені у таблиці 1.1.

Таблиця 1.1 – Інтервали та частоти

Границі інтервалів (нижня границя – верхня границя), мм	Частота m	Накопичувана частота $m_{\text{нак}}$	Відносна частота m/N	Відносна накопичувана частота $m_{\text{нак}} / N$
6,67-6,69	2	2	0,01	0,01
6,69-6,71	15	17	0,075	0,085
6,71-6,73	17	34	0,085	0,17
6,73-6,75	44	78	0,22	0,39
6,75-6,77	52	130	0,26	0,65
6,77-6,79	44	174	0,22	0,87
6,79-6,81	14	188	0,07	0,94
6,81-6,83	11	199	0,055	0,995
6,83-6,85	1	200	0,005	1
Σ	200		1	

Побудуємо гістограму (рис. 1.1):

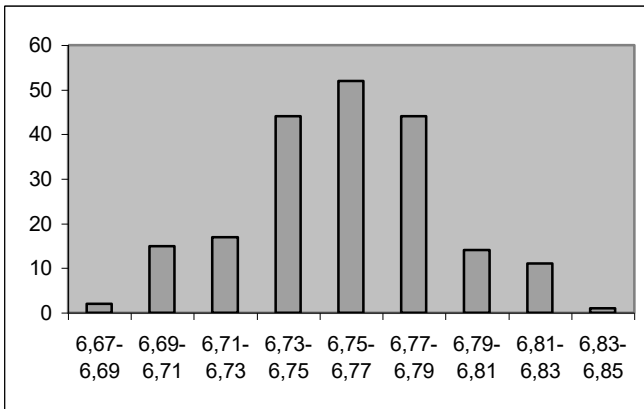


Рисунок 1.1 – Гістограма частот

Розрахуємо розмах: $R = x_{\max} - x_{\min} = 6,83 - 6,68 = 0,15$.

Математичне сподівання: $M\{x\} = \frac{1}{N} \sum_{i=1}^N x_i = 6,7533$.

Дисперсія: $S^2\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = 0,00097$.

Середньоквадратичне відхилення: $S\{x\} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} = 0,0311$.

1.4 Відомості про програмні пакети та мови програмування для опрацювання даних

1.4.1 Пакет Statgraphics

Неоціненну допомогу при аналізі статистичних даних може зробити статистичний графічний пакет *Statgraphics (Statistical Graphics System)*. Цей пакет відрізняється повнотою представлених статистичних методів, різноманітними графічними засобами, широкими можливостями оперування даними.

Пакет розрахований на фахівців, що добре знайомі із процедурами статистичного аналізу. Пакет *Statgraphics* є універсальним пакетом, що містить більшість стандартних статистичних методів.

У пакет включено більш 250 процедур обробки даних по наступних розділах математичної статистики:

- дисперсійний аналіз;
- аналіз часових рядів;
- описова статистика;
- контроль якості;
- багатомірний аналіз;
- непараметричний аналіз;
- планування експерименту;
- підбір розподілів;
- регресійний аналіз;
- лінійне програмування.

1.4.1.1 Загальні відомості про склад пакета Statgraphics

Пакет *Statgraphics* містить широко розгалужену ієрархічну систему меню. Кожний пункт головного меню включає підменю. У базовій системі функціонують наступні процедури:

- меню *Describe* складається зі статистичних методів аналізу по одній та більшій кількості змінних, процедури підбору розподілів, засобів табуляції даних і кросс-табуляції даних;

- меню *Compare* включає засоби порівняння двох і більшої кількості вибірок даних, процедури одно- та багатофакторного дисперсійного аналізу;

- меню *Relate* містить процедури простого, поліноміального і множинного регресійного аналізу;

Для розширення можливостей системи пропонуються додаткові модулі, ініціалізація яких відбувається через меню *Special*. До них належать:

- модуль *Quality Control (Контроль якості)* призначений для оцінки ефективності виробничого процесу і формування відповідних контрольних карт. У модулі добре організовані процедури для конструювання Парето - карт, аналізу можливостей процесу і побудови X - та R - контрольних карт;

- модуль *Experimental Design (Планування експерименту)* допомагає сформулювати критерій оптимальності плану експерименту, підібрати найкращий план, організувати збір та обробку потрібної інформації. Процедура взаємодії з модулем наступна: визначення факторів, вибір плану, генерація робочої таблиці для збору і запису даних, підбір моделі, інтерпретація результатів;

- модуль *Time-Series Analysis (Аналіз часових рядів)* містить описові методи, процедури згладжування рядів, сезонної декомпозиції і прогнозування. Даний модуль допомагає побачити чисту картину динамічних даних. Доцільно розпочати роботу з описових методів, для того щоб отримати перше візуальне представлення. Далі можна зробити більш точний опис динамічного ряду з урахуванням сезонних ефектів, циклічних змін, трендів, помилок, викидів або точок злому даних. Результати подаються в табличній формі або на зручних до сприйняття графіках;

- модуль *Multivariate Methods (Багатомірні методи)* призначений для вивчення і розкриття взаємовідносин множини факторів (змінних). Якщо користувач займається дослідженнями у фізиці, соціології, медицині чи інших галузях, де об'єкти досліджень характеризуються великою кількістю ознак, даний модуль допоможе сортувати і групувати дані, визначити відносини між змінними, висунути і перевірити різноманітні гіпотези. Для цього в модулі функціонують п'ять

потужних процедур, що забезпечують проведення кластерного аналізу, аналізу по методу головних компонент, факторного, дискримінантного і канонічного кореляційного аналізу;

– модуль *Advanced Regression* (*Розширений регресійний аналіз*) окрім базисних процедур регресійного аналізу включає різноманітні калібровочні моделі, процедури порівняння ліній регресії, відбору найкращих регресійних моделей, нелінійну множинну регресію, ридж-регресію і логістичну регресію.

1.4.2 Мова програмування R та середовище розробки R-Studio

R – це мова програмування й середовище для статистичних обчислень і графічного аналізу. R є мовою для аналізу даних із відкритим кодом, яка підтримується великою та активною дослідницькою спільнотою у всьому світі.

У R є багато особливостей та позитивних характеристик, які дозволяють рекомендувати саме її:

– R розповсюджується вільно, а середовище розробки R-Studio – вільно та безкоштовно;

– R має потужний статистичний апарат, у якому реалізовані всі способи аналізу даних;

– R має сучасні графічні можливості. Для візуалізації складних даних в R реалізовані різноманітні й потужні методи аналізу даних;

– отримання даних із різних джерел у доступному для використання вигляді може бути складним завданням. В R є можливість імпортувати дані з різних джерел, включно з текстовими файлами, системами управління базами даних, іншими статистичними програмами та спеціалізованими сховищами даних. R може також записувати дані у форматах усіх цих систем;

– R є платформою для простого написання програм, що реалізують нові статистичні методи;

– потужна спільнота та часте оновлення бібліотек та функцій.

Завантажити та встановити R останньої актуальної версії можна з офіційного сайту (рис. 1.2): <https://www.r-project.org/>



Рисунок 1.2 – Головна сторінка сайту R Project

Завантажити та встановити середовище розробки R-Studio останньої актуальної версії можна з офіційного сайту (рис. 1.3): <https://rstudio.com/products/rstudio/download/#download>

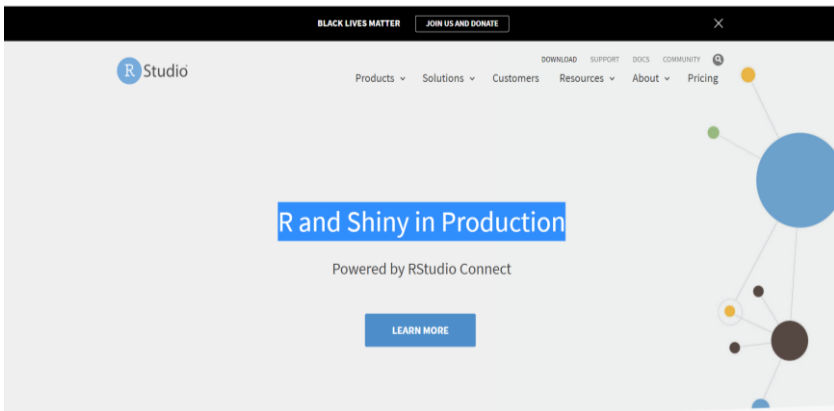


Рисунок 1.3 – Головна сторінка сайту R-Studio

RStudio є безкоштовним інтегрованим середовищем розробки (IDE) для R (рис.1.4). Цей програмний продукт з відкритим вихідним кодом для мови програмування R, яка призначена для статистичної обробки даних і роботи з графікою, робить роботу з R дуже зручною.

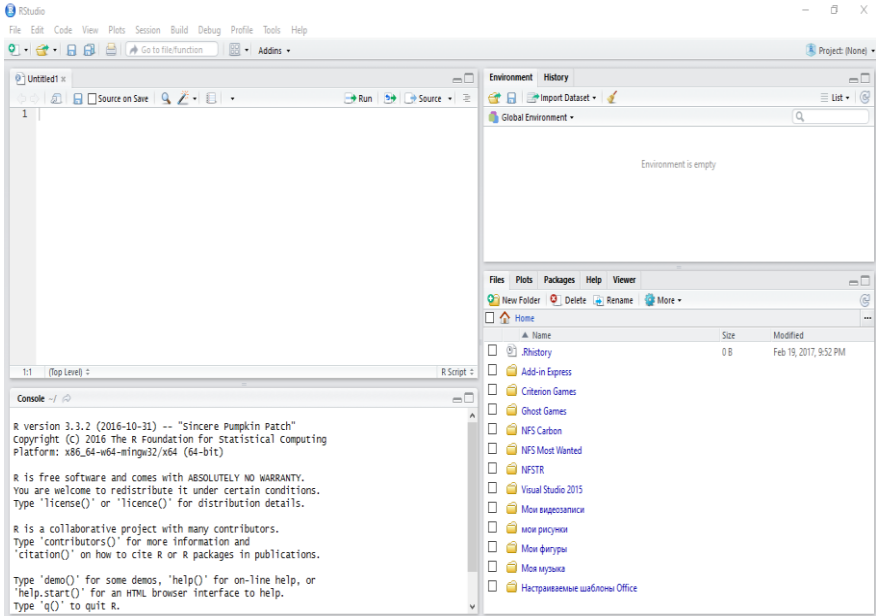


Рисунок 1.4 – Середовище розробки R-Studio

Середовище розробки R-Studio складається з наступних компонентів:

- панель меню;
- панель інструментів;
- консоль;
- запрошення (командний рядок);
- панель, що містить історію і робочий простір;
- панель з графіками.

Консоль RStudio надає цілий ряд опцій, що полегшують роботу з мовою R. Пропонуємо ознайомитися з ними нижче. Наприклад, автоматичне завершення коду: набираючи початок команди середовище пропонує користувачу продовження (рис.1.5).

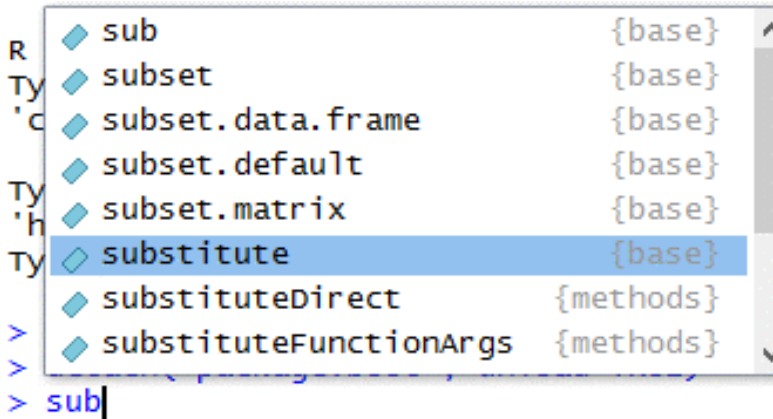


Рисунок 1.5 – Демонстрація автоматичного завершення коду в консолі

Повернення до попередніх команд. За допомогою сполучень клавіш `Ctrl + ↑` та `Ctrl + ↓` можна переходити до раніше викланих команд (рис. 1.6).

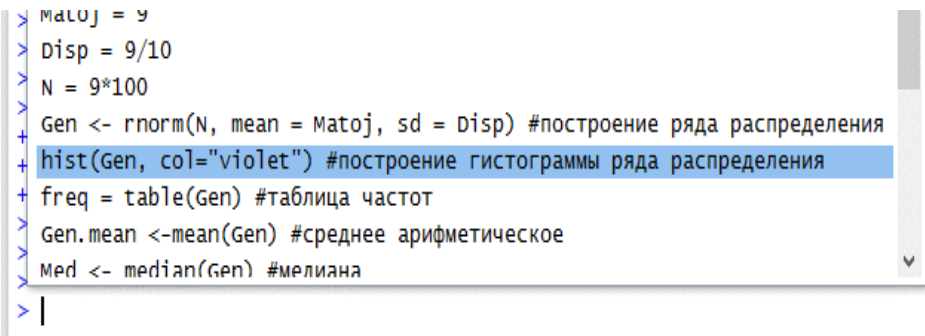


Рисунок 1.6 – Перехід до раніше викланих команд в консолі

«Гарячі» клавіші. Переглянути гарячі клавіші можливо викликавши `HELP – R HELP`, або обрати потрібний пункт меню `Code, View` або інше (рис.1.7).

Move Focus to Source	Ctrl+1
Move Focus to Console	Ctrl+2
Move Focus to Help	Ctrl+3
<hr/>	
Show History	Ctrl+4
Show Files	Ctrl+5
Show Plots	Ctrl+6
Show Packages	Ctrl+7
Show Environment	Ctrl+8
Show Viewer	Ctrl+9
Show Connections	Ctrl+F5

Рисунок 1.7 – Список «гарячих» клавіш пункту меню View

Створення нових скриптів. Для створення нових скриптів необхідно обрати File – New File → R Script (рис. 1.8). Якщо проект буде мати декілька R файлів, то спочатку New Project, а потім New File.

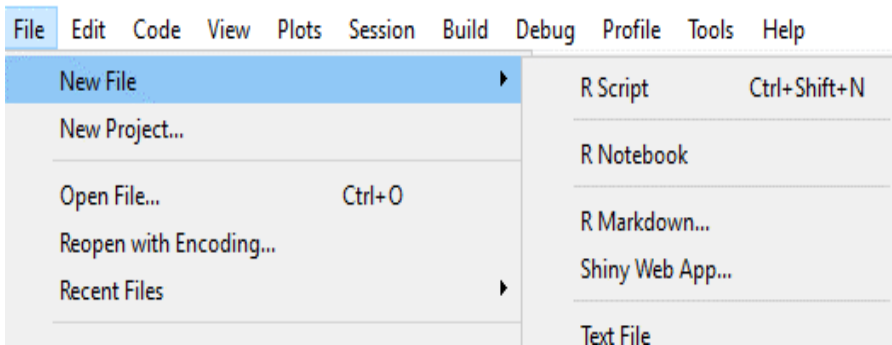


Рисунок 1.8 – Створення нового файлу

Створення функцій в коді. Щоб створити функцію треба виділити частину коду та обрати пункт Extract Function (рис.1.9).

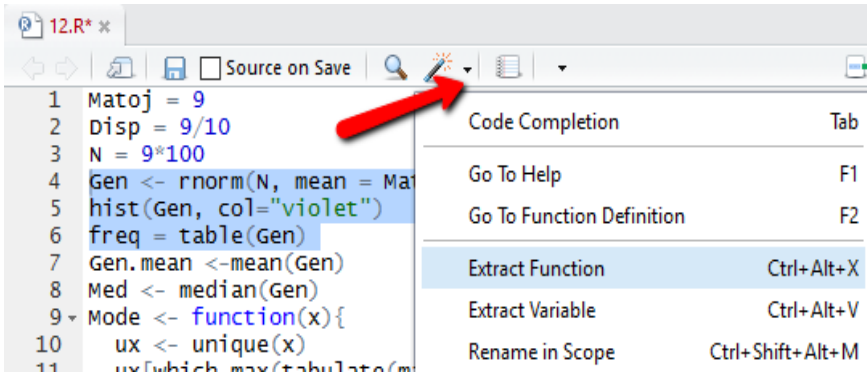


Рисунок 1.9 – Створення функції з виділеного коду

Пошук та заміна частин коду. Щоб знайти та/або замінити частину тексту треба викликати вікно пошуку за допомогою меню Edit -> Find and Replace, або Ctrl + F (рис. 1.10).

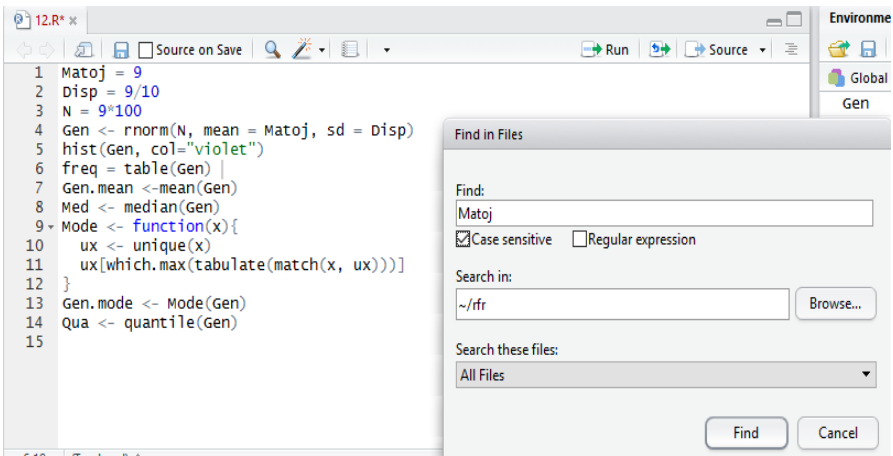


Рисунок 1.10 – Пошук частини коду в тексті

Коментування коду. Для коментування/зняття коментування з фрагменту коду – виділити його та натиснути пункт Comment/Uncomment Lines (рис. 1.11).

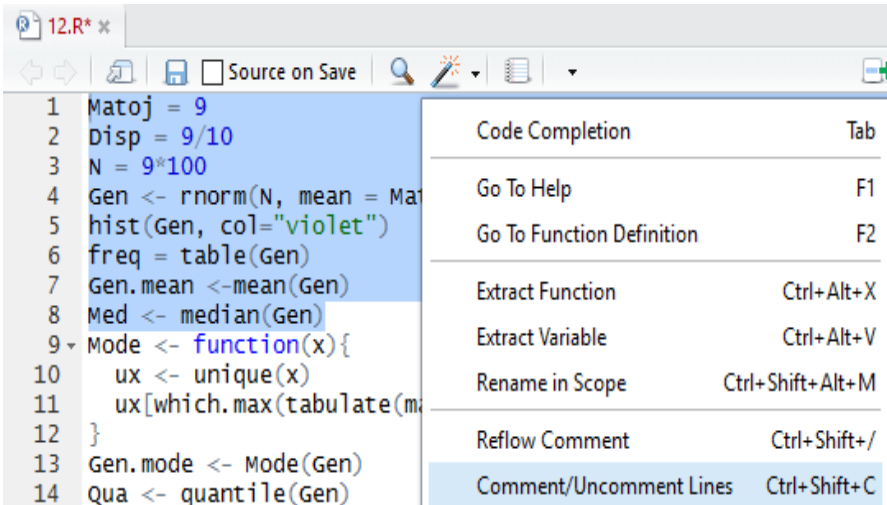


Рисунок 1.11 – Коментування коду

Виконання коду. Щоб виконати код – виділіть потрібну частину та натисніть Run. Для виконання поточного рядку – натисніть Ctrl + Enter. Після цього редактор автоматично перейде на наступний рядок. Для виконання усіх рядків – натисніть Ctrl + Shift + Enter.

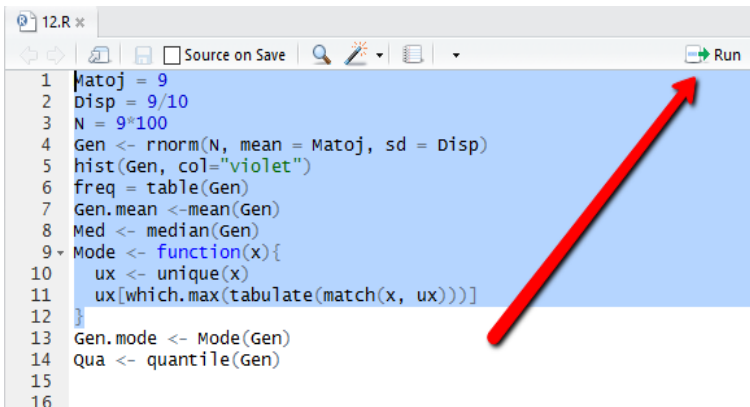


Рисунок 1.12 – Виконання коду

1.5 Порядок виконання роботи

1.5.1 Приклад статистичного аналізу даних у пакеті Statgraphics

Розглянемо приклад виконання аналізу в пакеті статистичної обробки даних Statgraphics. Для чого треба згенерувати дані для нормального розподілу з наступними значеннями: **RNORMAL(100;50;4)**, де 100 – кількість значень, 50 – математичне очікування, 4 – дисперсія. Далі необхідно вибрати пункт **Describe – Numeric Data – One-Variable Analysis**, натиснути на передньому плані кнопку **Tabular Option** та вибрати перші 4 пункти із запропонованих (рис.1.13): **Analysis summary** – аналіз вибірки, **Summary Statistics for Col_1** – сумарна статистика для змінної, **Percentiles for Col_1** – відсоткова статистика, **Frequency Tabulation** – таблиця частот (рис. 1.14).

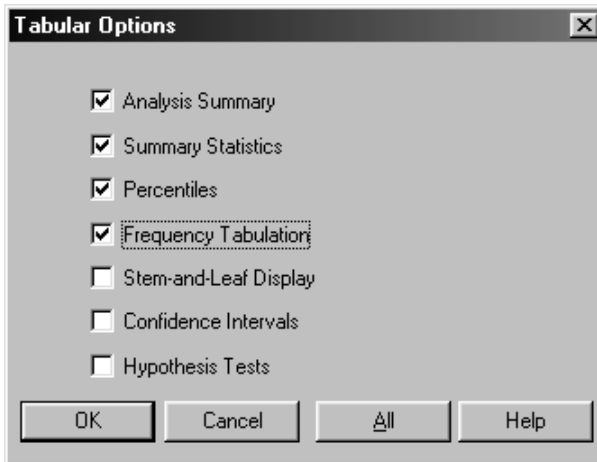


Рисунок 1.13 – Список опцій

Пункти Lower Limit and Upper Limit визначають межі кожного з інтервалів, Midpoint – його середнє значення. Пункт Frequency відображає кількість значень, що знаходяться в межах кожного з інтервалів.

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		38,0		0	0,0000	0	0,00
1	38,0	41,0	39,5	2	0,0200	2	0,02
2	41,0	44,0	42,5	4	0,0400	6	0,06
3	44,0	47,0	45,5	15	0,1500	21	0,21
4	47,0	50,0	48,5	27	0,2700	48	0,48
5	50,0	53,0	51,5	30	0,3000	78	0,78
6	53,0	56,0	54,5	17	0,1700	95	0,95
7	56,0	59,0	57,5	3	0,0300	98	0,98
8	59,0	62,0	60,5	2	0,0200	100	1,00
above	62,0			0	0,0000	100	1,00

Mean = 50,0576 Standard deviation = 3,96222

Рисунок 1.14 – Таблица частот для Col_1

Relative Frequency – це відносна частота, яка отримується, якщо попередні табличні значення поділити на кількість спостережень. **Cumulative Frequency** – це накопичена частота, отримана послідовним складанням відповідних частот. Аналогічно розраховується накопичена відносна частота (**Cum. Rel. Frequency**).

Далі треба натиснути на пункті меню **Graphical Option** (кнопка з зображенням графіка) і вибрати з наданого списку пункт **Frequency Histogram** (рис. 1.15).

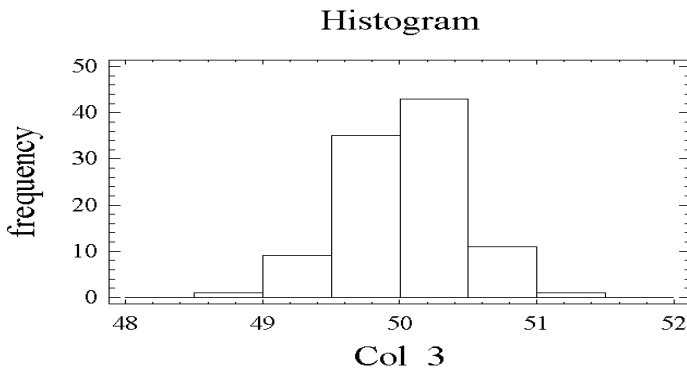


Рисунок 1.15 – Гістограма

1.5.2 Первинна обробка даних з використанням мови R та середовища R-Studio

1.5.2.1 Основні функції та команди, необхідні для виконання роботи

В роботі необхідно використати стандартні функції та команди R, які наведені у таблиці 1.2.

Таблиця 1.2 – Перелік стандартних функцій та команд R

Назва функції	Опис
<i>mean()</i>	Функція для знаходження середнього арифметичного, вона приймає параметр ряду розподілу.
<i>sd()</i>	Функція для пошуку стандартного відхилення, вона приймає параметр ряду розподілу.
<i>rnorm()</i>	Функція слугує для випадкової генерації нормально розподілених чисел. Вона має параметри: <i>N</i> – заданий розмір, <i>mean</i> – середнє значення, <i>sd</i> – стандартне відхилення.
<i>hist()</i>	Функція для створення гістограми. Вона може приймати ті ж параметри, що і у функції для створення графіків <i>plot()</i> . Одна з них: <i>col</i> – слугує для задання кольору стовпчиків або інших елементів графіків. Слугує оцінкою щільності відповідного розподілу.
<i>table()</i>	Функція визначає таблицю частот відповідних рівнів.
<i>median()</i>	Функція що повертає медіану, та приймає ті ж параметри, що й <i>mean()</i> .
<i>which.max()</i>	Функція знаходить порядкові номери елементів з максимальним значенням, а якщо елементів декілька, то буде повернуто номер першого такого елемента.
<i>unique()</i>	Функція повертає вектор, таблицю, або масив, але з однаковими елементами.
<i>tabulate()</i>	Функція бере ціле значення з вектору та підраховує кількість разів яке кожне ціле знаходиться в ньому.
<i>match()</i>	Функція повертає вектор тієї довжини, що і вектор з

	елементами у місці пошуку.
<i>quantile()</i>	Функція розраховує квантілі, приймає параметр ряду розподілу.
<i>print()</i>	Функція виводить на екран об'єкт, це функція загально-го призначення – конкретний результат її роботи буде залежати від класу об'єкта.
<i>cat()</i>	Функція більш розширена, ніж <i>print</i> і теж служить для виведення інформації на консоль.

Перевірка умов в мові *R* виконується наступним чином:

```
if (умова) {
```

```
    виконується, якщо умова вірна
```

```
} else {
```

```
    виконується, якщо умова не вірна
```

```
}
```

Циклічні вирази із заданою великою кількістю ітерацій виконуються за допомогою конструкції:

```
for (<змінна> in <вираз-1>)
```

```
<вираз-2>
```

Тут результатом <вираз-1> повинен бути вектор, а <змінна> на кожній ітерації циклу приймає значення чергового елемента цього вектора. кількість ітерацій дорівнює кількості елементів у векторі.

Відповідно, в <вираз-2> може використовуватися <змінна>, яка буде змінної (лічильником) циклу:

```
sum = 0
```

```
for (i in 1:20)
```

```
    sum = sum + i;
```

```
sum
```

[1] 210

Аналогічно працює цикл *while*.

```
while (<умова>) <вираз>
```

Переривання циклу здійснюється командою *break*. Для переривання поточної ітерації і переходу до наступної служить команда *next*.

1.5.2.2 Приклад реалізації у R-Studio

Згенерувати стовпець даних на основі наступної інформації $N = 900$, $\mu = 9$, $\sigma^2 = 0.9$, де N – кількість дослідів, μ – математичне очікування, σ^2 – дисперсія. Створити нормальний ряд розподілу на основі цих даних. Побудувати гістограму, таблицю частот, отримати описові статистики ряду розподілу.

```

Matoj = 9
Disp = 9/10
N = 9*100
Gen <- rnorm(N, mean = Matoj, sd = Disp) #побудова ряду розпо-
ділу
hist(Gen, col="violet") #побудова гістограми ряду розподілу
freq = table(Gen) #таблиця частот
Gen.mean <- mean(Gen) #середнє арифметичне
Med <- median(Gen) #медіана
Mode <- function(x){
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Gen.mode <- Mode(Gen) #мода
Qua <- quantile(Gen) #розмах (в процентах)

```

Результатом буде побудована гістограма (рис. 1.16) та дані (рис. 1.17). Таблиця частот, середнє арифметичне та інші дані можна переглянути у Global Environmental та вивести на консоль раніше розглянутими у пункті 1.5.2.1. функціями.

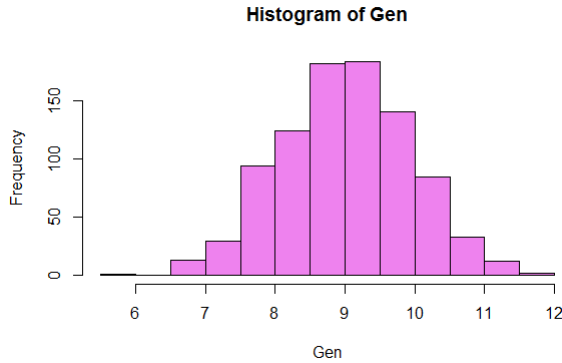


Рисунок 1.16 – Гістограма нормального розподілу

Global Environment	
values	
Disp	0.9
freq	'table' int [1:900(1d)] 1 1 1 1 1 1 1 1 1 ...
Gen	num [1:900] 9.77 7.17 9.68 8.99 10.22 ...
Gen.mean	9.00345852740242
Gen.mode	9.77118723912186
k	900L
Matoj	9
Med	9.02447287134164
Med2	9.06123718789205
N	900
Nya	12.3157844979733
Qua	Named num [1:5] 5.85 8.37 9.02 9.65 11.7
sum	8142.73930732055

Рисунок 1.17 – Результати аналізу

Приведемо приклад коду з використанням функцій, написаних самостійно:

```

Matoj = 9
Disp = 9/10
N = 9*100
cat("Мат. очікування = ")
cat( Matoj, "\n")
cat("Дисперсія = ")

```

```

cat(Disp, "\n")
cat("N = ")
cat(N, "\n")
Gen <- rnorm(N, mean = Matoj, sd = Disp) #побудова рядка роз-
поділу
hist(Gen, col="violet") #побудова гістограми рядка розподілу
freq = table(Gen) #таблиця частот
Gen <- sort(Gen, decreasing = FALSE) #сортування за спаданням
cat("Вибірка:", "\n")
print(Gen)
cat("таблиця частот:", "\n")
print(freq)

#розрахунок середнього арифметичного
sum = 0
for(k in 1:N){
  sum <- sum+Gen[k]
}
Gen.mean <- sum/N
cat("середнє арифметичне:", "\n")
print(Gen.mean)

#медіана
# N – парне, отже
Med2 <- (Gen[N/2] + Gen[N/2-1])/2
cat("медіана:", "\n")
print(Med2)
Mode <- function(x){
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Gen.mode <- Mode(Gen) #мода
cat("мода:", "\n")
print(Gen.mode)

#квантиль
inqua <- function(L, a){
  if (a == 1){

```

```

K <- (a * N)
  if ((K+1)<(N*a)) {
    m <- L[K+1]
    print(m)
  }
  else if((K+1)==(N*a))
  {
    m <- (L[K] + L[K+1])/2
    print(m)
  }
  else if((K+1)>(N*a))
  {
    m <- L[K]
    print(m)
  }
}
else{
  K <- (a * N) + 1
  if ((K+1)<(N*a)) {
    m <- L[K+1]
    print(m)
  }
  else if((K+1)==(N*a))
  {
    m <- (L[K] + L[K+1])/2
    print(m)
  }
  else if((K+1)>(N*a))
  {
    m <- L[K]
    print(m)
  }
}
}
cat("квантиль: ", "\n")
Quak <- function(L){
  cat("0%", "\n")
  R <- inqua(L,0)

```

```

cat("25%", "\n")
R <- inqua(L,0.25)
cat("50%", "\n")
R <- inqua(L,0.5)
cat("75%", "\n")
R <- inqua(L, 0.75)
cat("100%", "\n")
R <- inqua(L,1)
}
Nya <- Quak(Gen)

```

У даному коді функції середнього арифметичного, медіани та іншого написані вручну за допомогою базових команд мови R.

Функція *Quak(L)* використовується для виведення квантиля. У середині функції для зручності розрахунків була використана функція *inqu(a, L, a)*, яка власне і розраховує квантиль залежно від параметрів, що вводяться.

В результаті інформація про вибірку та результати будуть знаходитись на консолі (рис. 1.18 – 1.19).

```

Console ~/rti/ ↵
> cat("Вибірка:", "\n")
Вибірка:
> print(Gen)
[1] 5.772343 6.069365 6.367105 6.493743 6.557616 6.632246 6.690484 6.745241
[9] 6.812176 6.818157 6.851777 6.875082 6.885645 6.977262 6.997254 7.027105
[17] 7.055197 7.058575 7.084307 7.092464 7.111665 7.113568 7.158686 7.176832
[25] 7.182069 7.211556 7.289364 7.298448 7.396644 7.402025 7.404692 7.405302
[33] 7.410545 7.427586 7.462828 7.466267 7.490867 7.505435 7.507177 7.525326
[41] 7.533458 7.534222 7.535516 7.536922 7.537255 7.544346 7.550691 7.555716
[49] 7.570464 7.572515 7.578495 7.584823 7.585516 7.592551 7.605502 7.606006
[57] 7.635003 7.643131 7.645106 7.662661 7.668688 7.701786 7.709237 7.719382
[65] 7.720834 7.722181 7.722770 7.731209 7.741458 7.745065 7.749153 7.768944
[73] 7.769469 7.790771 7.807095 7.807977 7.810175 7.811538 7.811650 7.814208
[81] 7.818395 7.819513 7.821942 7.834108 7.834497 7.842406 7.843966 7.844880
[89] 7.854069 7.857776 7.859436 7.863713 7.867367 7.870093 7.870350 7.873031
[97] 7.884142 7.892579 7.905719 7.913631 7.917069 7.933899 7.936983 7.938589

```

Рисунок 1.18 – Дані про вибірку на консолі

```

Console ~/rfr/
+ k <- inqua(L, 0.75)
+ cat("100%", "\n")
+ R <- inqua(L,1)
+ }
> Nya <- Quak(Gen)
0%
[1] 5.772343
25%
[1] 8.415897
50%
[1] 9.020881
75%
[1] 9.632315
100%
[1] 12.13423
> |

```

Рисунок 1.19 – Дані про вибірку на консолі (продовження)

1.6 Завдання на лабораторну роботу

1.6.1. Дослідити можливості програмного пакету опрацювання даних Statgraphics та мови програмування R.

1.6.2. Ознайомитися з загальними положеннями теорії ймовірностей та статистичної обробки даних.

1.6.3. Згенерувати стовпець на основі наступної інформації: $N = Var * 10$, $\mu = Var$, $\sigma^2 = Var / 10$, де Var – номер варіанта, N – кількість дослідів, μ – математичне сподівання, σ^2 – дисперсія. Для пакету Statgraphics використати функцію $Rnormal(N, \mu, \sigma^2)$.

1.6.4. Зберегти отриману вибірку у форматі *.txt.

1.6.5. Побудувати гістограму, таблицю частот, отримати описові статистики ряду розподілу.

1.6.6. Зберегти результати аналізу у форматі *.rtf.

1.6.7. Побудувати гістограму, таблицю частот, отримати описові статистики ряду розподілу з використанням внутрішніх функцій мови R або реалізувати функції на мові R самостійно.

1.6.8. Зробити висновок.

1.6.9. Оформити звіт.

1.6.10. Відповісти на контрольні питання.

1.7 Зміст звіту

1.7.1 Назва і мета роботи.

1.7.2 Постановка завдання.

1.7.3 Вихідні дані.

1.7.4 Результати аналізу, отримані в пакеті Statgraphics та з використання мови програмування R, зокрема таблиця частот, гістограми і описові статистики.

1.7.5 Висновки по результатам аналізу.

1.8 Контрольні запитання

1.8.1 Робота з даними у системі Statgraphics.

1.8.2 Призначення і використання різних типів вікон.

1.8.3 Загальні і унікальні властивості Statgraphics.

1.8.4 Базова система статистичних процедур пакету Statgraphics.

1.8.5 Основні характеристики розподілу ймовірностей. Записати аналітичні вирази.

1.8.6 Що таке квантіль, мода, медіана?

1.8.7 Як визначити знак коефіцієнта асиметрії з вигляду графіка щільності ймовірності?

1.8.8 Як визначити знак коефіцієнта ексцесу з вигляду графіка щільності ймовірності?

1.8.9 Що таке ряди розподілу?

1.8.10 Які характеристики розподілу ви знаєте?

1.8.11 Як будується гістограма?

1.8.12 Що таке таблиця частот?

1.8.13 На що впливає ширина інтервалу?

2 ЛАБОРАТОРНА РОБОТА № 2

СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

2.1 Мета роботи

Вивчити методику статистичної перевірки гіпотез. Отримати основні характеристики розподілу ймовірностей випадкової величини та перевірити гіпотезу про закон розподілу вибірки з використанням пакету Statgraphics та мови програмування R.

2.2 Короткі теоретичні відомості

2.2.1 Нормальний розподіл

Нормальний розподіл відноситься до числа найбільш розповсюджених та важливих, він часто використовується для наближеного опису багатьох випадкових явищ. Повнота теоретичних дослідів, а також порівняно прості математичні властивості роблять нормальний закон розподілу найбільш привабливим та зручним у використанні, оскільки: по-перше, можна використовувати нормальний закон у якості першого наближення; по-друге, легко підібрати таке перетворення досліджуваної величини, яке видозмінить вихідний "ненормальний" закон розподілу, та перетворить його у нормальний.

Нормальний закон розподілу характеризується щільністю ймовірності виду:

$$\varphi(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma_x^2} \right], -\infty < x < \infty, \quad (2.1)$$

де μ , σ^2 – відповідно математичне сподівання та дисперсія випадкової величини x .

Зміст цієї форми запису означає наступне: якщо випадкова величина x на числовій вісі може прийняти значення, що знаходиться між x_1 та x_2 , то інтеграл від функції $\varphi(x)$ у границях від x_1 до x_2 є ймовірністю того, що випадкова величина розташована у інтервалі між x_1 та x_2 .

$$\int_{x_1}^{x_2} \varphi(x) dx = P(x_1 < x < x_2). \quad (2.2)$$

Теорема А.М. Ляпунова доводить, що сума достатньо великої кількості незалежних або слабозалежних випадкових величин підкоряється нормальному закону. Тут ми маємо приклад з похибками вимірювань – якщо похибка вимірювань створюється багатьма причинами (факторами), комбінації яких випадкові, то вона розподілена за нормальним законом.

Нормальний закон розподілу може бути заданий власними числовими характеристиками (параметрами) μ та σ^2 , що є відповідно математичним сподіванням та дисперсією випадкової величини X . Зміст параметрів нормального закону розподілу наведено на рис. 2.1.

Відзначимо, що $\varphi(x)$ прямує до нуля при $x \rightarrow -\infty$ та $x \rightarrow \infty$. Графік функції $\varphi(x)$ симетричний відносно точки μ . При цьому у цій точці функція $\varphi(x)$ досягає свого максимуму.

Параметр μ характеризує положення графіку функції на числовій вісі. Параметр σ ($\sigma > 0$) характеризує ступінь стиснення або розтягання графіку щільності.

У випадку нормального розподілу, гістограми виходять симетричними (рис. 2.1). За точкою максимуму можна визначити модальне значення (значення, що найчастіше зустрічається). У випадку симетричної гістограми модальне значення співпадає з медіаною та середнім арифметичним.

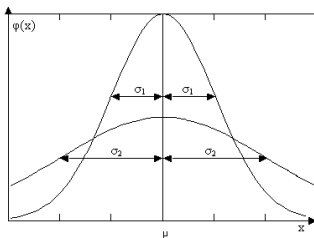


Рисунок 2.1– Щільність нормального розподілу з однаковим середнім μ та різними значеннями дисперсії σ_1, σ_2

При нормальному законі розподілу між інтервалом у якому знаходиться випадкова величина (довірчий інтервал) та ймовірністю, що відповідає цьому інтервалу (довірча ймовірність), існує певне співвідношення (рис.2.2), а саме:

$$P(\mu - \sigma_x \leq x \leq \mu + \sigma_x) = 0,67,$$

$$P(\mu - 2\sigma_x \leq x \leq \mu + 2\sigma_x) = 0,95,$$

$$P(\mu - 3\sigma_x \leq x \leq \mu + 3\sigma_x) = 0,99.$$

Закон трьох сигм полягає в тому, що, якщо випадкова величина розподілена нормально, то абсолютна величина її відхилення від математичного очікування не перебільшує потроєного середнього квадратичного відхилення.

На практиці правило трьох сигм застосовують так: якщо розподіл випадкової величини є невідомим, але умова, що вказана в наведеному прикладі виконується, то існує підстава вважати, що випадкова величина розподілена нормально; в іншому випадку вона не розподілена нормально.

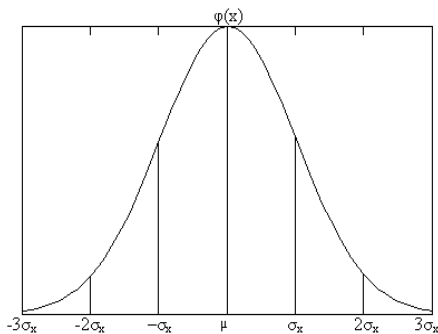


Рисунок 2.2 – Графічна інтерпретація правила трьох сигм

2.2.2 Перевірка гіпотези про розподіл випадкової величини

Для перевірки гіпотези про розподіл випадкової величини використовують критерій χ^2 – Пірсона, критерій Колмагорова та ін. Далі розглянемо використання критерію χ^2 .

Критерій Пірсона не доводить справедливості гіпотези, а лише встановлює при прийнятому рівні значущості q її згоду або незгоду з даними спостережень.

Припустимо, що генеральна сукупність розподілена відповідно до нормального розподілу, ми маємо емпіричні частоти, що отримані відповідно до лабораторної роботи №1, та розрахуємо теоретичні частоти m_i .

Для розрахунку інтегралу (2.2) скористаємося функцією Лапласа. Нормуємо границі інтервалів за формулами:

$$z_i = \frac{x_i - \bar{x}}{S\{x\}}, \quad z_{i+1} = \frac{x_{i+1} - \bar{x}}{S\{x\}}. \quad (2.3)$$

де $z_i = z_{i+1}$;

$S\{x\}$ – середнє квадратичне відхилення змінної x .

При нормуванні границь інтервалів найменше значення z_i визначаємо рівним $-\infty$ ($z_1 = -\infty$), а найбільше: $z_k = +\infty$.

Теоретичні ймовірності P_i того, що x потрапляє в інтервал (x_i, x_{i+1}) розраховуємо за допомогою функції Лапласа $\phi(z)$ відповідно до рівняння:

$$P_i = \phi(z_{i+1}) - \phi(z_i). \quad (2.4)$$

Теоретичні частоти нормального розподілу розрахуємо з виразу (2.5):

$$m_i = NP_i \quad (2.5)$$

При рівні значущості q необхідно перевірити нульову гіпотезу H_0 : генеральна сукупність розподілена відповідно до нормального закону розподілу.

В якості критерію перевірки нульової гіпотези приймається випадкова величина

$$\chi^2 = \frac{(m'_1 - m_1)^2}{m_1} + \frac{(m'_2 - m_2)^2}{m_2} + \dots + \frac{(m'_k - m_k)^2}{m_k}$$

або

$$\chi^2 = \sum_{i=1}^k \frac{(m'_i - m_i)^2}{m_i}, \quad (2.6)$$

де k – кількість інтервалів ряду розподілу.

Ця величина випадкова, оскільки в різних дослідах вона приймає різні, заздалегідь невідомі значення. Чим менше відрізняються емпіричні та теоретичні частоти, тим менше значення критерію (2.6), і, отже, він у відомій мірі характеризує близькість емпіричного і теоретичного розподілів. Зведення до квадрату різниць частот усувається можливістю взаємного погашення додатних і від'ємних різниць.

При необмеженому зростанні об'єму вибірки ($N \rightarrow \infty$) закон розподілу випадкової величини (2.6), незалежно від того, якому закону розподілу підпорядкована генеральна сукупність, прямує до закону розподілу χ^2 з f ступенями свободи. Тому випадкова величина (2.6) позначена як χ^2 , а сам критерій називають критерієм згоди “хі-квадрат”.

Число ступенів свободи знаходять по рівності $f = k - 1 - l$, де l – кількість параметрів припущеного розподілу, що оцінені за даними вибірки, і викликана тим, що є додаткове обмеження:

$$\sum_{i=1}^k m_i = \sum_{i=1}^k m'_i = N, \quad (2.7)$$

тобто теоретична кількість елементів сукупності повинна бути рівною фактичній кількості елементів.

Оскільки в даному випадку припущений розподіл являється нормальним, то оцінюють два параметри (математичне очікування і середньоквадратичне відхилення), тому $l = 2$, і число ступенів свободи $f = k - 3$, якщо розрахункове значення критерію (2.6) виявилось ме-

ніше критичного $\chi_{кр}^2$, що знаходять за таблицями для відповідного рівня значимості q і числа ступенів свободи f , тобто якщо

$$\chi^2 < \chi_{кр}^2 = \chi_{q, f}^2, \quad (2.8)$$

то не існує підстав відхилити нульову гіпотезу про нормальний розподіл. В протилежному випадку (при $\chi^2 > \chi_{кр}^2$) нульова гіпотеза відхиляється.

При перевірці гіпотези про нормальність розподілу існує правило, згідно з яким загальна кількість елементів вибірки повинна бути $N \geq 50$, а кількість елементів, що потрапили у будь-який i -й інтервал (тобто значення емпіричних частот m'_i), повинна бути:

$$m'_i \geq 5, (i = \overline{1, k}) \quad (2.9)$$

Якщо в крайні інтервали потрапляє менша кількість елементів, то вони об'єднуються з сусідніми інтервалами. Внутрішні інтервали об'єднувати забороняється. Загальна кількість інтервалів k_1 , що залишилися після об'єднання, повинна задовольняти умові: $k_1 \geq 4$. Інакше число ступенів свободи f (2.8) виявиться рівним нулю, і гіпотезу неможливо буде перевірити.

2.2.3 Приклади перевірки гіпотез та побудови довірчих інтервалів

Приклад 2.1. Побудувати довірчий 95% інтервал для математичного сподівання вибірки з попереднього завдання за допомогою критерію Ст'юдента.

Розв'язання

З формули для обчислення середньоквадратичного відхилення

$$S\{x\} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} \quad \text{отримаємо: } S\{x\} = 0,0311.$$

Знайдемо значення коефіцієнта Ст'юдента. Розрахуємо $q = 1 - 0,95 = 0,05$; $f = N - 1 = 200 - 1 = 199$. З таблиці отримаємо коефіцієнт Ст'юдента: $t_{kp}(0.05, 199) = 1,6525$.

Отримаємо значення довірчого інтервалу за формулою $\delta = \frac{1}{\sqrt{N}} t_{kp} S\{x\}$: $\delta = \frac{1}{\sqrt{200}} * 1,6525 * 0,0311 = 0,0003$.

Побудуємо довірчий інтервал для математичного сподівання:

$$6,7530 \leq M\{x\} \leq 6,7536$$

Приклад 2.2. Задані дві вибірки:

Вибірка 1: 0,958; 0,909; 0,859; 0,863; 0,811; 0,877; 0,798; 0,855; 0,788; 0,821; 0,830; 0,718; 0,642; 0,658; 0,753; 0,692;

Вибірка 2: 36,2; 39,8; 14,3; 46,8; 46,8; 49,7; 58,1; 62,3; 70,6; 71,1; 71,3; 83,2; 83,6; 90,1; 99,5; 111,2.

Перевірити рівність дисперсій вибірок за допомогою критерію Фішера. Вручну визначити розрахункове значення критерію за вибіркою та табличне значення критерію для рівнів значущості $q=0.05$. Для кожного випадку визначити критичну область та область прийняття гіпотези. Зробити відповідні висновки.

Розв'язання

Використовуючи формулу для обчислення дисперсії, отримаємо, що дисперсія першої вибірки дорівнює:

$$S_1^2\{x\} = \frac{1}{\sqrt{N-1}} \sum_{i=1}^N (x_i - \bar{x})^2 = 0,008035.$$

Для другої вибірки дисперсія дорівнює:

$$S_2^2\{x\} = \frac{1}{\sqrt{N-1}} \sum_{i=1}^N (x_i - \bar{x})^2 = 648,3185.$$

Знайдемо табличне значення критерію Фішера $F_{kp}(q, f_1, f_2)$, де $f_1 = N_1 - 1$, $f_2 = N_2 - 1$, N_1 – кількість екземплярів першої вибірки, N_2 – кількість екземплярів другої вибірки.

Для визначення критичної області необхідно знайти критичну точку $K_{кр} = F_{кр}(0,05;15;15) = 2,403$ та критерій Фішера за формулою

$$F = \frac{S_1^2\{x\}}{S_2^2\{x\}}, \text{ де у чисельник підставляється завжди більша дисперсія:}$$

$$K = F = \frac{S_1^2\{x\}}{S_2^2\{x\}} = \frac{648,3185}{0,008035} = 8068,8.$$

Оскільки отримали значення $F > F_{кр} (8068,8 > 2,403)$, то ми відкидаємо нульову гіпотезу про рівність дисперсій двох вибірок з рівнем значущості 0,05.

Приклад 2.3. Перевірити однорідність дисперсій 4 вибірок по 16 вимірів за допомогою критерію Кохрена. Традиційним способом (вручну) визначити розрахункове значення критерію за вибіркою та табличне значення критерію для рівнів значущості $q=0.05$. Для кожного випадку визначити критичну область та область прийняття гіпотези. Зробити відповідні висновки.

Розв'язання

Задані чотири вибірки:

Вибірка 1: 0,958; 0,909; 0,859; 0,863; 0,811; 0,877; 0,798; 0,855; 0,788; 0,821; 0,830; 0,718; 0,642; 0,658; 0,753; 0,692;

Вибірка 2: 36,2; 39,8; 14,3; 46,8; 46,8; 49,7; 58,1; 62,3; 70,6; 71,1; 71,3; 83,2; 83,6; 90,1; 99,5; 111,2.

Вибірка 3: 0,671; 0,537; 0,549; 0,685; 0,543; 0,597; 0,701; 0,684; 0,647; 0,727; 0,777; 0,661; 0,659; 0,701; 0,699; 0,678.

Вибірка 4: 10,1; 10,9; 11,3; 12,4; 20,1; 18,2; 14,4; 19,3; 16,7; 17,5; 13,4; 25,4; 17,1; 16,8; 15,1; 19,9.

Для кожної вибірки знайдемо дисперсію за формулою:

$$S1^2\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = 0,008035,$$

$$S2^2\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = 648,3185,$$

$$S3^2\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = 0,004658,$$

$$S4^2\{x\} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 = 16,4185.$$

Знайдемо значення критерію Кохрена:

$$G = \frac{S_j^2\{x\}_{\max}}{\sum_{i=1}^N S^2\{x\}} = \frac{648,3185}{(0,008035 + 648,3185 + 0,004658 + 16,4185)} = \frac{648,3185}{664,7497} = 0,9753.$$

Знайдемо табличне значення критерію Кохрена $G_{кр}(q, f_1, f_2)$ для рівня значущості $q=0,05$, де $f_1 = N_1 - 1$, $f_2 = f_{знач}$ – кількість дисперсій, де N_1 – кількість екземплярів вибірок:

$$G_{кр}(0,05;15;4) = 0,2419.$$

Для визначення критичної області необхідно знайти критичну точку $K_{кр} = G_{кр}(0,05;15;4) = 0,2419$ для рівня значущості $q=0,05$. Так як ми отримали значення $G > G_{кр}(0,9753 > 0,2419)$, то ми відкидаємо нульову гіпотезу про однорідність дисперсій заданих чотирьох вибірок з рівнем значущості 0,05.

2.3 Порядок виконання роботи

2.3.1 Виконання у пакеті Statgraphics

Перевіримо вибірку на розподіл відповідно до нормального розподілу у пакеті Statgraphics.

Крок 1. Натисніть двічі на стовпець Col_1 системи Statgraphics. З'явиться вікно, зображене на рис. 2.5. Вибрати пункт Formula. У рядку вводу записати: $Rnormal(N; \mu; \sigma^2)$, де N – кількість дослідів, μ – середнє квадратичне, σ^2 – дисперсія, відповідно до завдання.

Modify Column

Name: Col_1

Comment:

Width: 13

Type

- ☐ Numeric
- ☐ Character
- ☐ Integer
- ☐ Date
- ☐ Month
- ☐ Quarter
- ☐ Time (HH:MM)
- ☐ Time (HH:MM:SS)
- ☐ Fixed Decimal: 2
- ☒ Formula

Buttons: OK, Cancel, Define..., Help

Рисунок 2.5 – Вікно з параметрами стовпця даних

Після натискання клавіші ОК, отримаємо вікно з початковими даними, що зображено на рис. 2.6:

	Col_1	Col_2
1	10,967133729	
2	8,88972420734	
3	9,82030369458	
4	10,0290438753	
5	9,14838458446	
6	10,7419088632	
7	10,7291148081	
8	10,292370645	
9	9,73503691756	
10	9,27326990174	
11	9,82107309206	
12	11,7073624153	
13	11,6756186384	
14	8,51756651716	
15	11,252218179	
16	11,9568060326	
17	11,5189756573	
18	9,162425866	
19	10,26543296	

Рисунок 2.6 – Згенеровані значення для стовпця Col_1

Крок 2. Вибрати пункт меню Describe → Distributions → Distribution Fitting (Uncensored Data) (Перевірка розподілу). Перенести значення Col_1 зі стовпця зліва у строку вводу Data (рис. 2.7). Натиснути кнопку ОК.

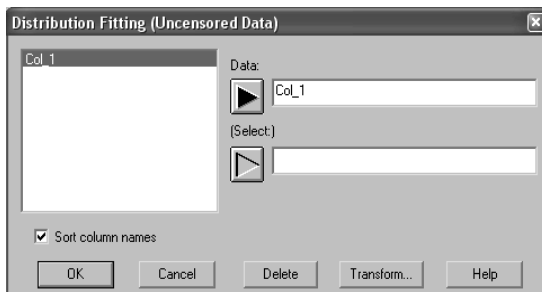


Рисунок 2.7 – Вікно Distribution Fitting

Крок 3. Отримаємо гістограму (рис. 2.8) і дані, що були отримані в результаті процесу перевірки розподілу на відповідність нормальності (рис. 2.9).

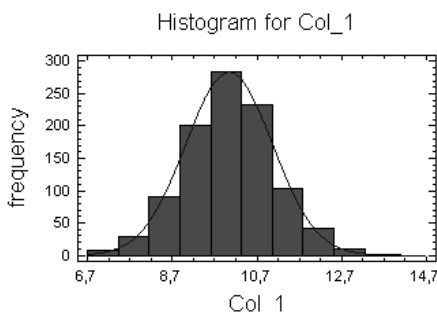


Рисунок 2.8 – Гістограма і перевірка її на нормальність

На рис. 2.9 поля **Lower Limit** and **Upper Limit** означають нижню і верхню границі інтервалів. На кожному з інтервалів розраховується частота, що спостерігається (експериментальна) – **Observed Frequency** та частота, що очікується (теоретична) – **Expected Frequency**. Останнім полем таблиці є значення критерію хі-квадрат

(**Chi - Square**). Як видно з таблиці, що зображена на рис. 2.9, **STATGRAPHICS** аналізує отриманий результат автоматично. Судячи з того, що значення **P-value** є більшим, ніж 0,1 ($P\text{-value}=0.36247$), ми не можемо відхилити нульову гіпотезу при рівні значущості 0,05. Іншими словами, ми приймаємо гіпотезу, що закон розподілу вибірки відповідає нормальному.

Goodness-of-Fit Tests for Col_1

Chi-Square Test					
	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	8,6809	8,6809	89	90,91	0,04
	8,6809	9,11816	90	90,91	0,01
	9,11816	9,42953	87	90,91	0,17
	9,42953	9,69168	87	90,91	0,17
	9,69168	9,93204	95	90,91	0,18
	9,93204	10,1661	95	90,92	0,18
	10,1661	10,4064	104	90,91	1,89
	10,4064	10,6686	101	90,91	1,12
	10,6686	10,9799	80	90,91	1,31
	10,9799	11,4172	74	90,91	3,14
above	11,4172		98	90,91	0,55

Chi-Square = 8,76541 with 8 d.f. P-Value = 0,36247

Estimated Kolmogorov statistic DPLUS = 0,0250974
 Estimated Kolmogorov statistic DMINUS = 0,0135843
 Estimated overall statistic DN = 0,0250974

Рисунок 2.9 – χ^2 тест для Col_1

2.3.2 Статистична перевірка гіпотез засобами мови R

В роботі необхідно використати наступні стандартні функції та команди R. Функція `array()` – призначена для створення багатовимірних масивів.

Приклад:

```
aux=array(rep(0,60),dim=c(3,4,5)) # створюємо набір з 5 матриць,
що складаються з 3 рядків, 4 стовпців, заповнених нулями
```

Функція *list()* – призначена для створення списків. В список можна включати поєднання будь-яких типів даних.

Приклад створення та заповнення списку:

```
list1 <- c("A", "B", "C")
```

```
list2 <- c(FALSE, TRUE)
```

```
my.list <- list(Text= list1, Logic= list2)
```

Звертання до елементів списку:

```
> my.list$Text[2]
```

```
[1] "B"
```

Функція *density()* – оцінює щільність ймовірності.

Приклад використання функції:

```
plot(density(my.array),main="Density estimate of data")
```

Функція *shapiro.test()* – призначена для виконання тесту Шапіро-Уїлка. Функції передається вибірка у вигляді масива.

Функція *chisq.test()* – призначена для перевірки критерію узгодженості Пірсона. Функції передається вибірка у вигляді масива.

Функція *t.test()* – призначена для виконання тесту Ст'юденту. Функції передається дві вибірки у вигляді масива.

Функція *install.packages()* – завантажує пакет зі стороннього джерела.

Функція *data()* – завантажує таблицю з даними.

Функція *attach()* – додає таблицю з даними до списку наявних змінних.

Функція *tapply()* – застосовує функцію до кожної сукупності значень, створеної відповідно до рівнів певного фактора.

2.3.2.1 Перевірка гіпотез у R-Studio для однієї вибірки

Виконаємо перевірку гіпотез у R-Studio для однієї вибірки. Для цього заповнюємо одновимірний масив 25 випадковими числами:

```
my.array <- array( rnorm(25, mean = 25, sd = 1),dim = c(25,1),  
dimnames = list(c(1:25) , c("Col1") ))
```

```
# виводимо вміст масиву
```

```
my.array
```

```
hist(my.array,main="Histogram of observed data")
# будемо гістограму за щільністю
plot(density(my.array),main="Density estimate of data")
```

```
#для виконання тесту Шапіро-Уїлка виконаємо функцію
shapiro.test(), де як аргумент передаємо масив
shapiro.test(my.array)
# у масив my1 заносимо ймовірності розподілення масиву my
my1.array<- rnorm(my.array, mean = 25, sd = 1)
# для пошуку критерію узгодженості Пірсона використовуємо
функцію chisq.test
chisq.test(my1.array)
# для виконання тесту Колмогорова-Смирнова використовуємо
функцію ks.test
ks.test(my.array, my1.array)
```

Результатом буде побудована гістограма (рис. 2.10). Результати тесту Колмогорова – Смирнова та тесту Шапіро-Уїлка будуть виведені на консоль (рис. 2.11).

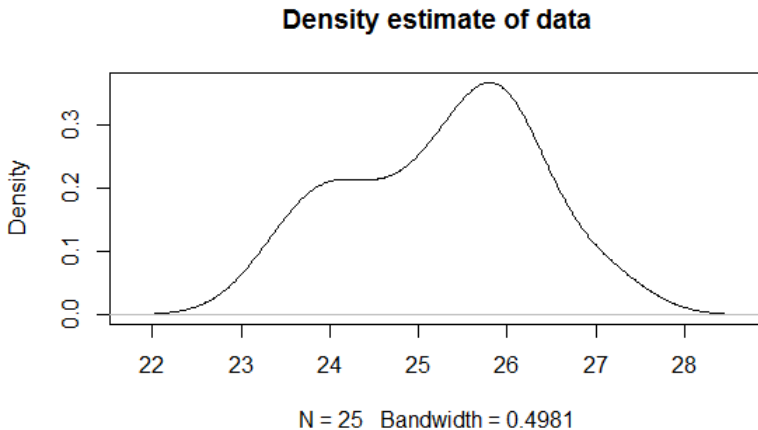


Рисунок 2.10 – Гістограма за щільністю

```

shapiro-wilk normality test

data:  my.array
W = 0.95496, p-value = 0.3232

chi-squared test for given probabilities

data:  my1.array
X-squared = 4.164, df = 24, p-value = 1

```

Рисунок 2.11 – Результати перевірки гіпотез

2.3.2.2 Перевірка гіпотез у R-Studio для двох вибірок

Виконаємо перевірку гіпотез у R-Studio двох вибірок. Результати наведено на рисунку 2.12.

```

# Набрати в консолі для роботи прикладу
# > install.packages("ISwR")

```

```

library(ISwR)      #Завантаження ISwR
data(energy)       #Завантажує таблицю energy
attach(energy)     #Додає таблицю energy
energy             #Виводить energy
tapply(expend, stature, mean)  # застосовує функцію mean до
expend, за допомогою у stature

```

```

# побудова діаграми розмаху
boxplot(expend ~ stature,
        xlab = "Тип будови тіла",
        ylab = "Витрати енергії",
        main = "T-test",
        col = "coral", data = energy)
#t-тест Стьюдента
t.test(expend ~ stature)

```



```

> boxplot(expend ~ stature,
+         xlab = "Тип будови тіла",
+         ylab = "Витрати енергії",
+         main = "T-test",
+         col = "coral", data = energy)
> t.test(expend ~ stature)

welch Two Sample t-test

data:  expend by stature
t = -3.8555, df = 15.919, p-value = 0.001411
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.459167 -1.004081
sample estimates:
mean in group lean mean in group obese
      8.066154           10.297778

```

Рисунок 2.12 – Результати t-Testy

2.4 Завдання на лабораторну роботу

2.4.1 Використовуючи рекомендовану літературу та ці методичні вказівки вивчити основні поняття та застосування методики перевірки гіпотез для оцінювання параметрів випадкових величин, роботу статистичного пакету програм Statgraphics та мови програмування R, для перевірки статистичних гіпотез розподілу випадкових величин.

2.4.2 Вивчити загальні положення теорії статистичної перевірки гіпотез.

2.4.3 Згенерувати стовпець даних на основі наступної інформації: $N = Var * 100$, $\mu = Var$, $\sigma^2 = Var/10$, де Var – номер варіанта, N – кількість дослідів, μ – математичне сподівання, σ^2 – дисперсія.

2.4.4 Зберегти отриману вибірку у форматі .xls (Excel).

2.4.5 Превірити гіпотезу про нормальний розподіл вибірки, використовуючи критерій Пірсона – χ^2 і критерій Колмогорова з використанням внутрішніх функцій мови R.

2.4.6 Зробити висновок.

2.4.7 Оформити звіт.

2.4.8 Відповісти на контрольні запитання.

2.5 Зміст звіту

2.5.1 Мета роботи.

2.5.2 Теоретичний аналіз щодо критеріїв перевірки статистичних гіпотез.

2.5.3 Індивідуальне завдання.

2.5.4 Отримані результати обчислень та графіки.

2.5.5 Аналіз отриманих результатів та висновки.

2.6 Контрольні запитання

2.6.1 Які ви знаєте параметри нормального закону розподілу випадкових величин?

2.6.2 Наведіть аналітичний вираз щільності ймовірності для нормального закону розподілу.

2.6.3 Накресліть функцію щільності ймовірності для нормального закону.

2.6.4 Чому дорівнюють коефіцієнт асиметрії та коефіцієнт ексцесу для нормального закону?

2.6.5 Яким критерієм визначається закон розподілу випадкової величини? Записати аналітичний вираз.

2.6.6 Як змінюється вигляд гістограми при зміні величини інтервалу?

2.6.7 Як використовується правило трьох сігм для визначення закону розподілу випадкової величини?

2.6.8 Необхідність застосування статистичних методів обробки результатів спостережень.

2.6.9 У чому полягає основний принцип статистичних гіпотез?

2.6.10 Які гіпотези можна перевірити за допомогою критерія Пірсона, Колмогорова, Фішера, Стюдента, Кохрена?

2.6.11 Як визначити довірчі інтервали для математичного сподівання?

3 ЛАБОРАТОРНА РОБОТА № 3 ДИСПЕРСІЙНИЙ АНАЛІЗ

3.1 Мета роботи

Вивчити методи дисперсійного аналізу, провести дослідження ефекту дії одного та двох якісних факторів на одну змінну відгуку відповідно, використовуючи для цього пакети статистичних програм.

3.2 Короткі теоретичні відомості

3.2.1 Постановка завдання

При дослідженні залежностей найбільш простим виявляється випадок впливу одного фактора на кінцевий результат, коли цей фактор приймає лише кінцеву кількість значень (рівнів). Такі задачі часто зустрічаються на практиці і називаються *задачами однофакторного дисперсійного аналізу*.

Типовий приклад – порівняння за результатами декількох різних способів дії, спрямованих на досягнення однієї мети:

- отримання експериментальних даних при дослідженні одного й того ж процесу декількома приладами;
- отримання визначених знань студентами при використанні декількох підручників.

Для опису задач однофакторного дисперсійного аналізу використовують наступну термінологію:

- *фактор* або *фактори* (якщо їх декілька) – те, що впливає на кінцевий результат (в наведених нижче прикладах факторами є поняття “прилад”, “підручник”). Розрізняють кількісні фактори, що можливо вимірювати на деякій шкалі (час) та якісні фактори, що вимірюються категоріями (колір, тип приладу);
- *рівень фактора* або *спосіб обробки* – конкретна реалізація фактора (визначений прилад або обраний підручник);
- *відгук* – значення ознаки, що вимірюється (тобто величина результату).

Для оцінювання впливу факторів на відгук необхідні експериментальні статистичні дані, що отримують наступним способом: кожний з

K рівнів фактора застосовують кілька разів до об'єкту, який досліджується, і реєструють результати. Підсумком випробувань є K вибірок.

Найбільш зручним способом подання даних є таблиця 3.1. У залежності від кількості факторів, що впливають, (в таблиці 3.1 – один фактор), говорять, що дані зведені в таблицю, з одним, двома і т.д. факторами.

Таблиця 3.1 – Подання статистичних даних

Рівні фактора	1	2	...	K
Результати вимі- рів	X_{11}	X_{12}	...	X_{1K}
	X_{21}	X_{22}	...	X_{2K}

	$X_{n_1 1}$	$X_{n_2 2}$...	$X_{n_k K}$

n_1, \dots, n_k – об'єми вибірок;

$N = n_1 + n_2 + \dots + n_k$ – загальна кількість спостережень.

Однією з головних кінцевих цілей у задачах однофакторного аналізу є оцінка величини впливу конкретного рівня фактора на відгук, що досліджується. Ця задача може бути сформульована у формі порівняння впливу двох чи декількох рівнів фактора між собою, тобто оцінки відмінності (контрасти) між дією різних рівнів фактора. Але перш ніж судити про кількісний вплив фактора на вимірювану ознаку, необхідно визначити, чи є такий вплив узагалі. Чи не можна пояснити розбіжності спостережених у досліді значень для різних рівнів фактора дією чистої випадковості? Адже внутрішньо властива явищу змінюваність приводить до того, що результати виявляються різними навіть при незмінному значенні фактора (тобто в кожному стовпці табл. 3.1). Можливо, тією ж причиною можна пояснити і відмінність між її стовпцями? Статистичною мовою це припущення означає, що всі дані табл. 3.1 належать одному й тому ж самому розподілу. Це припущення, як правило, називають *нульовою гіпотезою* і позначають H_0 . Для перевірки нульової гіпотези можуть бути використані різні критерії: як традиційні, що спираються на припущення про нормальність розподілу даних (F -відношення), так і непараметричні, що не потребують подібних припущень (рангові критерії Краскела-Уолліса, Джонкхієра та ін.).

Якщо нульова гіпотеза про відсутність ефектів обробки відкидається, то проводиться оцінка дії цих ефектів чи контрастів між ними і будуються довірчі інтервали для цих характеристик.

Якщо ж критерії не дозволяють відкинути нульову гіпотезу про відсутність ефектів обробки, то звичайно на цьому аналіз може бути закінчений. Але іноді висновок про відсутність ефектів обробки нас не може влаштувати, оскільки він суперечить теоретичним передумовам чи результатам попередніх досліджень. Тоді необхідно з'ясувати, чи немає ще яких-небудь факторів, що впливають на наявні спостереження. Можливо, вплив ефекту обробки не вдалося знайти лише тому, що його вплив є непомітним на фоні розходжень, викликаних дією неврахованого нами фактора. Наприклад, при вивченні впливу способів обробки ґрунту на врожайність таким фактором може бути тип ґрунту.

3.2.2 Дисперсійний аналіз

Для опису даних табл. 3.1 у більшості випадків виявляється прийнятною адитивна модель. Вона припускає, що значення відгуку x_{ij} можна подати у вигляді суми внеску (впливу) фактора і незалежної від внесків факторів випадкової величини. Інакше кажучи, кожне спостереження x_{ij} є сумою вигляду:

$$x_{ij} = a_j + e_{ij}, j = 1, \dots, k, i = 1, \dots, n, \quad (3.1)$$

де a_1, a_2, \dots, a_k – не випадкові невідомі величини, що є результатом дії відповідних обробок;

e_{ij} – незалежні однаково розподілені випадкові величини, що відображають внутрішню властиву спостереженням зміну.

Випадкові величини e_{ij} безпосередньо не спостерігаються, нам відомі лише значення x_{ij} .

При розгляді адитивної моделі однофакторного аналізу передбачається безперервність закону розподілу величин e_{ij} , при тому, що e_{ij} – незалежні й однаково розподілені. Часто про розподіл e_{ij} можна сказати більше, а саме, величини $e_{ij} \sim N(0, \sigma^2)$, тобто мають нормальний розподіл з нульовим середнім і загальною для всіх дисперсією σ^2 , що нам невідома. Додаткова інформація про закон розподілу випадко-

вих величин e_{ij} дозволяє використовувати більш сильні методи в моделі однофакторного аналізу, як для перевірки гіпотез, так і для оцінки параметрів. Сукупність цих методів називається *однофакторним дисперсійним аналізом*.

Ця назва пов'язана з тим, що аналіз моделі (3.1) заснований на зіставленні двох оцінок дисперсії σ^2 . Одна з них діє поза залежністю від того, вірна чи ні гіпотеза $H_0: a_1 = \dots = a_k$. Інша оцінка істотно використовує це припущення. Вона дає близький до σ^2 результат тільки в тому випадку, якщо гіпотеза є вірною. Зіставляючи одна з одною ці дві оцінки, ми можемо встановити, що H_0 варто відкинути, якщо вони виявляються помітно (значимо) різними.

Кожна однорідна група табл. 3.1 (кожен її стовпець) дає оцінку σ^2 . Для цього треба за кожним стовпцем знайти вибіркву суму квадратів відхилень від середнього арифметичного. Нехай середнє арифметичне для кожного стовпця дорівнює:

$$x_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}, j = 1, \dots, k \quad (3.2)$$

і далі обчислимо $\sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$. Аналізуючи одну нормальну вибірку, ми знайшли, що таку суму квадратів можна представити у вигляді добутку $\sigma^2 \chi^2$, де випадкова величина χ^2 має розподіл χ^2 з $(n_j - 1)$ ступенями волі. Оскільки дані в різних стовпцях отримані незалежно, об'єднана сума квадратів $\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$ має розподіл $\sigma^2 \chi^2$ з $(N - k)$ ступенями волі. Звідси отримаємо першу (основну) оцінку σ^2 :

$$\sigma^{2*} = \frac{1}{N - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2 \quad (3.3)$$

У висновку не було згадування про гіпотезу H_0 , отже, $\sigma^{2*} \approx \sigma^2$ незалежно від того, вірна гіпотеза H_0 чи ні.

Щоб отримати іншу оцінку σ^2 , звернемося знову до стовпців табл. 3.1, точніше – до їхніх середніх значень $x_{.j}$. Відповідно до властивостей нормального розподілу:

$$x_{.j} \sim N(a_j, \sigma^2 / n_j) . \quad (3.4)$$

Крім того, $x_{.j}$ і $\sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$ статистично незалежні. Знайдемо центр сукупності (3.4) з урахуванням «ваг» середніх значень n_j , тобто знайдемо, при якому z досягається мінімум виразу

$$\sum_{j=1}^k (x_{.j} - z)^2 n_j \rightarrow \min_z . \quad (3.5)$$

За допомогою стандартних засобів математичного аналізу легко побачити, що мінімум (3.5) досягається при $z = \bar{x}$, де

$$\bar{x} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} . \quad (3.6)$$

Відзначимо, що при виконанні гіпотези H_0 значення виразу (3.5) при $z = \bar{x}$ має розподіл $\sigma^2 \chi^2(k-1)$, де $\chi^2(k-1)$ – розподіл хі-квадрат з $(k-1)$ ступенями волі. Звідси знаходимо другу оцінку для σ^2 :

$$\sigma^{2**} = \frac{1}{k-1} \sum_{j=1}^k n_j (x_{.j} - \bar{x})^2 . \quad (3.7)$$

Оскільки випадкові величини $x_{.j}$ незалежні від (3.3), те ж вірно і для їхніх комбінацій. Тому оцінка (3.7) є незалежною від (3.3).

При порушенні H_0 оцінка σ^{2**} має тенденцію до зростання, тим більшому, чим більше відхилення від H_0 .

Оскільки ми маємо для оцінки σ^2 дві незалежні оцінки, що мають при гіпотезі H_0 розподіл хі-квадрат, їх частка $F = \sigma^{2**} / \sigma^{2*}$, чи, докладніше вираз (3.8):

$$F = \frac{\frac{1}{k-1} \sum_{j=1}^k n_j (x_{\cdot j} - \bar{x})^2}{\frac{1}{N-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{\cdot j})^2}, \quad (3.8)$$

повинен мати F -розподіл з $(k-1, N-k)$ ступенями волі. Відзначимо, що статистика (3.8) вже не залежить від σ^2 . Вираз (3.8) одержує тим більшу тенденцію до зростання, чим сильніше порушується гіпотеза H_0 . Тому проти H_0 говорять великі (неправдоподібно великі) значення F , розраховані за спостереженнями, далі — $F_{\text{набл.}}$. Отже, для перевірки H_0 треба було б обчислити $P(F \geq F_{\text{набл.}})$, тобто ймовірність одержати за рахунок дії випадковості значення статистики F , яке є більшим або рівним $F_{\text{набл.}}$. Гіпотезу H_0 варто відкинути, якщо ймовірність $P(F \geq F_{\text{набл.}})$ є малою.

3.2.3 Зв'язок двохфакторного й однофакторного аналізу

Розглянемо задачу про дію на вимірювану величину (відгук) двох факторів. У цій задачі ми припускаємо, що на відгук можуть впливати два фактори, кожен з яких приймає кінцеве число значень (рівнів), і цікавимося тим, як впливають ці фактори на досліджуваний відгук і чи впливають взагалі. Такі задачі характерні як для промислових і технологічних експериментів, так і для гуманітарних досліджень.

Буває, що в рамках однофакторної моделі вплив фактора, що нас цікавить, не виявляється, хоча змістовні розуміння вказують, що такий вплив повинний бути. Іноді цей вплив виявляється, але точність висновків про кількісний бік цього впливу недостатній. Причиною такого явища може бути великий внутригруповий розкид, на фоні якого дія фактора залишається непомітною чи майже непомітною. Дуже часто цей розкид викликається не тільки випадковими причинами, але також дією ще одного фактора. Якщо ми в змозі вказати такий фактор, тоді можна спробувати включити його в модель, щоб зменшити статистичну неоднорідність спостережень і завдяки цьому виявити дію на відгук закономірних причин. Звичайно, не завжди вдається поправити справу введенням одного фактора, що “заважає”, і переходом до двохфакторних схем, як вище. Іноді приходится розглядати і трьох-, і багатфакторні моделі. Задум у всіх цих випадках залишається таким самим.

До задач двохфакторного чи багатфакторного аналізу часто приводять також дослідження з оптимізації технологічних процесів. При цьому найчастіше заздалегідь відомо, що обидва фактори впливають на відгук, а дослідника цікавить чисельна оцінка цього впливу з метою вибору оптимального рівня факторів.

Іноді фактори розділяють на важливі і ті, що заважають, але це зовсім не обов'язково. У ряді задач фактори змістовно рівноправні для експериментатора. Ці нюанси мало впливають на статистичні моделі, вони можуть позначитися тільки на постановках статистичних питань.

3.2.4 Таблиця вихідних даних для двохфакторного аналізу

Розглянемо, як змінюється таблиця даних однофакторного аналізу, при включенні в модель дії фактора, що заважає.

Назвемо головний фактор фактором A , а фактор, що заважає, – фактором B . Нехай фактор A приймає k , а фактор B – n різних значень. Фактор B розбиває всі об'єкти спостереження на n блоків, кожен блок утворює спостереження, проведені при одному рівні фактора B . У блоці відгуки можуть значимо відрізнитися тільки за рахунок застосування до них різних *обробок*, тобто за рахунок різних рівнів фактора A . Рівні фактора A (обробки) відображаються в таблиці по стовпцях, а рівні фактора B (блоки) – по рядках. Традиційна термінологія “блок-обробка” у застосуванні до факторів B і A склалася як результат різного відношення до цих факторів, один із яких є фактором, що заважає, а інший – визначальним.

Основну таблицю вихідних даних для двохфакторного аналізу наведено у табл. 3.2.

Таблиця 3.2 – Основна таблиця вихідних даних для двохфакторного аналізу

Блоки	Обробки			
	1	2	...	k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}		...	x_{2k}
...
N	x_{n1}	x_{n2}	...	x_{nk}

Таблиця 3.2, що містить $n \times k$ спостережень (по одному спостереженню в клітинці), є основною таблицею двохфакторного аналізу. Її відмінність від таблиці однофакторного аналізу полягає в тому, що спостереження в будь-якому стовпці не є однорідними, тобто можуть не утворювати вибірки (якщо вплив фактора, що заважає, є значимим). Для опису такої двохфакторної таблиці вимагаються більш складні ймовірнісні моделі, ніж для однофакторного аналізу.

3.2.5 Адитивна модель даних двохфакторного експерименту при незалежній дії факторів

Для опису даних таблиці 3.2 двохфакторного експерименту в більшості випадків виявляється прийнятною адитивна модель. Вона припускає, що значення відгуку x_{ij} є сумою самостійних внесків відповідних рівнів кожного з факторів і незалежних від цих факторів випадкових величин. Останні відображають внутрішню зміну відгуку при фіксованих рівнях факторів, що може викликатися різними причинами.

Таким чином, кожне спостереження x_{ij} можна подати у вигляді:

$$x_{ij} = b_i + t_j + e_{ij}; \quad i = 1, \dots, n; \quad j = 1, \dots, k. \quad (3.9)$$

При цьому числа b_i, \dots, b_n є результатами впливу на відгук фактора B , що заважає, дія якого розбиває всі дані на блоки. Тому величини b_i, \dots, b_n називають *ефектами блоків*. Числа t_1, \dots, t_k відображають дію на відгук фактора A , який нас цікавить, і називається *ефектами обробки*. Щодо випадкових величин e_{ij} передбачається, що вони однаково розподілені і незалежні в сукупності. Різні методи двохфакторного аналізу потребують від їхнього розподілу або тільки безперервності, або приналежності до нормального сімейства розподілів $N(0, \sigma^2)$ із середнім 0 і деякою невідомою дисперсією σ^2 .

Відзначимо, що навіть у випадку справедливості виразу (3.9) величини внесків факторів b_i і t_j не можуть бути відновлені однозначно. Дійсно, збільшення всіх b_i на одну константу й одночасне зменшення всіх t_j на цю ж константу залишає вираз (3.9) незмінним. Для однозначної визначеності внесків факторів зручно перейти до подання спостережень у вигляді:

$$x_{ij} = \mu + \beta_i + \tau_j + e_{ij}; \quad i = 1, \dots, n; j = 1, \dots, k, \quad (3.10)$$

вважаючи, що $\sum_{i=1}^n \beta_i = 0$, $\sum_{j=1}^k \tau_j = 0$. При цьому параметр μ інтерпретується як середнє значення, властиве усім величинам x_{ij} , а β_i та τ_j – як відхилення від μ у результаті дії факторів B та A .

Як і у випадку однофакторного аналізу, доцільно в першу чергу перевірити гіпотезу про значущість ефектів обробки. Сформулюємо нульову гіпотезу у вигляді: $H_0: \tau_1 = \tau_2 = \dots = \tau_k = 0$. Іншими словами, припустимо, що вплив фактора A відсутній.

3.2.6 Двохфакторний дисперсійний аналіз

Якщо є підстави припускати, що випадкові величини e_{ij} у моделі двофакторного аналізу (3.10) мають нормальний розподіл з нульовим середнім і невідомою однаковою при всіх i та j дисперсією σ^2 , можна запропонувати критерій для перевірки гіпотези $H_0: \tau_1 = \tau_2 = \dots = \tau_k = 0$ та побудувати ефективні оцінки параметрів μ , τ_j і β_i . Методи, які використовуються для цього, є аналогічними тим, що були розглянуті при вирішенні задач однофакторного дисперсійного аналізу.

Так само, як і в задачі однофакторного дисперсійного аналізу, перевірка гіпотези H_0 ґрунтується на порівнянні двох незалежних оцінок σ^2 . При цьому одна з оцінок σ^{2*} діє поза залежністю від того, чи є вірною гіпотеза H_0 , а інша – σ^{2**} – тільки у випадку справедливості гіпотези.

Оптимальна в класі незміщених оцінок оцінка σ^{2*} може бути отримана за допомогою методу найменших квадратів. Для цього спочатку оцінимо невідомі значення параметрів μ , β_i і τ_j у моделі (3.9). А саме, знайдемо значення $\hat{\mu}$, $\hat{\beta}_i$ і $\hat{\tau}_j$ такі, що при них досягає мінімуму вираз (3.11):

$$\sum_{i,j} (x_{ij} - \mu - \beta_i - \tau_j)^2 \quad (3.11)$$

за умовою, що $\sum_{i=1}^n \beta_i = \sum_{j=1}^k \tau_j = 0$. Мінімальна величина (3.10) дорівнює $\sum_{i,j} (x_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)^2$ та виражає розкид спостережень щодо підібраних значень, що очікуються.

Розв'язання задачі (3.10) здійснюється стандартними методами математичного аналізу і приводить до наступних оцінок $\hat{\mu}$, $\hat{\beta}_i$ та $\hat{\tau}_j$:

$$\begin{aligned}\hat{\mu} = x_{..} &= \frac{1}{nk} \sum_{i=1}^n \sum_{j=1}^k x_{ij}; \hat{\beta}_i = x_{i.} - x_{..} = \frac{1}{k} \sum_{j=1}^k x_{ij} - x_{..}; \\ \hat{\tau}_j &= x_{.j} - x_{..} = \frac{1}{n} \sum_{i=1}^n x_{ij} - x_{..}\end{aligned}\quad (3.12)$$

Отримані оцінки параметрів моделі мають наступні розподіли:

$$\hat{\mu} \sim N(\mu, \frac{\sigma^2}{nk}); \hat{\beta}_j \sim N(\beta_j, \frac{\sigma^2(n-1)}{nk}); \hat{\tau}_j \sim N(\tau_j, \frac{\sigma^2(k-1)}{nk}).$$

Для одержання оцінки σ^{2*} можна використовувати величину:

$$\sum_{i=1}^n \sum_{j=1}^k \left(x_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j \right)^2 = \sum_{i=1}^n \sum_{j=1}^k \left(x_{ij} - x_{i.} - x_{.j} + x_{..} \right)^2,$$

що має розподіл $\sigma^2 \chi^2$ з числом ступенів волі $nk - (n-1) - (k-1) - 1 = (n-1)(k-1)$. Сама оцінка σ^{2*} дорівнює:

$$\sigma^{2*} = \frac{1}{(n-1)(k-1)} \sum_{i=1}^n \sum_{j=1}^k \left(x_{ij} - x_{i.} - x_{.j} + x_{..} \right)^2. \quad (3.13)$$

Вираз (3.13) дає незміщену оцінку σ^{2*} , що є справедливою як при виконанні гіпотези H_0 , так і при її порушенні.

Для одержання другої оцінки величини σ^2 , незалежної від оцінки σ^{2*} , скористаємося тим, що випадкові величини $x_{.1}, \dots, x_{.k}$, є середніми значеннями за відповідними стовпцями таблиці двохфакторного аналізу, при нульовій гіпотезі незалежні й однаково розподілені за но-

рмальним законом $N(\mu, \sigma^2 / n)$. На їх основі ми стандартним способом можемо сконструювати статистику для оцінки σ^2 : $n \sum_{j=1}^k (x_{.j} - x_{..})^2$, що має розподіл $\sigma^2 \chi^2$ з $(k-1)$ ступенями волі. При цьому сама оцінка σ^{2**} є:

$$\sigma^{2**} = \frac{n}{k-1} \sum_{j=1}^k (x_{.j} - x_{..})^2. \quad (3.14)$$

При H_0 вираз (3.13) теж дає незміщену оцінку σ^2 . При порушенні ж H_0 статистика (3.13) приймає тенденцію до збільшення – тим більшу, чим більше розходження між ефектами обробки $\tau_1, \tau_2, \dots, \tau_k$.

Складаючи F -відношення двох оцінок дисперсій, отримаємо:

$$F = \frac{\frac{n}{k-1} \sum_{j=1}^k (x_{.j} - x_{..})^2}{\frac{1}{(n-1)(k-1)} \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - x_{i.} - x_{.j} + x_{..})^2}. \quad (3.15)$$

При гіпотезі величина F має F -розподіл з числом ступенів волі $(k-1)$ та $(n-1)(k-1)$. Критерій для перевірки гіпотези H_0 має при цьому наступний вигляд:

- відкинути гіпотезу H_0 на рівні значимості α , якщо $F \geq F_{1-\alpha}$;
- не відкидати гіпотезу H_0 на рівні значимості α , якщо $F < F_{1-\alpha}$.

Тут $F_{1-\alpha}$ позначає квантиль рівня $(1 - \alpha)$ F -розподілу з числом ступенів волі $((k-1)$ та $(n-1)(k-1))$.

3.2.7 Порядок виконання роботи

3.2.7.1 Порядок виконання однофакторного дисперсійного аналізу в пакеті Statgraphics

Проілюструємо виконання однофакторного ДА в пакеті Statgraphics Plus на основі наступного приклада. Для з'ясування впливу грошового стимулювання на продуктивність праці шістьом однорі-

дним групам з п'яти чоловік кожна, були запропоновані задачі однакової трудності. Задачі пропонувалися кожному випробовуваному незалежно від всіх інших. Групи відрізняються між собою величиною грошової винагороди за розв'язувану задачу. У табл. 3.3 показана кількість вирішених задач членами кожної групи.

Таблиця 3.3 – Величина винагороди (від меншої до більшої)

Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Перевіримо гіпотезу про відсутність впливу грошової винагороди на число задач, що розв'язуються.

Для виконання дисперсійного аналізу скористаємося пакетом Statgraphics.

Перетворимо вихідні дані до вигляду, представленому на рис.3.1 і занесемо в елементарну таблицю.

	Group	Label	Col_3
1	10	1	
2	11	1	
3	9	1	
4	13	1	
5	7	1	
6	8	2	
7	10	2	
8	16	2	
9	13	2	
10	12	2	
11	12	3	
12	17	3	
13	14	3	
14	9	3	
15	16	3	
16	12	4	
17	15	4	
18	16	4	
19	16	4	
20	19	4	
21	24	5	
22	16	5	
23	22	5	
24	18	5	
25	20	5	
26	19	6	
27	18	6	
28	27	6	
29	25	6	
30	24	6	

Рисунок 3.1 – Вигляд таблиці з даними для однофакторного дисперсійного аналізу

Виконання аналізу. Для виконання однофакторного дисперсійного аналізу необхідно вибрати Compare ► Analysis of Variance ► One-Way ANOVA (однофакторний дисперсійний аналіз) у рядку меню. З'явиться діалогове вікно (див. рис. 3.2).

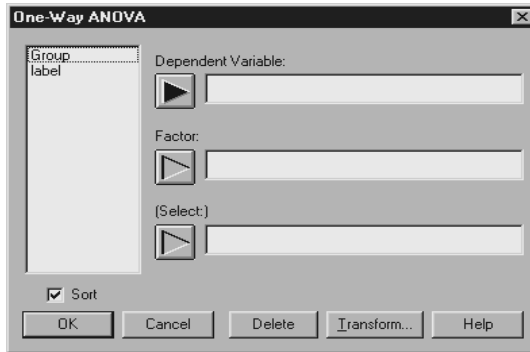


Рисунок 3.2 – Діалогове вікно однофакторного дисперсійного аналізу

Дане діалогове вікно містить текстові поля Dependent Variable (Залежна Змінна), Factor (Фактор), Select (Вибір), а також перемикач Sort (Сортувати).


Після натискання кнопки ОК у діалоговому вікні One-Way ANOVA на екран будуть видані підсумки аналізу (Analysis Summary) (див. рис. 3.3).



Рисунок 3.3 – Вікно підсумків аналізу

У перекладі з англійської мови зміст вікна наступний:

- підсумки Аналізу;
- залежна змінна: Group;
- фактор: label;
- кількість спостережень: 30;
- кількість рівнів: 6.

Далі необхідно натиснути кнопку  табличних (текстових) опцій Tabular Options на панелі інструментів вікна аналізу, після чого з'явиться діалогове вікно доступних текстових результатів (див. рис. 3.4).

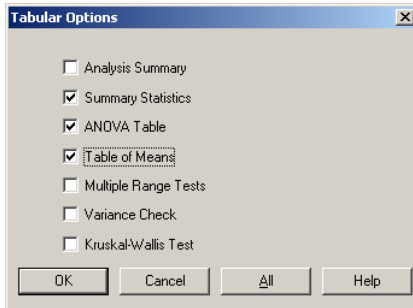


Рисунок 3.4 – Діалогове вікно доступних табличних опцій

Розглянемо призначення необхідних для нашого аналізу пунктів даного вікна.

Analysis Summary (Підсумки Аналізу) – розглянуті вище.

Summary Statistics (Підсумкова Статистика) – відображається статистична інформація щодо центра, розкиду, і форми даних: кількість значень (Count) змінної, середнє значення (Average), дисперсія (Variance), стандартне відхилення (Standard deviation), загальний підсумок (Total). Ця інформація є корисною, коли виникає необхідність визначити, чи можуть відповідні дані використовуватися в інших статистичних аналізах або дані необхідно перетворити.

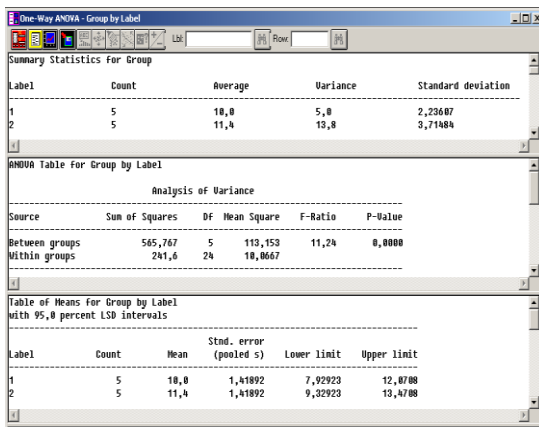
ANOVA Table (Таблиця Дисперсійного Аналізу) – відображається стандартний дисперсійний аналіз (ANOVA) у вигляді таблиці. Дана таблиця має значення для визначення відмінності (дисперсії) між групами та у межах груп. Сума квадратів між групами визначає відмінність серед середніх значень різних рівнів фактора. Загальна сума квадратів визначає відмінність усіх даних навколо головного серед-

нього значення. Кожне середньоквадратичне – це сума квадратів значень, що спостерігаються (джерела змін), розділена на число ступенів волі (df) джерела. F-відношення – це середньоквадратичне значення між групами, розділене на середньоквадратичне значення в межах груп. *P*-значення визначає рівень значущості. Маленькі рівні значимості (менше, ніж 0,05) вказують, що середні значення вибірок значимо відрізняються.

Table of Means (Таблиця Середніх Значень) – відображається таблиця, що показує кількість спостережень (Count) на кожному рівні, середнє значення (Average), об'єднані стандартні помилки, нижні і верхні границі довірчих інтервалів для середніх значень.

Multiple Range Tests (Множинні Порівняння) – відображається аналіз множинних порівнянь. Дані тести досліджують розходження між рівнями фактора. Дана опція є корисною, коли необхідно вибрати попарні порівняння між середніми значеннями для різних рівнів кожного фактора.

Для нашого приклада необхідно вибрати пункти Summary Statistics (Підсумкова Статистика), ANOVA Table (Таблиця Дисперсійного Аналізу), Table of Means (Таблиця Середніх Значень), Variance Check (Перевірка Дисперсії), відзначивши їх галочками, і натиснути кнопку ОК. На екрані відобразиться вікно аналізу, у якому будуть наведені результати аналізу в згорнутому (мінімізованому) вигляді (див. рис. 3.5).



Summary Statistics for Group

Label	Count	Average	Variance	Standard deviation
1	5	10,0	5,0	2,23607
2	5	11,4	10,8	3,27188

ANOVA Table for Group by Label

Analysis of Variance

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-Value
Between groups	505,767	5	101,153	11,24	0,0000
Within groups	241,6	24	10,0667		

Table of Means for Group by Label

With 95,0 percent LSD intervals

Label	Count	Mean	Std. error (pooled s)	Lower limit	Upper limit
1	5	10,0	1,41492	7,92923	12,0708
2	5	11,4	1,41492	9,32923	13,4708

Рисунок 3.5 – Текстові результати аналізу в згорнутому вигляді

Розглянемо підсумкову статистику (див. рис. 3.6).

Summary Statistics for Group		
label	Count	Average
1	5	10,0
2	5	11,8
3	5	13,6
4	5	15,6
5	5	20,0
6	5	22,6
Total	30	15,6
label	Variance	Standard deviation
1	5,0	2,23607
2	9,2	3,03315
3	10,3	3,20936
4	6,3	2,50998
5	10,0	3,16228
6	15,3	3,91152
Total	28,1103	5,30192
label	Minimum	Maximum
1	7,0	13,0
2	8,0	16,0
3	9,0	17,0
4	12,0	19,0
5	16,0	24,0
6	18,0	27,0
Total	7,0	27,0
label	Std. skewness	Std. kurtosis
1	0,0	0,0912071
2	0,206093	-0,0636335
3	-0,553072	-0,311059
4	-0,170962	0,686551
5	0,0	-0,547723
6	-0,260463	-1,14708
Total	0,929109	-0,570408
label	Sum	
1	50,0	
2	59,0	
3	68,0	
4	78,0	
5	100,0	
6	113,0	
Total	468,0	

Рисунок 3.6 – Підсумкова статистика (Summary Statistics)

У таблиці підсумкової статистики показані статистичні дані для кожної з груп, що відповідають одному рівню фактора, а також аналогічні дані для загального набору (значення в рядку Total):

- label – рівень фактора, спосіб обробки (назва даного стовпця визначається за ім'ям відповідної змінної кодів рівнів);
- Count – кількість спостережень на даному рівні;
- Average – середнє значення на даному рівні;
- Variance – дисперсія на даному рівні;
- Standard Deviation – стандартне відхилення на даному рівні;
- Minimum – найменше значення на даному рівні;
- Maximum – найбільше значення на даному рівні;
- Std. Skewness – нормований коефіцієнт асиметрії на даному рівні;
- Std. Kurtosis – нормований коефіцієнт ексцесу на даному рівні;
- Sum – сума значень на даному рівні .

Розглянемо таблицю ДА (див. рис. 3.7). Її призначення – дати відповідь на питання про наявність значимого впливу рівнів факторів на досліджуваний відгук чи, іншими словами, про присутність ефектів обробки.

ANOVA Table for Group by label

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	590,0	5	118,16	12,64	0,0000
Within groups	224,4	24	9,35		
Total (Corr.)	815,2	29			

Рисунок 3.7 – Базова таблиця дисперсійного аналізу

Наведемо переклад термінів, що фігурують у даній таблиці:

- Source – джерело варіації;
- Between groups – між групами;
- Within groups – усередині груп;
- Total (Corr.) – разом (скореговане значення);
- Sum of Squares – сума квадратів;
- Df – ступені волі;
- Mean Square – середні квадрати (дисперсії);
- F-Ratio – F -відношення (коефіцієнт Фішера);
- P-Value – рівень значимості.

Пояснимо значення величин, що містяться в таблиці (рис. 3.7). У рядку *Between groups* виводяться характеристики, пов'язані з дією фактора, що аналізується: сума квадратів між групами $\sum_{j=1}^k n_j (x_{.j} - \bar{x})^2$

(визначення величин $x_{.j}$ і \bar{x} подані в 3.2), відповідна кількість сту-

пенів волі (*Df*) і частка цих величин, тобто оцінка σ^{2**} (3.6). У рядку *Within groups* виводяться сума квадратів усередині груп

$\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - x_{.j})^2$, відповідна кількість ступенів волі і частка цих вели-

чин, тобто оцінка σ^{2*} (3.2). У рядку *Total (Corr.)* виводиться сума квадратів відхилень спостережень від їхнього середнього значення і число ступенів волі розподілу цієї величини при виконанні нульової гіпотези. У стовпці *F-Ratio* виводиться значення *F*-статистики (3.7), а в стовпці *P-Value* – її рівень значимості. Якщо ця величина близька до нуля, то є підстава відкинути нульову гіпотезу.

У нашому випадку *p-value* < 0,05, тому ми відкидаємо H_0 . І робимо висновок, що фактор впливає на відгук.

3.2.7.2 Порядок виконання двохфакторного дисперсійного аналізу в мові програмування R та середовища R-Studio

В роботі необхідно використати бібліотеки R з функціями, які наведені у таблиці 3.1.

Таблиця 3.1 – Перелік стандартних функцій та команд R

Назва бібліотеки	Опис
<i>tools</i>	Допоміжна бібліотека, для додаткового для встановлення та запуску додаткового інструментарію.
<i>HSAUR2</i>	Функції, набори даних, аналіз та приклади для другого видання посібника «Керівництво по статистичному аналізу з використанням R».
<i>ggplot2</i>	Набір інструментів для візуалізації даних, використовуючи граматику графіки
<i>doBy</i>	Набір інструментів для роботи із груповими статистиками.

Для встановлення бібліотеки використовується конструкція:

```
install.packages("lib_name")
```

Для запуску бібліотеки використовується конструкція:

```
library(lib_name)
```

Як приклад розглянемо результати реального експерименту [], в якому лабораторним щурам приблизно однакового віку і ваги протягом певного часу давали корм з різним вмістом білка (фактор type з двома рівнями: низький вміст – Low, і високий вміст – High). Крім того, корми розрізнялися за походженням білка (фактор source з двома рівнями: beef – яловичина, і cereal – злаки). Наприкінці експерименту було виміряно приріст ваги у щурів (weightgain) у кожній з цих груп. Таблиця з даними з цього експерименту входить до складу пакету HSAUR2.

Для початку запустимо бібліотеку HSAUR2 та підтягнемо таблицю даних weightgain.

```
library(HSAUR2)
data(weightgain)
#Переглянемо структуру таблиці
str(weightgain)
```

```
'data.frame': 40 obs. of 3 variables:
 $ source      : Factor w/ 2 levels "Beef","cereal": 1 1 1 1 1 1 1 1 1 ...
 $ type        : Factor w/ 2 levels "High","Low": 2 2 2 2 2 2 2 2 2 ...
 $ weightgain: int 90 76 90 64 86 51 72 90 95 78 ...
```

Рисunek 3.8 – Структура таблиці weightgain

Перед виконанням будь-якого статистичного аналізу корисно розглянути дані на графіку:

```
library(ggplot2)
ggplot(data = weightgain, aes(x = type, y = weightgain)) +
geom_boxplot(aes(fill = source))
```

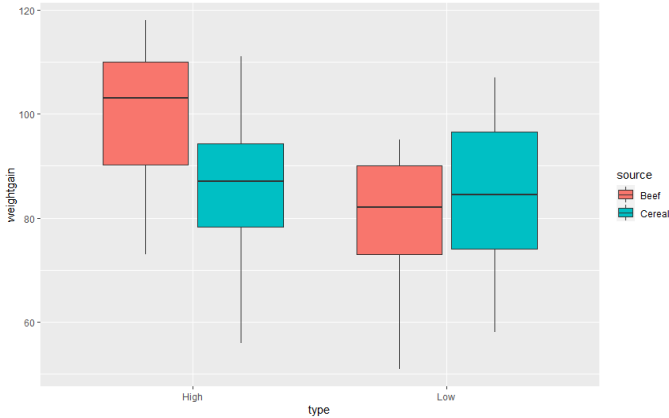


Рисунок 3.9 – Графік розподілення даних

Додатково ознайомимося зі зведеними описовими статистиками:

```
require(doBy)
summaryBy(weightgain ~ type + source, data = weightgain,
FUN = c(mean, sd, length))
```

	type	source	weightgain.mean	weightgain.sd	weightgain.length
1	High	Beef	100.0	15.13642	10
2	High	Cereal	85.9	15.02184	10
3	Low	Beef	79.2	13.88684	10
4	Low	Cereal	83.9	15.70881	10

Рисунок 3.10 – Описові характеристики

Середні значення приросту ваги в досліджених групах помітно варіюють (видно, наприклад, що приріст ваги у тварин, яких годували кормом з низьким вмістом білка тваринного походження, виявився істотно нижче, ніж у групі «High – Beef»). Завдання двофакторного дисперсійного аналізу – з'ясувати, чи пов'язані спостережувані відмінності в прирості ваги з досліджуваними факторами, або ці відмінності випадкові і не мають ніякого відношення до вмісту білка в кормі і його походження.

Корисним прийомом, що дозволяє краще зрозуміти аналізовані ефекти, є також побудова графіку дизайну експерименту (англ. design

plot). На такому графіку відображаються середні значення змінної-відгуку відповідно до кожного рівня досліджуваних факторів:

`plot.design(weightgain)`

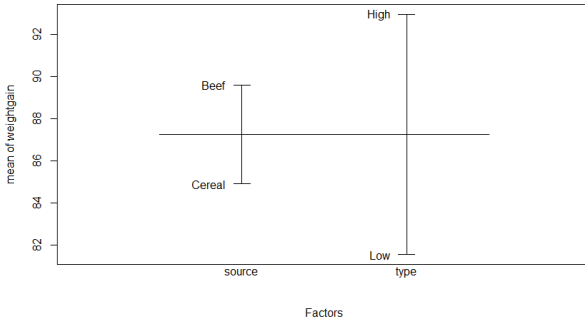


Рисунок 3.11 – Графік дизайну експерименту

З отриманого графіка видно, що найбільша різниця в середніх приростах ваги шурів пов'язана з рівнем вмісту білка в кормі, тоді як ефект джерела походження білка виражений в меншій мірі.

Розглянутий експеримент ми можемо віднести до так званого повнофакторного експерименту, оскільки в ньому реалізуються всі можливі поєднання наявних рівнів факторів. Значна перевага такого дизайну експерименту полягає в тому, що він дозволяє з'ясувати наявність взаємодії між досліджуваними факторами. В рамках дисперсійного аналізу, під взаємодією розуміють таку ситуацію, коли змінна-відгук поводить себе по-різному при різних поєднаннях досліджуваних факторів. Зрозуміти цю концепцію допоможе «графік взаємодій» (interaction plot), який в R можна побудувати за допомогою базової функції:

```
with(weightgain, interaction.plot(x.factor = type, trace.factor = source, response = weightgain))
```

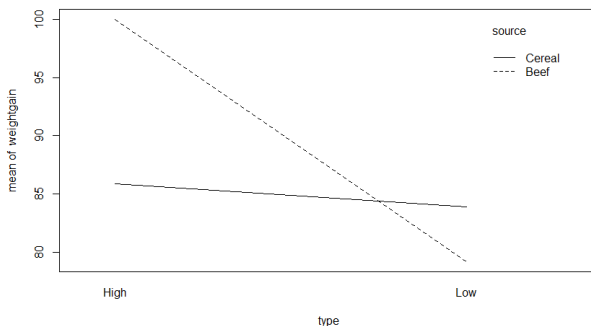


Рисунок 3.12 – Графік взаємодій

З наведеного малюнка видно, що при високому вмісті білка в кормі, приріст ваги шурів в середньому також високий, але за умови, що цей білок має тваринне походження. Якщо ж вміст білка низький, то ситуація змінюється на протилежну – приріст виявляється трохи вище (хоча і не набагато) в групі шурів, які отримували корм рослинного походження.

Крім того, що розглянутий експеримент є повнофакторним, у всіх чотирьох групах є також однакове число шурів (по 10 в кожній), тобто ми маємо справу зі збалансованим набором даних. Як було показано раніше, для аналізу збалансованих наборів даних ми можемо застосувати класичний спосіб розкладання загальної дисперсії в даних на окремі складові, реалізований у функції `aov()`:

```
M1 <- aov(weightgain ~ source + type + source:type, data =
weightgain)
```

```
summary(M1)
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
source  1    221    220.9   0.988 0.3269
type    1   1300   1299.6   5.812 0.0211 *
source:type 1    884    883.6   3.952 0.0545 .
Residuals 36   8049    223.6

---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.13 – Виконання дисперсійного аналізу за допомогою функції `aov()`

В цілому, можна зробити висновок про відсутність статистично значущої зв'язку між приростом ваги шурів і джерелом білка в кормі ($P = 0,3269$), тоді як вплив рівня вмісту білка в кормі виявилося зна-

чим ($P = 0,0211$). Взаємодія між джерелом походження білка і рівнем його вмісту в кормі незначна ($P = 0,0545$), що, в принципі, узгоджується з результатом аналізу наведеного вище графіка взаємодій.

Зверніть увагу на те, як у формулі, поданій на функцію `aov()`, було задано взаємодію між двома факторами: спочатку були наведені два головних фактори, розділені знаком «+», а потім до них додано вираз «source:type». Це стандартний синтаксис для такого роду аналізу в R. однак наведену формулу можна було б також скоротити до `weightgain ~ source * type` – результат виявився б ідентичним.

При аналізі незбалансованих наборів даних, спосіб виконання дисперсійного аналізу, реалізований у функції `aov()`, буде давати зміщені оцінки Р-значень. У таких випадках слід використовувати функцію `lm()`. Перевага цієї функції полягає ще і в тому, що вона дозволяє краще зрозуміти, де саме лежать відмінності між порівнюваними групами. Застосуємо функцію `lm()` щодо даних по приросту ваги у шурів:

```
M2 <- lm(weightgain ~ type*source, data = weightgain)
summary(M2)
lm(formula = weightgain ~ type * source, data = weightgain)
```

```
Call:
lm(formula = weightgain ~ type * source, data = weightgain)

Residuals:
    Min       1Q   Median       3Q      Max
-29.90  -9.90   2.05  10.85  25.10

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      100.000     4.729   21.148 < 2e-16 ***
typeLow           -20.800     6.687   -3.110  0.00364 **
sourceCereal     -14.100     6.687   -2.109  0.04201 *
typeLow:sourceCereal  18.800     9.457    1.988  0.05447 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.95 on 36 degrees of freedom
Multiple R-squared:  0.23,    Adjusted R-squared:  0.1658
F-statistic: 3.584 on 3 and 36 DF, p-value: 0.02297
```

Рисунок 3.14 – Виконання дисперсійного аналізу за допомогою функції `lm()`

Відповідно до розглянутих раніше принципів, отримані параметри моделі ми інтерпретуємо наступним чином. У першому рядку таблиці із параметрами моделі (Intercept) представлена інформація, що відноситься до середнього значення приросту ваги в групі шурів, яким

давали корм з високим вмістом білка (High) тваринного походження (Beef) (для простоти позначимо цю базову групу як «High – Beef»). Бачимо, що середній приріст ваги в цій групі склав 100 г, і що цей приріст значимо відрізняється від 0 (стандартна помилка = 4,729 г, значення t-критерію Стюдента = 21,148, Р-значення << 0,0001). Другий рядок (typeLow) містить оцінку величини ефекту, пов'язаного з низьким вмістом в кормі білка тваринного походження: в цій групі щурів приріст ваги виявився істотно нижче (на 20.8 г, $P = 0,00364$), ніж в групі «High – Beef» ($100 - 20,8 = 79,2$ г). З третього рядка таблиці (sourceCereal) ми дізнаємося, що у щурів, яким давали корм з високим вмістом білка рослинного походження приріст ваги був статистично значимо нижче (на 14,1 г, $P = 0,04201$), ніж в групі «High – Beef» ($100 - 14,1 = 85,9$ г). Нарешті, в четвертому рядку ми знаходимо оцінку ефекту взаємодії між факторами «type» та «source» – 18,8 г. Склавши перші три коефіцієнти з ефектом взаємодії, отримаємо середнє значення приросту ваги щурів в групі, яким давали корм з низьким вмістом білка рослинного походження: $100 - 20,8 - 14,1 + 18,8 = 83,9$. Як бачимо, цей середній приріст лише незначно ($P = 0,05447$) перевищив такий у групі щурів, яким давали корм з низьким вмістом білка рослинного походження (83,9 та 79,2). Як бачимо, отримані результати добре узгоджуються з враженнями, які ми отримали в ході розрахунку описових статистик по кожній групі, а також при виконанні графічного розвідувального аналізу даних.

Результати дисперсійного аналізу, виконаного за допомогою функції `lm()`, можна представити також і у вигляді класичної ANOVA-таблиці. Для цього відповідну модель необхідно подати на функцію `anova()`:

```
anova(M2)
```

```
Analysis of Variance Table

Response: weightgain
Df Sum Sq Mean Sq F value Pr(>F)
type      1 1299.6   1299.60    5.8123  0.02114 *
source     1   220.9    220.90    0.9879  0.32688
type:source 1   883.6    883.60    3.9518  0.05447 .
Residuals 36 8049.4    223.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.15 – Представлення результатів у вигляді класичної ANOVA-таблиці

В ході специфікації моделі M2 ми спочатку вказали фактор «type», а потім «source» (weightgain ~ type*source). Спробуємо змінити порядок.

```
M3 <- lm(weightgain ~ source*type, data = weightgain)
anova(M3)
```

```
Analysis of Variance Table

Response: weightgain
      Df Sum Sq Mean Sq F value    Pr(>F)
source  1  220.9   220.90   0.9879  0.32688
type    1 1299.6  1299.60   5.8123  0.02114 *
source:type 1  883.6   883.60   3.9518  0.05447 .
Residuals 36 8049.4   223.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.16 – Зміна порядку перерахування факторів у формулі моделі

Результати ідентичні тим, що були отримані для моделі M2. Єдина різниця полягає лише в тому, що рядки, що містять інформацію по факторам «type» і «source» помінялися місцями. Однак, як було зазначено вище, такий повний збіг результатів буде спостерігатися тільки при роботі зі збалансованими наборами даних. Для демонстрації цього твердження змінимо вихідні дані шляхом видалення, наприклад, перших 6 спостережень і останніх 7 спостережень:

```
set.seed(123)
weightgain2 <- weightgain[-c(1:6, 34:40), ]
# Моделі з різним порядком вказівки предикторів:
M4 <- lm(weightgain ~ type*source, data = weightgain2)
M5 <- lm(weightgain ~ source*type, data = weightgain2)
# ANOVA-таблиця для моделі M4
anova(M4)
```

```
Analysis of Variance Table

Response: weightgain
      Df Sum Sq Mean Sq F value    Pr(>F)
type    1  758.0   758.02   3.1655  0.08843 .
source  1  584.8   584.76   2.4419  0.13179
type:source 1  744.5   744.54   3.1092  0.09113 .
Residuals 23 5507.6   239.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.17 – ANOVA-таблиця для моделі M4

ANOVA- таблиця для моделі M5

anova(M5)

```
Analysis of Variance Table

Response: weightgain
          Df Sum Sq Mean Sq F value Pr(>F)
source    1 1188.8  1188.83   4.9645 0.03594 *
type      1   153.9   153.95   0.6429 0.43087
source:type 1   744.5   744.54   3.1092 0.09113 .
Residuals 23 5507.6   239.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Рисунок 3.18 – ANOVA-таблиця для моделі M5

Як бачимо, висновки, зроблені на підставі цих двох нових моделей виявляються абсолютно різними: жоден з параметрів моделі M4 не є статистично значущим, тоді як модель M5 вказує на існування істотного ефекту джерела походження білка. Така різниця обумовлена алгоритмом, за яким відбувається декомпозиція загальної дисперсії при аналізі незбалансованих наборів даних. Тож як визначити, яка з останніх двох моделей відображає дійсність.

Значною мірою це визначається тим, як саме ми хочемо представити результати аналізу. Якщо дисперсійний аналіз розглядати як окремий випадок загальної лінійної моделі і представляти його результати у вигляді таблиці з оцінками параметрів такої моделі, тоді ніякої різниці між M4 і M5 немає (відповідні оцінки лише змінюють своє положення в таблиці результатів-це ще раз підкреслює універсальність концепції «загальна лінійна модель»):

summary(M4)

```
Call:
lm(formula = weightgain ~ type * source, data = weightgain2)

Residuals:
    Min       1Q   Median       3Q      Max
-27.00 -10.82    2.00   11.18   23.10

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    100.000      4.893   20.435 3.02e-16 ***
typeLow        -16.250      9.155   -1.775  0.0891 .
sourceCereal   -24.000     10.187   -2.356  0.0274 *
typeLow:sourceCereal  24.150     13.696    1.763  0.0911 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.47 on 23 degrees of freedom
Multiple R-squared:  0.2748,    Adjusted R-squared:  0.1802
F-statistic: 2.906 on 3 and 23 DF,  p-value: 0.05643
```

Рисунок 3.19 – Представлення дисперсійного аналізу M4 як окремого випадку загальної лінійної моделі і представлення його результатів у вигляді таблиці з оцінками параметрів такої моделі

summary(M5)

```

Call:
lm(formula = weightgain ~ source * type, data = weightgain2)

Residuals:
    Min       1Q   Median       3Q      Max
-27.00 -10.82   2.00  11.18  23.10

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    100.000     4.893   20.435 3.02e-16 ***
sourceCereal    -24.000     10.187   -2.356  0.0274 *
typeLow         -16.250     9.155   -1.775  0.0891 .
sourceCereal:typeLow  24.150     13.696   1.763  0.0911 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.47 on 23 degrees of freedom
Multiple R-squared:  0.2748,    Adjusted R-squared:  0.1802
F-statistic: 2.906 on 3 and 23 DF,  p-value: 0.05643

```

Рисунок 3.20 – Представлення дисперсійного аналізу M5 як окремого випадку загальної лінійної моделі і представлення його результатів у вигляді таблиці з оцінками параметрів такої моделі

Якщо ж ми хочемо представити результати аналізу незбалансованих даних у вигляді класичної ANOVA-таблиці, слід приділити особливу увагу розвідувального аналізу даних. Особливо корисним, зокрема, буде графік дизайну експерименту. Той фактор, який, судячи з такого графіку, має найбільший вплив на досліджувану змінну-відгук, слід включати в модель першим. Для нашої видозміненої таблиці з даними отримуємо:

```
plot.design(weightgain2)
```

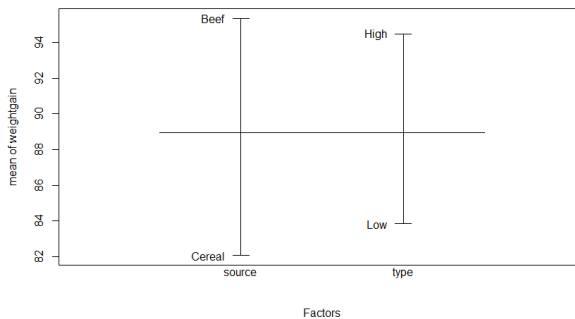


Рисунок 3.21 – Графік дизайну експерименту для таблиці weightgain2

Оскільки джерело походження тепер має більший ефект на приріст ваги щурів, ніж рівень вмісту білка, перевагу слід віддати моделі M5.

3.3 Завдання на лабораторну роботу

3.3.1 Використовуючи рекомендовану літературу та дані методичні вказівки, вивчити основні положення дисперсійного аналізу (ДА).

3.3.2 Для однофакторного ДА згенерувати дані таким чином: кількість груп $n = Var$, кількість вимірів у кожній групі

$$k = \begin{cases} Var, Var < 6 \\ Round(Var / 2), 6 \leq Var < 15 \\ Round(Var * 2 / 10), Var \geq 15 \end{cases}; \text{кількість вирішених задач – для}$$

пакета Statgraphics – $Rnormal(n*k, v1, Var/10)$.

3.3.3 Зберегти отриману вибірку у форматі .xls(Excel).

3.3.4 Для двохфакторного дисперсійного аналізу, з використанням мови програмування R:

Бібліотека	Вибірка	Приклад функцій
datarium	jobsatisfaction	<code>data("jobsatisfaction", package = "datarium")</code>
		<code>summaryBy(score ~ gender + education_level, data = jobsatisfaction, FUN = c(mean, sd, length))</code>
		<code>with(jobsatisfaction, interaction.plot(x.factor = education_level, trace.factor = gender, response = score))</code>
		<code>Model<-lm(score ~ gender*education_level, data = jobsatisfaction)</code>

3.3.5 Виконати ДА. Зробити висновки.

3.4 Зміст звіту

- 3.4.1 Назва та мета роботи.
- 3.4.2 Постанова завдання.
- 3.4.3 Таблиця вихідних даних.
- 3.4.4 Основні формули розрахунку параметрів та критеріїв.
- 3.4.5 Таблиці “Підсумкова статистика” та “Базова таблиця дисперсійного аналізу” з коментаріями.
- 3.4.6 Програма та результати виконання ДА з використанням мови програмування R та середовища R-Studio.
- 3.4.7 Висновки за результатами аналізу.

3.5 Контрольні запитання

- 3.5.1 Постановка задачі однофакторного дисперсійного аналізу.
- 3.5.2 Терміни і визначення, які використовуються при однофакторному аналізі.
- 3.5.3 Накреслити форму таблиці вихідних статистичних даних.
- 3.5.4 Що таке адитивна модель? Записати її аналітичний вираз.
- 3.5.5 Запишіть аналітичні вирази для обчислення середнього по стовпцям та по всій таблиці вихідних даних.
- 3.5.6 Запишіть аналітичні вирази для обчислення дисперсії по стовпчикам та по всій таблиці вихідних даних.
- 3.5.7 Запишіть аналітичні вирази для обчислення критерію Фішера.
- 3.5.8 В яких випадках нульова гіпотеза відкидається?
- 3.5.9 Зв'язок задач двохфакторного та однофакторного дисперсійного аналізу
- 3.5.10 Постановка задачі двохфакторного дисперсійного аналізу.
- 3.5.11 Накреслити форму таблиці вихідних статистичних даних.
- 3.5.12 Що таке адитивна модель? Записати її аналітичний вираз.
- 3.5.13 Запишіть аналітичні вирази для обчислення оцінок параметрів моделі.
- 3.5.14 Запишіть аналітичні вирази для обчислення дисперсії σ^{2*} та σ^{2**} .
- 3.5.15 Запишіть аналітичний вираз для обчислення критерію Фішера.

4 ЛАБОРАТОРНА РОБОТА № 4 МЕТОДИ ВИВЧЕННЯ ВЗАЄМОЗВ'ЯЗКІВ

4.1 Мета роботи

Вивчити методику кореляційного та лінійного регресійного аналізу. Ознайомитися з можливостями пакетів Statgraphics та мови програмування R для вирішення задач кореляційного та регресійного аналізу.

4.2 Короткі теоретичні відомості

Перед дослідженням конкретного виду зв'язку між змінними, тобто перед оцінюванням невідомих параметрів Θ у співвідношенні типу (4.1):

$$M(y|x) = f(x; \Theta) \quad (4.1)$$

де M – математичне сподівання; x – змінна; y – відгук; θ – параметри моделі, необхідно з'ясувати, чи існує взагалі цей зв'язок. Якщо так, то необхідно встановити ступінь тісноти цього зв'язку.

4.2.1 Аналіз парних зв'язків: коефіцієнт парної кореляції

Величина r , що визначається співвідношенням (4.2), називається *коефіцієнтом кореляції* і характеризує ступінь тісноти зв'язку між випадковими компонентами y та x .

За допомогою безпосередніх обчислень, що спираються на формулу для щільності двовимірного нормального закону, можна показати, що

$$r = \frac{M[(x - M(x))(y - M(y))]}{\sqrt{D(x) \cdot D(y)}} = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y}, \quad (4.2)$$

де коваріація $\text{cov}(x, y)$ – другий центральний змішаний момент двовимірної випадкової величини (x, y) , а σ_x і σ_y – середньоквадратичні (безумовні) відхилення відносно компонент y та x .

Якщо коефіцієнт кореляції є додатнім, то це означає однаковий характер тенденції взаємопов'язаної зміни випадкових компонент у та x : зі збільшенням x ми спостерігаємо збільшення відповідних індивідуальних значень y , а отже, збільшується умовне математичне очікування $M(y|x=x)$. Від'ємне значення r говорить про протилежну тенденцію взаємопов'язаної зміни компонент x та y (зі збільшенням x зменшується $M(y|x=x)$).

4.2.2 Вибіркове значення коефіцієнта кореляції

Вибіркове значення \hat{r} коефіцієнта кореляції (тобто статистична оцінка \hat{r} невідомого r) розраховується за вхідними даними за формулою:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4.3)$$

$$\text{де } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ та } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Теоретичний і вибірковий коефіцієнти кореляції можуть бути формально обчислені для будь-якої двовимірної системи спостережень; вони є вимірювачами ступеня *лінійного* статистичного зв'язку між ознаками, що аналізуються.

Але у випадку сукупного нормального розподілу досліджуваних випадкових величин x та y , коефіцієнт r має чіткий смисл як характеристика тісноти зв'язку між ними. Якщо $|r| = 1$, то це говорить про суто функціональну лінійну залежність між досліджуваними величинами, а якщо $r = 0$, то це свідчить про відсутність лінійної залежності. При $|r| > 0,75$ лінійний зв'язок вважається суттєвим.

У всіх інших випадках (розподілення x та y відхиляється від нормального, одна із досліджуваних величин не є випадковою, тощо) коефіцієнт кореляції можна використовувати лише в якості однієї із можливих характеристик ступеню тісноти зв'язку.

4.2.3 Кореляційне відношення

Найбільш зручною є ситуація, в якій характер вибірових даних допускає їх групування по осі пояснювальної змінної і можливість обчислення так званих “окремих” середніх ординат \bar{y}_i всередині кожного (i -го) інтервалу групування. Нехай таке групування виконано. При цьому k – кількість інтервалів групування по осі абсцис; m_i ($i = 1, 2, \dots, k$) – кількість вибірових точок, що потрапили в i -ий інтервал;

$\bar{y}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$ – середнє значення ординат точок, що потрапили в i -ий інтервал групування. Тоді вибірковою оцінкою дисперсії σ_f^2 буде величина (4.4):

$$s_{y(x)}^2 = \frac{1}{n} \sum_{i=1}^k m_i (\bar{y}_i - \bar{y})^2, \quad (4.4)$$

$$\text{де загальне середнє } \bar{y} = \frac{\sum_{i=1}^k m_i \bar{y}_i}{n}.$$

У цьому випадку ми можемо отримати вираз для *кореляційного відношення* залежної змінної y за незалежною змінною x :

$$R_{y \cdot x}^2 = \frac{S_{y(x)}^2}{S_y^2}, \quad (4.5)$$

де вибіркова дисперсія s_y^2 індивідуальних результатів спостереження y_{ij} навколо загального середнього \bar{y} обчислюється за формулою:

$$s_y^2 = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{m_i} (y_{ij} - \bar{y})^2.$$

Кореляційне відношення несиметричне по відношенню до досліджуваних змінних, тобто $R_{y \cdot x} \neq R_{x \cdot y}$.

Кореляційне відношення є величиною невід'ємною.

В іншому властивості кореляційного відношення багато в чому схожі на властивості коефіцієнта кореляції: кореляційне відношення не може бути більшим одиниці.

Якщо $|R| = 1$, то це означає наявність однозначного функціонального зв'язку між y та x , і, навпаки, однозначний функціональний зв'язок між x і y говорить про те, що $|R| = 1$. Відсутність кореляційного зв'язку між x і y означає, що умовні середні $\overline{y_i}$ зберігають постійне значення, що дорівнює загальному середньому \overline{y} , а тому $R_{y/x} = 0$. Якщо $R_{y/x} = 0$, то $\overline{y_i} = \overline{y}$, і, як наслідок, часткові середні $\overline{y_i}$ не залежать від x .

4.2.4 Перевірка гіпотези про відсутність кореляційного зв'язку

Для побудови відповідного критерію перевірки даної гіпотези використаємо факт наближеної $F(k-1, n-k)$ -розподіленості випадкової величини:

$$F(0) = \frac{R_{y \cdot x}^2}{1 - R_{y \cdot x}^2} \cdot \frac{n-k}{k-1},$$

яке є справедливим в припущенні, що $R_{y \cdot x} = 0$ і що умовні розподіли залежної змінної $y(x)$ при будь-якому фіксованому x описуються нормальним законом з постійною дисперсією σ^2 .

Тому, якщо виявиться, що:

$$\frac{R_{y \cdot x}^2}{1 - R_{y \cdot x}^2} \cdot \frac{n-k}{k-1} > v_{\alpha}^2(k-1, n-k),$$

то гіпотеза про відсутність кореляційного зв'язку між x та y відкидається з рівнем значимості α (тут $\nu_{\alpha}^2(k-1, n-k) - 100\alpha\%$ -на точка F-розподілу з числом ступенів волі чисельника $k-1$ і знаменника $n-k$ знаходиться з таблиці). При виконанні зворотної нерівності значення кореляційного відношення $R_{y,x}$ є статистично незначущим, тобто робиться висновок про відсутність кореляційного зв'язку між x та y .

4.2.5 Лінійний регресійний аналіз

Лінійний регресійний аналіз поєднує широке коло задач, пов'язаних з побудовою функціональних залежностей між двома групами числових змінних: x_1, \dots, x_p та y_1, \dots, y_q . Для стислості ми об'єднаємо x_1, \dots, x_p в багатовимірну змінну x , а y_1, \dots, y_q – у змінну y , і будемо говорити про дослідження залежності між x та y . При цьому ми будемо вважати x незалежною змінною, що впливає на значення y . У зв'язку з цим ми будемо називати y – *відгуком*, а x – *факторами*, що впливають на відгук.

Постанова завдання. Статистичний підхід до задачі побудови (точніше, відновлення) функціональної залежності y від x ґрунтується на припущенні, що нам відомі деякі експериментальні дані (x_i, y_i) , де y_i – значення відгуку при заданому значенні фактора x_i , і змінюється від 1 до n . Пари значень (x_i, y_i) часто називають результатом одного виміру, а n – числом вимірів.

Ми будемо припускати, що значення відгуку y , що спостерігається в досліді, можна уявно розділити на дві частини: одна з них закономірно залежить від x , тобто є функцією від x ; інша частина – випадкова по відношенню до x . Позначимо першу через $f(x)$, другу через ε і представимо відгук у вигляді:

$$y = f(x) + \varepsilon \quad (4.6)$$

де ε – деяка випадкова величина.

Припустимо, що функція $f(x, \theta)$ має вид $\beta_0 + \beta_1 x$, тобто лінійно залежить від параметрів $\theta = (\beta_0, \beta_1)$. Ця задача має назву *лінійного регресійного аналізу*. У цьому випадку співвідношення (4.6) приймає вигляд (4.7):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (4.7)$$

де x_1, \dots, x_n – задані числа (значення фактора); y_1, \dots, y_n – спостережені значення відгуку; $\varepsilon_1, \dots, \varepsilon_n$ – незалежні однаково розподілені випадкові величини.

При вирішенні задачі лінійного регресійного аналізу використовуються два основні підходи: непараметричний і гаусовський, вони розрізняються характером припущень щодо закону розподілу випадкових величин ε . Розглянемо Гаусову модель простої лінійної регресії.

Для цього необхідно виконання наступних передумов:

- незалежність спостережень;
- однакова розподіленість помилок;
- помилки ε_i розподілені за нормальним законом $N(0, \sigma^2)$ з деякою невідомою дисперсією σ^2 .

4.2.6 Метод найменших квадратів

При виборі методів визначення параметрів регресійної моделі можна керуватися різними підходами.

Визначення. *Методом найменших квадратів* називається спосіб підбору параметрів регресійної моделі, виходячи з мінімізації суми квадратів залишків:

$$\sum_{i=1}^n [y_i - \beta_0 - \beta_1 x_i]^2 \rightarrow \min_{\beta_0, \beta_1}. \quad (4.8)$$

Геометричний зміст метода найменших квадратів проілюстровано на рис. 4.1.

Оцінки b_0 та b_1 обчислюємо за формулами:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad (i = 1..n), \quad (4.9)$$

$$\text{де } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Розв'язання рівняння щодо вільного члена (відрізка на осі ординат при $x = 0$) b_0 дає:

$$b_0 = \bar{y} - b_1 \bar{x}. \quad (4.10)$$

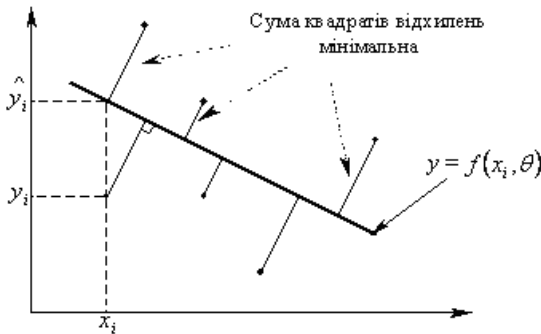


Рисунок 4.1 – Геометричний зміст методу найменших квадратів

4.2.7 Перевірка гіпотези про коефіцієнт нахилу

При вирішенні задачі лінійної регресії виникає питання про рівність нулю коефіцієнта нахилу – перевірка значущості коефіцієнта. Зі статистичної точки зору це означає перевірку гіпотези $H_0: b_1 = 0$. Важливість цієї гіпотези пояснюється тим, що в цьому випадку змінна y змінюється чисто випадково, не залежачи від значення x .

Для перевірки значущості коефіцієнтів регресії часто використовують критерій Стюдента, який в цьому випадку розраховується:

$$t = \frac{b_i}{s \sqrt{c_{ii}}}, \quad (4.11)$$

де b_i – i -й коефіцієнт рівняння регресії; s – середнє квадратичне відхилення; c_{ii} – діагональний коефіцієнт матриці дисперсій-

коваріацій, тобто матриці $(X^T X)^{-1}$, яка утворюється в процесі розв'язання системи рівнянь.

Якщо $t > t_{\alpha, V}$ (де $t_{\alpha, V}$ – табличне значення критерію Стюдента з рівнем значущості α і V ступенями волі), то коефіцієнт є значимим, у зворотному випадку – ні.

4.2.8 Аналіз рівняння регресії

Тепер ми вивчимо питання про те, яка точність може бути приписана нашій оцінці лінії регресії, що оцінюється дисперсією σ^2 помилок спостереження. Незміщену оцінку s^2 для σ^2 можливо отримати через остаточну суму квадратів (табл. 4.1).

Таблиця 4.1 – Таблица ДА

Джерело варіації	Число ступенів волі	Суми квадратів SS	Середні квадрати
Обумовлена регресією SS ₃	1	$\sum_{i=1}^n (y_i - \bar{y})^2$	$MS_{рег}$
Відносно регресії (залишок) SS ₂	$n - 2$	$\sum_{i=1}^n (y_i - \hat{y})^2$	$s^2 = \frac{SS_2}{(n - 2)}$
Загальна, скорегована на середнє Y SS ₁	$n - 1$	$\sum_{i=1}^n (y_i - \bar{y})^2$	

Як вже відзначалося, побудована лінія регресії – це розрахункова лінія, заснована на деякій моделі чи припущеннях. Але припущення ми не повинні сліпо приймати, ми повинні розглядати їх як можливі. При деяких умовах можлива перевірка коректності моделі.

В широкому сенсі під адекватністю розуміється відповідність моделі процесу чи об'єкту, який вона описує, раніше визначеним умовам.

Адекватність моделі перевіряється за критерієм Фішера:

$$F = \frac{MS_{рег}}{s^2}. \quad (4.12)$$

Якщо $F > F_{табл}$ для обраного рівня значущості q та відповідних ступенях волі ($1, n - 2$) то модель вважається адекватною, і навпаки неадекватною при $F \leq F_{табл}$.

Необхідно відзначити, що ця перевірка являється формальною, тому заключне рішення про адекватність моделі треба приймати, виходячи з придатності моделі до практичного використання за всією сукупністю параметрів.

Придатність лінії регресії для прогнозування залежить від того, яка частина SS_1 щодо середнього приходить на SS_3 , обумовлену регресією, і яка – відповідає SS_2 щодо регресії. Ми будемо задоволені, якщо SS_3 , обумовлена регресія, буде набагато більше, ніж SS_2 щодо регресії, чи (що те ж саме) якщо відношення:

$$R^2 = \frac{(SS_3, \text{обумовлена регресією})}{(\text{повна } SS, \text{скорегована на } \bar{y})} = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}, \quad i = \overline{1, n} \quad (4.13)$$

не буде занадто відрізнятися від одиниці. Тоді R^2 вимірює «частку загального розкиду щодо середнього \bar{y} , що пояснюється регресією». Її часто виражають у відсотках, помножуючи на 100. Фактично R – це кореляція між y та \hat{y} , коефіцієнт R часто називають *множинним коефіцієнтом кореляції*.

Максимально можливе значення $R^2 = 1$ (чи 100%), коли всі значення x різні. Однак на практиці величина R^2 не може досягти 1, яка би гарна не була модель. Це обумовлено не якістю моделі, а помилкою відтворюваності.

4.2.9 Приклад побудови рівняння регресії

Приклад 4.1. Залежність між величинами y і x описується рівнянням регресії $y = a_0 + a_1x$. Розрахувати коефіцієнт кореляції. Отримати коефіцієнти рівняння регресії. Перевірити адекватність моделі за допомогою коефіцієнта Фішера. Числові значення y і x наведені у таблиці 4.2.

Таблиця 4.2 – Вхідні дані

X	0,5	1	1,5	2	2,5
Y	20	50	70	100	130
N	1	2	3	4	5

Розв'язання

Знайдемо середні значення \bar{x} та \bar{y} :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i = 1,5, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^n y_i = 74.$$

Для розрахунку коефіцієнта кореляції побудуємо таблицю 4.3.

Таблиця 4.3 – Розраховані дані для отримання коефіцієнта кореляції

№ п/п	x	y	$ x_i - \bar{x} $	$ y_i - \bar{y} $	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0,5	20	1	54	1	2916	54
2	1	50	0,5	24	0,25	576	12
3	1,5	70	0	4	0	16	0
4	2	100	0,5	26	0,25	676	13
5	2,5	130	1	56	1	3136	56
Σ	7,5	370	3	164	2,5	7320	135

Обчислимо середні квадратичні відхилення $S\{x\}$ та $S\{y\}$:

$$S\{x\} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{(0,5-1,5)^2 + (1-1,5)^2 + (1,5-1,5)^2 + (2-1,5)^2 + (2,5-1,5)^2}{4}} =$$

$$= \sqrt{\frac{2,5}{4}} \approx 0,79;$$

$$S\{y\} = \sqrt{\frac{(20-74)^2 + (50-74)^2 + (70-74)^2 + (100-74)^2 + (130-74)^2}{4}} =$$

$$= \sqrt{\frac{7320}{4}} \approx 42,78.$$

Розрахуємо коефіцієнт кореляції:

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{(N-1)S\{x\}S\{y\}} = \frac{135,0}{135,18} \approx 0,999$$

Таким чином коефіцієнт кореляції між величинами x та y дорівнює 0,999, що говорить про наявність функціонально лінійного зв'язку між величинами x та y .

Розрахуємо коефіцієнти рівняння регресії:

$$a_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{135}{2,5} = 54;$$

$$a_0 = \bar{y} - a_1 \bar{x} = 74 - 54 \cdot 1,5 = -7.$$

Таким чином, залежність між величинами y і x описується рівнянням регресії: $\hat{y} = -7 + 54\hat{x}$.

Побудуємо таблицю 4.5 для розрахунку критерію Фішера.

Таблиця 4.5 – Дані для розрахунку критерію Фішера

y_i	\hat{y}_i	$(y_i - \hat{y}_i)$	$(\hat{y}_i - \bar{y})$
20	20	0	-54
50	47	3	-27
70	74	-4	0
100	101	-1	27
130	128	2	54

Перевіримо адекватність моделі за допомогою критерію Фішера:

$$F = \frac{S_1^2}{S_2^2} = \frac{(N-2) \sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \hat{y})^2} = \frac{3 \cdot 7290}{30} = 729.$$

Табличне значення критерію Фішера для рівня значущості $q=0,05$ та відповідних ступенях волі $(1, n-2)$ дорівнює: $F_{табл}(1;3;0,05)=10,3$. Оскільки ми отримали значення $F > F_{табл}$, то модель можна вважати адекватною.

Отримаємо значення множинного коефіцієнта кореляції:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{7290}{7320} = 0,995,$$

яке свідчить про придатність лінії регресії для цілей передбачення.

4.3 Порядок виконання роботи

4.3.1 Постановка завдання

Як приклад розглянемо використання лінійного регресійного аналізу в задачі відновлення залежності між входом і виходом у вимірювально-реєструючій системі. Подібні задачі широко поширені в експериментальних дослідженнях, у багатьох предметних областях вони називаються по-різному: градування, калібрування, тарировка і т.д. Розглянемо розв'язання подібної задачі за допомогою градувального експерименту – на тензоваги робиться вплив еталонною силою (моментом сил) і фіксуються значення відгуку на виході системи. Варіюючи значення еталонної сили в межах робочого діапазону тензовагів, ми одержуємо дані, по яких потрібно відтворити вид залежності між входом і виходом вимірювальної системи.

Таблиця 4.2 – Дані каліброваного експерименту однієї компоненти тензозагів

Номер еталонного навантаження i		1	2	3	4	5	6
Значення еталонної сили x_i		0.0	0.2	0.4	0.6	0.8	1.0
Значення відгуку y_{ij}	J = 1	31.0	110.0	186.5	266.7	345.5	425.6
	J = 2	29.8	111.0	191.0	269.7	349.3	425.9
	J = 3	29.1	109.6	187.1	270.1	349.7	426.5
	J = 4	29.0	111.0	190.3	270.2	349.9	426.5
	J = 5	29.15	109.6	186.7	266.55	347.05	427.0
	J = 6	28.2	110.35	190.95	270.25	349.8	427.0
Середні значення \bar{y}_i		29.38	110.26	188.76	268.92	348.54	426.42
Значення s_i^2		0,894	0,408	4,858	3,191	3,364	0,326

У таблиці 4.2 приведені дані градуовального експерименту одного компоненту тензозагів, призначеного для вимірювання сили лобового опору. У ході експерименту значення еталонної сили x змінювалися від 0 до 1 кг із кроком 0,2 кг, і для кожного значення сили реєструвалося значення відгуку y у десятках мв. Вимірювання повторювалися 6 разів. У таблиці наведено також середні відгуки \bar{y}_i , і стандартні відхилення s_i^2 .

Потрібно побудувати рівняння регресії.

4.3.2 Розв'язання задачі

Методи регресійного аналізу в пакеті Statgraphics представлені в пункті Relate головного меню. У ньому знаходяться 3 процедури (рис.4.2):

- Simple Regression (проста регресія);
- Polynomial Regression (поліноміальна регресія);
- Multiple Regression (множинна регресія).

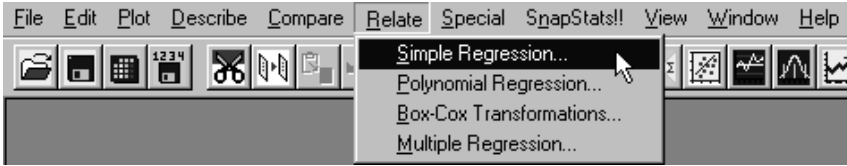


Рисунок 4.2 – Меню процедур регресійного аналізу

Для вирішення поставленої задачі ми припускаємо, що залежність буде мати вигляд прямої лінії і відповідно до цього вибираємо пункт **Simple Regression** (рис. 4.3).

Вихідні дані наведено на рисунку 4.3.

	X	Y	Col_3
1	0,0	29,38	
2	0,2	110,26	
3	0,4	188,76	
4	0,6	268,92	
5	0,8	348,54	
6	1,0	426,42	
7			

Рисунок 4.3 – Вихідні дані

В меню простої регресії потрібно вибрати спочатку поле Y (залежну змінну), а потім вибрати X – незалежну змінну (рис. 4.4). Після натискання кнопки ОК на екран будуть видані підсумки аналізу.

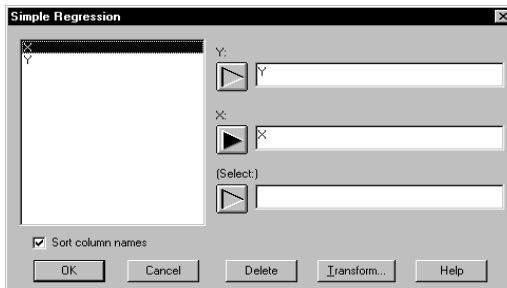


Рисунок 4.4 – Меню простої регресії

4.3.3 Результати

Вікно вибору текстових результатів можна викликати, натиснувши на другу зліва кнопку в панелі інструментів аналізу (Tabular Options). При цьому з'явиться вибір з пунктів (рис. 4.5):

- Analysis Summary – сумарний аналіз;
- Lack-of-Fit Test – перевірка невідповідності моделі;
- Forecasts – прогнози;
- Comparison of Alternative Models – порівняння альтернативних моделей;
- Unusual Residuals – незвичайні побічні впливи;
- Influential Points – точки впливу.

Нам потрібні пункти всі пункти для проведення аналізу (рис. 4.5).

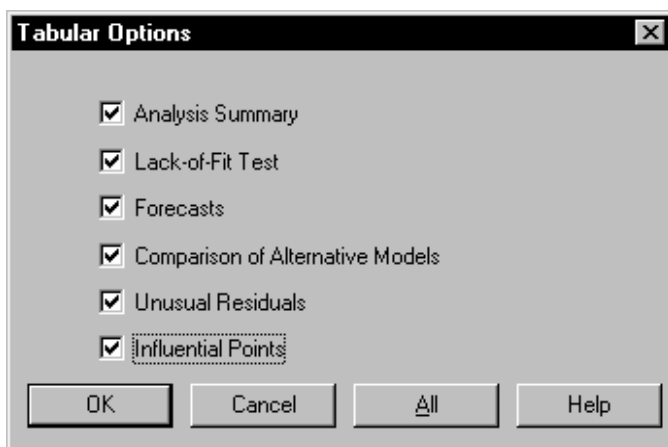


Рисунок 4.5 – Вікно вибору текстових результатів

Для аналізу результатів використовується таблиця з оцінкою коефіцієнтів регресійного рівняння та таблицею ДА (рис. 4.6).

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: Y
Independent variable: X

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	30,1276	0,580168	51,9291	0,0000
Slope	397,171	0,958116	414,534	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	110422,0	1	110422,0	171838,22	0,0000
Residual	2,57036	4	0,64259		
Total (Corr.)	110424,0	5			

Correlation Coefficient = 0,999988
R-squared = 99,9977 percent
R-squared (adjusted for d.f.) = 99,9971 percent
Standard Error of Est. = 0,801617
Mean absolute error = 0,620952
Durbin-Watson statistic = 2,31093 (P=0,1196)
Lag 1 residual autocorrelation = -0,414508

Рисунок 4.6 – Результати регресійного аналізу

На рис. 4.6 *Intercept* – вільний член – a (b_0); *Slope* – нахил – b (b_1); *Estimate* – оцінка значення параметра; *Standard Error* – середньоквадратична помилка; *T-statistic* – коефіцієнт Ст'юдента; *P-Value* – перевірка значущості зв'язку.

Стовпець *Standard Error* (стандартна помилка) містить значення стандартних помилок зазначених коефіцієнтів. Отримані рівні значущості говорять, що обидві оцінки значимо відрізняються від нуля. Таким чином $b_0 = 30,13$; $b_1 = 397,2$.

Дисперсійний аналіз показує підсумки перевірки побудованої моделі: *Sum of Squares* – сума квадратів відхилень; *Df* – ступінь волі; *Mean Square* – дисперсія; *F-ratio* – коефіцієнт Фішера; *P-Value* – характеристика ступеня значущості зв'язку між X і Y .

Таблиця ДА є базовою таблицею аналізу варіації і служить для оцінювання адекватності запропонованої моделі даних. З таблиці ДА видно, що отримане рівняння регресії є адекватним і має вигляд $y = 30,13 + 397,2 x$.

4.3.4 Відповідність моделі

Для аналізу відповідності моделі використовуємо: Correlation Coefficient – коефіцієнт парної кореляції Y від X ; R-Squared показує, який відсоток мінливості Y припускається через X ; Standard Error of Est. – середньоквадратична помилка моделювання, вплив зовнішніх факторів.

Отримане рівняння регресії пояснює 99% розкиду.

Отже робимо висновок, що модель можна використовувати для прогнозування.

Порівняння моделей: тут зазначено параметр R^2 , що визначає ступінь схожості моделі з експериментом. Треба вибрати модель з найбільшими значеннями, і потім натиснути правою кнопкою на поле аналізу. У контекстному меню вибрати пункт Analysis Options, де й установити галочку навпроти цієї моделі. Потім подивитися нові значення в Analysis Summary і Forecasts.

4.3.5 Графічні результати

Вікно вибору графічних результатів (рис. 4.7) можна викликати, натиснувши на третю зліва кнопку в панелі інструментів аналізу (Graphical Options).

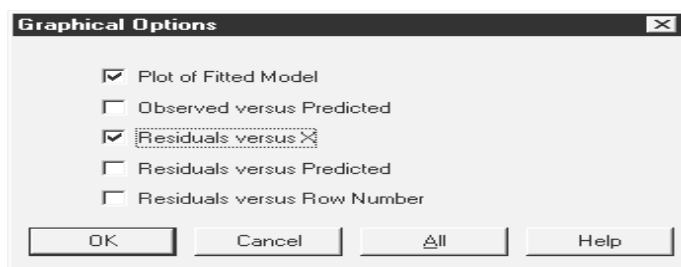


Рисунок 4.7 – Вікно вибору графічних результатів

При цьому з'явиться вибір з пунктів (рис. 4.7):

- Plot of Fitted Model – графік побудованої моделі;
- Observed versus Predicted – графік відповідності моделі й експерименту;

- Residuals versus X – залежність перешкод від X;
- Residuals versus Predicted – залежність перешкод від прогнозованого Y;
- Residuals versus Row Number – залежність перешкод від номера рядка.

Для графічного аналізу результатів використаємо графік моделі (Plot of Fitted Model) (рис. 4.8):

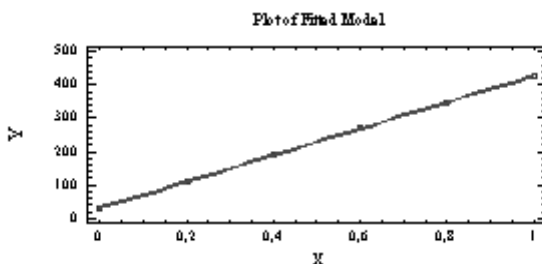


Рисунок 4.8 – Графік побудованої моделі

З графіку на рис. 4.8 бачимо, що модель підібрана добре, всі точки лежать на лінії.

4.3.6 Кореляційний та регресійний аналіз з використанням мови R

4.3.6.1 Основні функції та команди, необхідні для виконання роботи

Функція $cor(x)$ – застосовується для обчислення кореляційної матриці.

Функція $cor(x, y)$ – застосовується для обчислення коефіцієнта кореляції.

Приклад:

```
cor(x, y = NULL, use = "everything",
method = c("pearson", "kendall", "spearman"))
```

де x – матриця чи таблиця;

y – необов'язковий аргумент, матриця чи таблиця;

use – необов'язковий параметр, що надає можливість обчислення кореляції за відсутності даних (може приймати такі значення: "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs");

method – визначає тип кореляції. Може приймати такі значення: pearson, spearman(для непараметричних даних) або kendall.

Функція *str()* – для з'ясування структури об'єкту.

Функція *summary()* – загального призначення. Швидкий розрахунок основних параметрів описової статистики. Таких як мінімальне, максимальне значення, розрахунок дисперсії, середнього значення та квантилів.

Функція *with()* – слугує показником на вибірку, з якою треба працювати іншій функції.

Функція *cor.test()* – схожа з *cor()*, проте, на відміну від *cor()*, виконує ще й оцінку статистичної значимості коефіцієнтів регресії, перевіряючи нульову гіпотезу про рівність їх нулю.

Функція *lm(formula = y ~ x)* призначена для задання формули простої лінійної регресії $y=a+bx$.

Функція *summary(lm(y~x))* призначена для отримання таблиці значимості коефіцієнтів рівняння лінійної регресії $y=a+bx$.

Функція *abline(lm(y~x))* призначена для побудови графіку простої лінійної регресії $y=a+bx$ методом найменших квадратів.

Функція *anova(lm(y~x))* призначена для побудови таблиці дисперсійного аналізу.

Функція *read.csv()* – призначена для читання значень, розділених комами з файлів формату ('.csv') .

Приклад:

```
MyData <- read.csv(file="TheDataIWantToReadIn.csv", header=TRUE, sep=",", fill=TRUE)
```

Аргумент *file* – місце розташування файлу.

Аргумент *header = TRUE* означає, що в файлі присутній заголовок.

Аргумент *sep* – символ роздільник. Значення в кожному рядку файлу розділені цим символом. За замовченням *sep = ""*.

Аргумент *fill = TRUE* означає, що в файлі відключені коментарі. Функція *view()* – відображає таблицю, зчитану з файлу формату .csv.

4.3.6.2 Розрахунок коефіцієнту кореляції у R-Studio

Розглянемо приклад розрахунку коефіцієнту кореляції засобами мови R.

```
X.vector <- 5:15 # вибірка X
Y.vector <- 7:17 # вибірка Y

cor(X.vector,Y.vector) # коефіцієнт кореляції
# результат виконання:
#[1] 1

z.vector <- -3*X.vector + 8
cor(X.vector,z.vector)

# результат виконання:
#[1] -1

# Створимо дві вибірки: роки та відсотки
year <- c(2010 , 2011 , 2012 , 2013 , 2014)
rate <- c(9.34 , 8.50 , 7.62 , 6.93 , 6.60)

# будуємо графік залежності проценту за кредитом від року

plot(year,rate, main="Процентна ставка на 4 роки для кредиту на
авто")

# розраховуємо коефіцієнт кореляції
cor(year,rate)
# [1] -0.9880813
```



Рисунок 4.9 – Графік залежності проценту за кредитом від року

4.3.6.3 Побудова лінійної регресії засобами мови R

Розглянемо приклад побудови лінійної регресії засобами мови

R.

```
x = scan('D:/x.txt')
y = scan('D:/y.txt')
print(cor.test(x, y, use="complete.obs"))
p.lm<-lm(formula = y~x)
print(summary(p.lm))
plot(x,y)
abline(lm(y~x))
```

Результати виконання програми:

Pearson's product-moment correlation

data: x and y

t = -1.0939, df = 10, p-value = 0.2996

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.7585132 0.3040154

sample estimates:

cor

-0.3269267

Call:

lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-5.1580	-1.7529	0.4109	1.7730	4.6695

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.57184	1.72650	0.910	0.384
x	-0.04138	0.03783	-1.094	0.300

Residual standard error: 2.881 on 10 degrees of freedom

Multiple R-squared: 0.1069, Adjusted R-squared: 0.01757

F-statistic: 1.197 on 1 and 10 DF, p-value: 0.2996

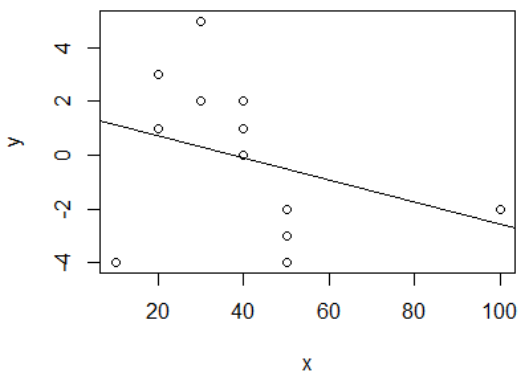


Рисунок 4.10 – Графік лінійної регресії

4.4 Завдання на лабораторну роботу

4.4.1 Одержати вихідні дані у викладача.

4.4.2 Використовуючи рекомендовану літературу та дані методичні вказівки, вивчити метод лінійного регресійного аналізу (РА) та кореляційного аналізу.

4.4.3 Вивчити можливості пакетів статистичного аналізу даних для вирішення задач РА.

4.4.4 Виконати РА, використовуючи дані, що отримані у викладача.

4.4.5 Оформити звіт.

4.4.6 Відповісти на контрольні питання.

4.5 Зміст звіту

4.5.1 Назва і мета роботи.

4.5.2 Постанова завдання.

4.5.3 Таблиця вихідних даних.

4.5.4 Основні формули розрахунку параметрів та критеріїв.

4.5.5 Таблиці кореляційного аналізу і графіки з коментаріями.

4.5.6 Висновки за результатами аналізу.

4.5.7 Вихідні дані.

4.5.8 Коефіцієнти рівнянь, таблиця ДА, рівняння регресії.

4.5.9 Графічна інтерпретація задачі регресійного аналізу.

4.5.10 Прогнози та таблиця R^2 .

4.5.11 Висновки.

4.6 Контрольні запитання

4.6.1 Що означають поняття статистична залежність, функціональна залежність, кореляційна залежність?

4.6.2 Коли використовується кореляційний аналіз?

4.6.3 Як визначається індекс кореляції? Що він показує?

4.6.4 Як визначається коваріація і коефіцієнт кореляції? Як визначити вибіркове значення коефіцієнта кореляції?

4.6.5 Як визначається кореляційне відношення? Який його фізичний зміст?

4.6.6 Яким чином перевіряється гіпотеза про відсутність кореляційного зв'язку?

4.6.7 Для чого призначений узагальнений засіб обчислення парних кореляційних характеристик?

4.6.8 Як розраховується узагальнений коефіцієнт кореляції?

4.6.9 В чому полягає мета регресійного аналізу?

4.6.10 Для чого використовується метод найменших квадратів?

4.6.11 Наведіть передумови регресійного аналізу.

4.6.12 Як виконується перевірка значущості оцінок коефіцієнтів рівняння регресії?

4.6.13 За допомогою якого критерію виконується перевірка адекватності рівняння регресії?

4.6.14 Як виконується вибір найкращої моделі рівняння регресії?

5 ЛАБОРАТОРНА РОБОТА № 5 РАНГОВИЙ АНАЛІЗ

5.1 Мета роботи

Вивчити методи рангового аналізу, використовуючи для цього пакети статистичних програм.

5.2 Короткі теоретичні відомості

В багатьох практичних задачах необхідно досліджувати об'єкти, що мають декілька (дві чи більше) ознак, і необхідно визначити, наскільки ці ознаки пов'язані між собою.

Методи визначення зв'язку ознак помітно відрізняються в залежності від вигляду шкали вимірювання цих ознак:

- для вивчення зв'язку ознак, що вимірювали у номінальній шкалі, наприклад, ознак виду “так чи ні”, застосовуються таблиці сполучень, статистика Фішера-Пірсона χ^2 , різні міри зв'язку ознак (коефіцієнти Юла, Крамера, Чупрова та ін.) та логарифмічно лінійні моделі;

- для ознак, що вимірювали у порядковій шкалі, – даних типу “краще-гірше”, тестових балів і т.д. – застосовуються ранжування та коефіцієнти кореляції Спірмена та Кендела;

- для даних, що вимірювали у кількісних шкалах, застосовуються коефіцієнт кореляції Пірсона та модель простої лінійної регресії.

Таким чином, першим кроком аналізу є класифікація типу даних, тобто відношення їх до тієї чи іншої шкали вимірювання – номінальної, порядкової чи кількісної.

5.2.1 Характеристика ознак

Методи аналізу кількісних даних ми розглядали у попередніх лабораторних роботах. Та якщо ми нічого не знаємо про розподіл спостережень, то безпосередньо використовувати для перевірки нульової гіпотези кількісні значення спостережень x_{ij} є проблематичним. У цьому випадку простіше за все спиратись у своїх висновках тільки на

відношення “більш-менш” між спостереженнями, оскільки вони не залежать від розподілу спостережень.

Взагалі конкретні числові значення величин x_{ij} є умовністю, а змістовий смисл мають лише відношення “більш-менш” між ними.

Отже, розглянемо методи аналізу даних, що засновані на рангах. Відповідні критерії для перевірки нульової гіпотези мають назву *рангові*, вони придатні для будь-яких неперервних розподілів спостережень. Більш того, вони придатні, коли вимірювання зроблені у порядковій шкалі, наприклад, є тестовими балами або експертними оцінками.

Рангом спостереження називають номер, який отримає це спостереження в упорядкованій сукупності всіх даних – після їх впорядкування за визначеним правилом (наприклад, від менших значень до більших і навпаки).

Процедура переходу від сукупності спостережень до послідовності їх рангів називається *ранжуванням*. Результат ранжування називається ранжировкою.

Ранг r_{ij} задає порядкове місце, що займає значення змінної x_{ij} у ряді розподілу.

5.2.2 Поняття рангової кореляції

Під *ранговою кореляцією* розуміють статистичний зв'язок між порядковими змінними. В статистичній практиці цей зв'язок аналізується на основі вхідних даних, що мають вигляд впорядкованостей (ранжировок) n , досліджуваних об'єктів за різними властивостями.

5.2.3 Ранговий коефіцієнт кореляції Спірмена

Близькість двох рядів чисел x_1, \dots, x_n та y_1, \dots, y_n характеризує величина (5.1):

$$\rho = \sum_{i=1}^n (x_i - y_i)^2. \quad (5.1)$$

Вона приймає найменше можливе значення $S = 0$ тоді і тільки тоді, коли послідовності повністю співпадають. Найбільше можливе значення $\rho = \frac{1}{3}(n^3 - n)$ величина S приймає, коли ці послідовності повністю протилежні (це означає, що для $x_i = 1$ значення $y_i = n$; для $x_i = 2$, які відповідають $y_i = n - 1$ і т. д.). Крім ступеня подібності послідовностей (x_1, \dots, x_n) та (y_1, \dots, y_n) , на ρ впливають також і численність групи n . Щоб послабити вплив змінної n , переходять до *коефіцієнту рангової кореляції Спірмена*:

$$r_s = 1 - \frac{6\rho}{n^3 - n}. \quad (5.2)$$

Коефіцієнт ρ за абсолютною величиною обмежений одиницею: $|r_s| \leq 1$. Свої граничні значення $r_s = \pm 1$ він приймає у вказаних вище випадках повної передбачуваності однієї рангової послідовності за другою.

Зазначимо, що значення ρ не залежить від початкової нумерації об'єктів. В якості такої часто буває зручним вибрати упорядкування за однією з ознак. Тоді послідовність рангів за цією ознакою перетвориться у послідовність 1, 2, ..., n . Другу послідовність позначимо, наприклад, z_1, \dots, z_n . При цьому:

$$\rho = \sum_{i=1}^n (x_i - y_i)^2 = \sum_{k=1}^n (k - z_k)^2. \quad (5.3)$$

Можна побачити, що коефіцієнт Спірмена є повним аналогом коефіцієнта парної кореляції.

Як приклад, оцінімо близькість рядів ($n = 10$) поданих у табл. 5.1 за допомогою коефіцієнту Спірмена. Для цього необхідно спочатку розрахувати величину $\rho = \sum_{i=1}^n (x_i - y_i)^2$. Далі за формулою

$$r_s = 1 - \frac{6\rho}{n^3 - n} \text{ виконати обчислення коефіцієнту Спірмена.}$$

Таблиця 5.1 – Розрахунок коефіцієнту Спірмена

X	Y	ρ_i
1,843201	1,721118	0,014904
2,864538	2,764027	0,010103
3,284455	3,164398	0,014414
4,079488	4,785251	0,498102
5,409823	5,759757	0,122454
6,826413	6,137924	0,474017
7,440344	7,828727	0,150842
8,440813	8,033664	0,16577
9,583976	9,052013	0,282984
10,76483	10,0566	0,50159
Σ		2,23518
r_s		0,986453

Згідно розрахункам, коефіцієнт Спірмена дорівнює $r_s = 0,986453$, що свідчить про високу близькість двох рядів.

5.2.4 Ранговий коефіцієнт кореляції Кендела

Другий коефіцієнт рангової кореляції отримав популярність після робіт М.Кендела. Цей коефіцієнт в якості міри подібності між двома ранжировками використовує мінімальну кількість перестановок сусідніх об'єктів, які треба зробити, щоб одне упорядкування об'єктів перетворити в інше.

В тому випадку, коли необхідно порівняти не дві змінні, а велику кількість, наприклад, при з'ясуванні узгодженості поглядів групи експертів, використовується *коефіцієнт конкордації*, запропонований Кенделом:

$$W = \frac{12}{m^3(n^3 - n)} \sum_{j=1}^n \left(\sum_{i=1}^m \left(R_{ij} - \frac{n+1}{2} \right) \right)^2, \quad (5.4)$$

де n – кількість аналізуємих об'єктів; m – кількість експертів;
 R_{ij} – ранг j -го об'єкта, присвоєний йому i -м експертом.

5.2.5 Ранговий однофакторний аналіз

Якщо ми нічого не знаємо про розподіл змінних, ми можемо використати ранговий однофакторний аналіз, що є непараметричним аналогом дисперсійного аналізу. Для проведення статистичного аналізу, перш за все, для зручності треба всі експериментальні дані звести до однієї таблиці (табл. 5.2).

Таблиця 5.2 – Представлення статистичних даних

Рівні фактора	1	2	...	k
Результати вимірів	x_{11}	x_{12}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2k}

	x_{n11}	x_{n22}	...	x_{nkk}

Тут n_1, \dots, n_k – об'єми вибірок; $N = n_1 + n_2 + \dots + n_k$ – загальна кількість спостережень.

При цьому вся інформація, що ми використовуємо в таблиці 5.2, містяться у тих *рангах*, що отримують числа x_{ij} при упорядкуванні всієї сукупності.

Впорядкуємо величини x_{ij} (немає різниці як – від більшої до меншої, чи від меншої до більшої). Позначимо через r_{ij} ранг числа x_{ij} у всій сукупності. Тоді табл. 5.2 перетворюється у табл. 5.3. Важливо відзначити, що при виконанні гіпотези H_0 будь-які можливі розташування рангів по місцям у табл. 5.3 рівно ймовірні.

Таблиця 5.3 – Представлення статистичних даних

Рівні фактора	1	2	...	k
Результати вимірювань	r_{11}	r_{12}	...	r_{1k}
	r_{21}	r_{22}	...	r_{2k}

	r_{n11}	r_{n22}	...	r_{nkk}

Згідно сформульованої стратегії аналізу виникає питання: чи не можна пояснити спостережене в досліді розміщення рангів у таблиці 5.3 дією чистої ймовірності? Це питання можна переформулювати у вигляді статистичної гіпотези H_0 про те, що всі k наведених вибірок (стовпці таблиці 5.3) однорідні, тобто є вибірками з одного й того ж розподілу. Наша задача – вказати статистичний критерій, за допомогою якого можна було б судити про справедливість висунутої гіпотези.

5.2.6 Критерій Краскала-Уолліса

Для кожної обробки j (тобто для кожного стовпця початкової таблиці) треба розрахувати

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad (5.5)$$

та

$$R_j = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}, \quad (5.6)$$

де $R_{.j}$ – середній ранг, розрахований по стовпцю. Якщо між стовпцями нема систематичних різниць, середні ранги $R_{.j}$, $j=1, \dots, k$ не повинні значно відрізнятися від середнього рангу, що розрахований за всією сукупністю $\|r_{ij}\|$. Зрозуміло, що останній дорівнює $(N+1)/2$. Тому величини

$$\left(R_1 - \frac{N+1}{2}\right)^2, \dots, \left(R_k - \frac{N+1}{2}\right)^2 \quad (5.7)$$

при H_0 у сукупності повинні бути невеликими. Складаючи загальну характеристику, розумно врахувати різницю у кількості спостережень для різних обробок та взяти в якості міри відхилення від чистої ймовірності величину:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(R_j - \frac{N+1}{2}\right)^2. \quad (5.8)$$

Ця величина називається *статистикою Краскала-Уолліса*. Множник $12/(N(N+1))$ присутній у її виразі як нормуючий, для забезпечення асимптотичної збіжності розподілу H до розподілу χ^2 -квадрат з числом ступенів волі $(k-1)$. Друга форма для обчислення H :

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1). \quad (5.9)$$

Отже нульова гіпотеза відкидається (на рівні значущості α), якщо $H_{набл} > \chi^2_{1-\alpha}$, де $\chi^2_{1-\alpha}$ – квантіль рівня $(1-\alpha)$ розподілу χ^2 -квадрат з $(k-1)$ ступенями волі. Таблиці розподілу статистики H при гіпотезі H_0 можна знайти у збірниках статистичних таблиць.

Як було зазначено вище, при присутності збігів (та використанні середніх рангів) теоретична схема діє як наближення, а надійність її висновків знижується тим більше, чим більше збігів, тому при великій кількості збігів, рекомендується використовувати модифіковану формулу статистики H' :

$$H' = \frac{H}{\left(1 - \sum_{j=1}^g T_j / [N^3 - N]\right)}, \quad (5.10)$$

де g – кількість груп співпадаючих спостережень, $T_j = (t_j^3 - t_j)$;
 t_j – кількість співпадаючих спостережень у групі з номером j .

5.2.7 Практичний приклад

Приклад 1. Проілюструймо застосовування наведеного вище критерію на наступному прикладі. Для з'ясування впливу грошового стимулювання на продуктивність праці шістьом однорідним групам з п'яти чоловік, кожній було запропоновано задачі однакової складності. Задачі пропонувалися кожному випробуваному незалежно від всіх інших. Групи відрізняються між собою величиною грошової винагороди за розв'язувану задачу. У табл. 5.4 наведено кількість вирішених задач членами кожної групи.

Таблиця 5.4 – Величина винагороди (від меншої до більшої)

Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Перевіримо гіпотезу про відсутність впливу грошової винагороди на кількість вирішених задач. Оскільки відгук вимірюється у порядковій шкалі будемо використовувати критерій Краскела-Уолліса.

У зв'язку з присутністю у таблиці 5.4 збігів ми повинні скористатися середніми рангами. Спочатку перепишемо всі результати вимірювань у зростаючому порядку: 7, 8, 9, 9, 10, 10, 11, 12, 12, 12, 13, 13, 14, 15, 16, 16, 16, 16, 17, 18, 18, 19, 19, 20, 22, 24, 24, 25, 27. Так, значення $x_{ij}=10$ зустрічається у таблиці 5.3 два рази, та при упорядкуванні “займає п'яте та шосте місце”, отже середній ранг $x_{ij}=10$ дорівнює 5.5. В результаті ранжування отримаємо таблицю 5.5. В двох ни-

жніх строках наведено суми рангів R_j та середні ранги $R_{.j} = R_j / n_j$ за стовпцями.

Таблиця 5.5 – Таблиця рангів спостережень

	Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
	5,5	2	9	9	27,5	23,5
	7	5,5	20	14	17	21,5
	3,5	17	13	17	26	30
	11,5	11,5	3,5	17	21,5	29
	1	9	17	23,5	25	27,5
$R_j =$	27,5	45	62,5	80,5	117	131,5
$R_{.j} =$	5,7	9	12,5	16,1	23,4	26,3

Для розрахунку статистики Краскела-Уолліса H зручніше використати формулу (5.9). Де $N = 30$ – кількість спостережень при заданому значенні фактору $n_j = 5$, $j=1,...,6$. Підставивши ці значення, отримаємо: $H = 17682 / 155 - 93 = 21,077$.

Із статистичних таблиць розподілу χ^2 знаходимо, що мінімальний рівень значущості α є трохи більшим за 0,001. Але, ці висновки є приблизними у зв'язку з тим, що в таблиці 5.5 була деяка кількість співпадаючих значень спостережень x_{ij} . Для врахування впливу зв'язків можна використати статистику H' (5.10). В нашому випадку маємо наступні вісім груп співпадаючих спостережень: 9, 9, 10, 10, 12, 12, 12, 13, 13, 16, 16, 16, 16, 16, 18, 18, 19, 19, 24, 24.

Таким чином: $T_1 = (2^3 - 2) = 6$, $T_2 = (2^3 - 2) = 6$, $T_3 = (3^3 - 3) = 24$, $T_4 = 6$, $T_5 = (5^3 - 5) = 120$, $T_6 = 6$, $T_7 = 6$, $T_8 = 6$. Знаменник дробу в (6.15) дорівнює: $1 - \sum_{j=1}^8 T_j / (30^3 - 30) = 1 - 6/899$, а саме значення

H' приблизно дорівнює 21,2186.

Оскільки скореговане значення H' статистики Краскела-Уолліса незначно відрізняється від значення H , ми можемо відхилити гіпотезу на мініальному рівні значущості біля 0,001.

Приклад 2. Проведемо розрахунок критерію Краскела-Уолліса для з'ясування впливу грошової стимуляції на кількість вирішених задач серед восьми груп по чотири чоловіка. Вхідні дані наведені у табл.5.6.

Таблиця 5.6 – Вхідні дані ($n = 4, N = 32$)

Групи							
№1	№2	№3	№4	№5	№6	№7	№8
0	2	4	4	7	7	5	8
2	2	2	3	5	6	7	7
1	2	4	3	4	7	7	8
0	3	1	2	6	6	7	9

Отже для розрахунку критерію нам необхідно привести табл.5.6 до виду рангової таблиці. Для цього необхідно відсортувати вимірювання від меншого значення до більшого та присвоїти кожному свій ранг (номер по порядку). Якщо декілька вимірювань повторюються, треба скласти ранги цих вимірювань та поділити на кількість повторів. Це й буде ранг кожного з повторюваних вимірювань (табл.5.7).

Відсортовані вимірювання: 0, 0, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 9.

Наприклад, для значення 0 ранг дорівнює сумі порядкових номерів цього значення $(1+2)$ поділену на кількість однакових значень (2), тобто ранг дорівнює $(1+2)/2 = 1,5$. Тим же чином розрахуємо наступні ранги (табл. 5.7).

$$\text{Ранг значення 1} = (3+4)/2 = 3,5$$

$$\text{Ранг значення 2} = (5+6+7+8+9+10)/6 = 7,5$$

$$\text{Ранг значення 3} = (11+12+13)/3 = 12$$

$$\text{Ранг значення 4} = (14+15+16+17)/4 = 15,5$$

$$\text{Ранг значення 5} = (18+19)/2 = 18,5$$

$$\text{Ранг значення 6} = (20+21+22)/3 = 21$$

$$\text{Ранг значення 7} = (23+24+25+26+27+28+29)/7 = 26$$

$$\text{Ранг значення 8} = (30+31)/2 = 30,5$$

$$\text{Ранг значення 9} = 32$$

Таблиця 5.7 – Присвоєння рангів вимірюванням

Значення	Кількість	Ранг
0	2	1,5
1	2	3,5
2	6	7,5
3	3	12
4	4	15,5
5	2	18,5
6	3	21
7	7	26
8	2	30,5
9	1	32

Отже тепер ми можемо побудувати рангову таблицю (табл.5.8).

Таблиця 5.8 – Рангова таблиця

Ранги	Групи							
	№1	№2	№3	№4	№5	№6	№7	№8
	1,5	7,5	15,5	15,5	26	26	18,5	30,5
	7,5	7,5	7,5	12	18,5	21	26	26
	3,5	7,5	15,5	12	15,5	26	26	30,5
	1,5	12	3,5	7,5	21	21	26	32
R_j	14	34,5	42	47	81	94	96,5	119
$R_{.j}$	3,5	8,625	10,5	11,75	20,25	23,5	24,125	29,75

Спочатку треба розрахувати $R_j = \sum_{i=1}^{n_j} r_{ij}$ та $R_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}$.

Результати наведені у табл.5.8. Далі розрахуємо критерій Краскела-Уолліса за формулою:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(R_{.j} - \frac{N+1}{2} \right)^2$$

Для цього спершу знайдемо суму (табл. 5.9).

Таблиця 5.9 – Сума у формулі Краскела-Уолліса (n_j)

№	$n_j \left(R_{.j} - \frac{N+1}{2} \right)^2$
1	676
2	248,06
3	144
4	90,25
5	56,25
6	196
7	232,56
8	702,25
Σ	2345,375

Підставивши всі необхідні значення отримаємо $H=26,66$.

Але при присутності досить великої кількості збігів надійність цього критерію знижується, тому рекомендується використовувати модифіковану формулу статистики H :

$$H' = \frac{H}{\left(1 - \frac{\sum_{j=1}^g T_j}{N^3 - N} \right)}$$

де g – кількість груп співпадаючих спостережень, $T_j = (t_j^3 - t_j)$,
 t_j – кількість співпадаючих спостережень у групі з номером j .

Отже в даному випадку $g=9$, T для груп =6, 6, 210, 24, 60, 6, 24, 336, 6.

Звідси сума по $T = 678$.

Підставивши значення отримаємо

$$H' = \frac{26.65199}{\left(1 - \frac{678}{32^3 - 32} \right)} = 27,21566.$$

5.3 Порядок виконання роботи

Проілюструємо виконання рангового аналізу Краскела-Уолліса з використанням мови програмування R та середовища розробки R-Studio.

В роботі необхідно використати бібліотеки R з функціями, які наведені у таблиці 5.10.

Таблиця 5.10 – Перелік стандартних функцій та команд R

Назва бібліотеки	Опис
<i>tidyverse</i>	для маніпулювання даними та візуалізації
<i>ggpubr</i>	для створення простих та легких звітів аналізу
<i>rstatix</i>	надає дружні функції-труби для простого статистичного аналізу.

Функції (або оператори) труби обробляють об'єкт даних, використовуючи послідовність операцій, передаючи результат одного кроку в якості введення для наступного кроку з використанням інфікс-операторів, а не більш типовий R-метод вкладених виклики функцій. Зверніть увагу, що метою операторів труб є підвищення читабельності сирцевого коду.

Будемо використовувати вбудований набір даних R під назвою PlantGrowth. Він містить масу досліджуваних рослин і дві різні умови обробки.

```
set.seed(1234)
```

```
PlantGrowth %>% sample_n_by(group, size = 1)
```

```
# A tibble: 3 x 2
  weight group
  <dbl> <fct>
1   5.14 ctr1
2   3.83 trt1
3   5.37 trt2
```

Рисунок 5.1 – Вибірка даних PlantGrowth

Перевпорядкуємо рівні вибірки:

```
PlantGrowth <- PlantGrowth %>% reorder_levels(group, order =
c("ctrl", "trt1", "trt2"))
```

Обчислено зведені статистики по групах:

```
PlantGrowth %>% group_by(group) %>%
get_summary_stats(weight, type = "common")
```

```
# A tibble: 3 x 11
  group variable    n  min  max median  iqr  mean   sd   se    ci
<fct> <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 ctrl  weight     10  4.17  6.11  5.16  0.743  5.03  0.583  0.184  0.417
2 trt1  weight     10  3.59  6.03  4.55  0.662  4.66  0.794  0.251  0.568
3 trt2  weight     10  4.92  6.31  5.44  0.467  5.53  0.443  0.14  0.317
```

Рисунок 5.2 – Зведені статистики за групами

Створимо діаграму розмаху «weight» за «group»:

```
ggboxplot(PlantGrowth, x = "group", y = "weight")
```

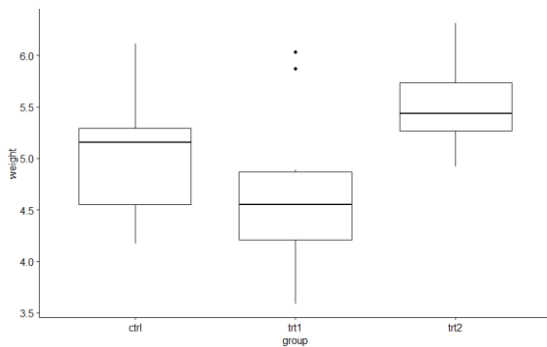


Рисунок 5.3 – Діаграма розмаху «weight» за «group»

Розглянемо чи існує будь-яка істотна різниця між середніми вагами рослин в трьох експериментальних умовах.

Будемо використовувати дружню до функцію-трубу `kruskal_test()` – функцію з бібліотеки `rstatix` – оболонку базової функції R `kruskal.test()`.

```
res.kruskal <- PlantGrowth %>% kruskal_test(weight ~ group)
```

```
res.kruskal
```

```
# A tibble: 1 x 6
  .y.      n statistic    df    p method
* <chr> <int>    <dbl> <int> <dbl> <chr>
1 weight    30     7.99     2 0.0184 kruskal-wallis
```

Рисунок 5.4 – Використання `kruskal_test()` для визначення різниці середніх ваг рослин в різних експериментальних умовах

Квадрат η^2 – швидкості навчання моделі, заснований на статистиці H , може бути використаний як міра розміру тестового ефекту Краскела-Уолліса. Він розраховується наступним чином: $\eta^2[H] = (H - k + 1) / (n - k)$; де H – значення, отримане в тесті Краскела-Уолліса; k – число груп; n – загальне число спостережень.

Ця квадратна оцінка приймає значення від 0 до 1 та помножається на 100, що вказує на відсоток дисперсії в залежній змінній, визначеної незалежною змінною.

Значення інтерпретації становлять: від 0,01 до 0,06 (малий ефект), від 0,06 до 0,14 (помірний ефект) та більше за 0,14 (великий ефект).

```
PlantGrowth %>% kruskal_effsize(weight ~ group)
```

```
# A tibble: 1 x 5
  .y.      n effsize method magnitude
* <chr> <int>    <dbl> <chr>    <ord>
1 weight    30  0.222 eta2[H] large
```

Рисунок 5.5 – Значення тестового ефекту Краскела-Уолліса (виявлено великий розмір ефекту $\eta^2[H] = 0,22$.)

З результатів тесту Краскела-Уолліса ми виявили, що існує значна різниця між групами, але ми не знаємо, які пари груп відрізняються.

За значущим тестом Крускала-Уолліса зазвичай слід тест Данна (Dunn's test), щоб визначити, які групи відрізняються. Також можна використовувати критерій Уїлкоксона (Wilcoxon signed-rank test) для обчислення попарних порівнянь між рівнями групи з поправкою на багаторазове тестування.

```
# Pairwise comparisons
```

```
pwc <- PlantGrowth %>% dunn_test(weight ~ group,
p.adjust.method = "bonferroni")
pwc
```

```
# A tibble: 3 x 9
  .y. group1 group2 n1 n2 statistic p p.adj p.adj.signif
* <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <chr>
1 weight ctrl trt1 10 10 -1.12 0.264 0.791 ns
2 weight ctrl trt2 10 10 1.69 0.0912 0.273 ns
3 weight trt1 trt2 10 10 2.81 0.00500 0.0150 *
```

Рисунок 5.6 – Попарні порівняння з використанням тесту Данна

```
pwc2 <- PlantGrowth %>% wilcox_test(weight ~ group,
p.adjust.method = "bonferroni")
pwc2
```

```
# A tibble: 3 x 9
  .y. group1 group2 n1 n2 statistic p p.adj p.adj.signif
* <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <chr>
1 weight ctrl trt1 10 10 67.5 0.199 0.597 ns
2 weight ctrl trt2 10 10 25 0.063 0.189 ns
3 weight trt1 trt2 10 10 16 0.009 0.027 *
```

Рисунок 5.7 – Попарні порівняння з використанням критерію Уїлкоксона

Попарне порівняння показує, що тільки trt1 та trt2 істотно розрізняються (тест Уїлкоксона, $P = 0,027$).

Статистично значущі відмінності між групами лікування були оцінені за допомогою критерію Краскела-Уолліса ($P = 0,018$). Попарний тест Уїлкоксона між групами показав, що достовірною була тільки різниця між групами trt1 та trt2 (тест Уїлкоксона, $P = 0,027$)

```
# Побудова діаграми розмаху для p-values
pwc <- pwc %>% add_xy_position(x = "group")
ggboxplot(PlantGrowth, x = "group", y = "weight") +
  stat_pvalue_manual(pwc, hide.ns = TRUE) +
  labs(
    subtitle = get_test_label(res.kruskal, detailed = TRUE),
    caption = get_pwc_label(pwc)
  )
```

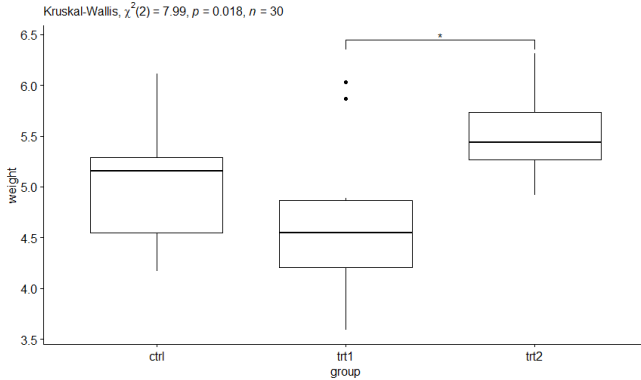


Рисунок 5.8 – Діаграма розмаху для рівней значущості

5.4 Завдання на лабораторну роботу

5.4.1 Для даних з лабораторної роботи (вибірка `jobsatisfaction`) перевірити, чи існує вплив фактора на відгук за допомогою критерію Краскела-Уолліса.

Бібліотека	Вибірка	Приклад функцій
datarium	jobsatisfaction	<code>jobsatisfaction %>% sample_n_by(education_level, size = 1)</code>
		<code>jobsatisfaction %>% group_by(education_level) %>% get_summary_stats(score, type = "common")</code>
		<code>ggboxplot(jobsatisfaction, x = "education_level", y = "score")</code>
		<code>res.krskal <- jobsatisfaction %>% kruskal_test(score ~ education_level)</code>

5.4.2 Проаналізувати отримані результати. Які виникли труднощі?

5.4.3 Оформити звіт.

5.5 Зміст звіту

- 5.5.1 Назва і мета роботи.
- 5.5.2 Постанова завдання.
- 5.5.3 Таблиця вихідних даних.
- 5.5.4 Основні формули для розрахунку критерію.
- 5.5.5 Результати аналізів.
- 5.5.6 Висновки за результатами аналізу.

5.6 Контрольні запитання

- 5.6.1 Коли доречним є використання непараметричних методів аналізу?
- 5.6.2 Постанова завдання рангового однофакторного аналізу Краскела-Уолліса.
- 5.6.3 Накреслити форму таблиці вихідних статистичних даних.
- 5.6.4 Накреслити для отриманих у викладача вихідних даних таблицю рангів спостережень.
- 5.6.5 Написати формулу статистики Краскела-Уолліса для випадку, коли у вихідних даних є багато однакових значень.
- 5.6.6 Написати формулу статистики Краскела-Уолліса для випадку, коли у вихідних даних немає однакових значень.
- 5.6.7 Яким чином робиться висновок про справедливість нульової гіпотези?
- 5.6.8 Як розраховується ранговий коефіцієнт кореляції Спірмена?
- 5.6.9 Написати формулу для розрахунку коефіцієнта кореляції Спірмена.
- 5.6.10 Коли доречним є використання коефіцієнта Кендалла?
- 5.6.11 Написати формулу для розрахунку коефіцієнта конкоддинації.

6 ЛАБОРАТОРНА РОБОТА №6 ПОВНИЙ ФАКТОРНИЙ ЕКСПЕРИМЕНТ

6.1 Мета роботи

Вивчити апарат математичного моделювання методом планування експерименту. Виконати статистичний аналіз рівнянь регресії, використовуючи статистичні пакети прикладних програм, побудувати математичну модель за експериментальними даними, отриманими при проведенні повного факторного експерименту типу 2^k .

6.2 Короткі теоретичні відомості

6.2.1 Постановка завдання

Планування експерименту – це вибір числа і умов проведення дослідів, необхідних і достатніх для розв’язання поставлених задач з необхідною точністю. При цьому важливим є:

- а) мінімізіція загальної кількості дослідів;
- б) одночасне варіювання всіма змінними, що визначають процес, за спеціальними правилами;
- в) використання математичного апарату, що формалізує багато дій дослідника.

Одним з таких методів є планування факторного експерименту. Сутність методу полягає у тому, що ще до проведення експерименту обирають вид математичної моделі, і вже для неї складають план експерименту таким чином, щоб він відповідав певним умовам. Відповідність плану цим умовам дозволяє обчислити коефіцієнти моделі за так званими ортогональними поліномами, що надає моделі найбільшу точність при мінімально можливому числі дослідів.

Експеримент називається *факторним*, тому що незалежні величини x_1, x_2, \dots, x_n , які цілеспрямовано змінюють в ході експеримента, прийнято називати *факторами впливу* (змінні на вході). Саму функцію $y = f(x)$ називають *відгуком* (параметр оптимізації).

Часто об’єкт дослідження розглядають у вигляді умовної схеми “чорного ящика” (рис. 6.1).

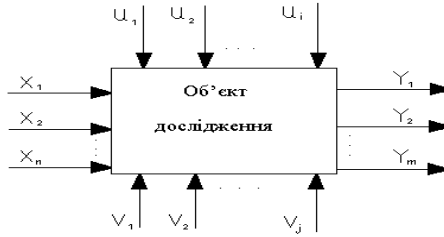


Рисунок 6.1 – Умовна схема «чорного ящика»

де y_1, y_2, \dots, y_m – відгуки, які залежать від параметрів впливу трьох типів, що позначені X, U, V ;

x_1, x_2, \dots, x_n – це спостережувані і керовані в процесі експерименту незалежні між собою змінні, що називаються факторами;

u_1, u_2, \dots, u_i – це спостережувані, але не керовані параметри;

v_1, v_2, \dots, v_j – це не спостережувані і не керовані параметри.

У загальному випадку відгук Y залежить від всіх груп змінних. Задача експериментатора полягає у тому, щоб знайти залежність відгука Y від факторів X в умовах впливу інших факторів.

Математична теорія планування експерименту – це наука про засоби складання економних експериментальних планів, які одночасно дозволяють здобути найбільшу кількість інформації про об'єкт, про засоби проведення експерименту, обробки експериментальних даних, використання отриманих результатів для оптимізації виробничих процесів. Існують два основні засоби збирання початкового статистичного матеріалу для наступного отримання математичної моделі: пасивний та активний експеримент.

Активний експеримент – це такий експеримент, матриця умов планування якого наперед досліджена. Матриця умов проведення експерименту називається матрицею планування і складається згідно з вимогами теорії планування експерименту (ТПЕ).

Пасивний експеримент – це експеримент, для якого умови проведення наперед не встановлюються, тобто експеримент виконується в режимі звичайної роботи об'єкту, що досліджується. Пасивний експеримент за якістю початкового матеріалу поступається активному.

Результат його дуже важко обробляти і якість отриманої моделі завжди є не дуже великою.

В даній лабораторній роботі розглядається активний експеримент, за результатами обробки початкових даних якого може бути отримана математична модель процесу :

$$y = a_0 + \sum_{i=1}^n a_i x_i + \sum_{i=1}^n a_{ii} x_i^2 + \sum_{\substack{i,l=1 \\ i \neq l}}^n a_{i,l} x_i x_l + \varepsilon, \quad (6.1)$$

де y – значення параметру оптимізації, який прогнозується;
 a_0, a_1, \dots, a_n – невідомі оцінки коефіцієнтів, які потрібно знайти; x_i – змінні на вході об'єкту; ε – помилка одиничного прогнозування.

У наступних розділах описані основні етапи дослідження при реалізації активного експерименту в порядку їх виконання.

6.2.2 Вибір параметру оптимізації

Залежні змінні (відгуки) мають відповідати наступним вимогам:

- а) мати фізичний смисл і достатньо повно характеризувати досліджуваний об'єкт, процес чи явище;
- б) бути відтворюваними, тобто при повторенні дослідів у номінально однакових умовах отримані значення повинні співпадати з точністю до помилки експерименту;
- в) кожному набору значень незалежних змінних повинно відповідати одне (з точністю до випадкової помилки) значення відгуку.

6.2.3 Вибір факторів

Незалежні змінні (фактори) повинні відповідати наступним основним вимогам:

- а) фактори мають бути керованими: можливість встановлювати та підтримувати необхідні значення у процесі експерименту;
- б) фактори не повинні залежати від інших змінних (можливість незалежно від решти факторів керувати кожною змінною);
- в) фактори мають бути детермінованими величинами, однозначними та відповідати вимогам сумісності.

6.2.4 Вибір координат базової точки

Координата базової точки $(x_{1,0}; x_{2,0}; \dots; x_{n,0})$, яку називають також *центром експерименту*, обирають у точці звичайного, номінального ведення процесу.

6.2.5 Вибір ступенів варіювання

Ступенем варіювання факторів називається відстань на координатній вісі між основним (базовим) і верхнім (або нижнім) рівнем.

Вибір ступенів варіювання λ_i має відповідати наступним вимогам.

Якщо на вхідні і вихідні змінні накладаються обмеження, то при варіюванні вхідних змінних не треба виходити за встановлені межі:

$$\left. \begin{aligned} x_{i,0} + \lambda_i &\leq x_{i\max} \\ x_{i,0} - \lambda_i &\geq x_{i\min} \end{aligned} \right\}, \quad (6.2)$$

де $x_{i,0}$ – базове, початкове значення фактора ; λ_i – ступінь варіювання; $x_{i\max}, x_{i\min}$ – відповідно найвища і найнижча межі.

Ступінь варіювання λ_i повинна суттєво перевищувати похибку вимірювання по x_i .

6.2.6 Отримання матриці планування. Визначення координат точок для пробних експериментів

Для отримання ортогональної матриці планування (МП) усі фактори мають бути пронормовані за формулою (6.3):

$$X_i = \frac{x_i - x_{i,0}}{\lambda_i}, \quad (6.3)$$

де X_i – нормоване значення фактора; x_i – натуральне значення фактору.

Тоді нормовані значення факторів для базового $X_{i,0}$, нижнього $X_{i,n}$ та верхнього $X_{i,e}$ рівнів варіювання дорівнюють:

$$X_{i,0} = \frac{x_{i,0} - x_{i,0}}{\lambda_i} = 0; \quad (6.4)$$

$$X_{i,n} = \frac{x_{i,n} - x_{i,0}}{\lambda_i} = \frac{x_{i,0} - \lambda_i - x_{i,0}}{\lambda_i} = -1; \quad (6.5)$$

$$X_{i,e} = \frac{x_{i,e} - x_{i,0}}{\lambda_i} = \frac{x_{i,0} + \lambda_i - x_{i,0}}{\lambda_i} = +1, \quad (6.6)$$

$$\text{де} \quad \left. \begin{array}{l} x_{i,e} = x_{i,0} + \lambda_i \\ x_{i,n} = x_{i,0} - \lambda_i \end{array} \right\}.$$

Матриця планування включає в себе тільки вектор-стовпці для незалежних (лінійних) нормованих факторів. Вектор-стовпець x_0 потрібен для розрахунку вільного члена a_0 у рівнянні регресії.

В МП для $n=2$ (табл. 6.1) наведені всі можливі комбінації рівнів (їх $2^2 = 4$). У загальному випадку, коли кількість факторів дорівнює n , а рівнів варіювання k , кількість усіх можливих комбінацій визначається формулою:

$$N = k^n. \quad (6.7)$$

Таблиця 6.1 – Матриця планування для експерименту 2^2

№ досліду	X_0	X_1	X_2	y
1	+1	-1	-1	y_1
2	+1	+1	-1	y_2
3	+1	-1	+1	y_3
4	+1	+1	+1	y_4

При дворівневому плануванні ($k = 2$), очевидно $N = 2^n$.

Якщо реалізуються усі 2^n комбінацій рівнів варіювання факторів, то такий експеримент називають *повним факторним експериментом* (ПФЕ), а МП, в якій передбачені всі 2^n комбінацій рівнів, – *повною матрицею планування* (ММП). Після заміни x_i на нормовані X_i за формулою (6.3) рівняння регресії (6.1) приймає вигляд:

$$y = a_0^* + \sum_{i=1}^n a_i^* X_i + \sum_{\substack{i,l=1 \\ i \neq l}}^n a_{i,l}^* X_i X_l + \varepsilon, \quad (6.8)$$

де $a_0^*, a_i^*, a_{i,l}^*$ – оцінки коефіцієнтів нормованого рівняння регресії, які відрізняються від оцінок коефіцієнтів a_i рівняння у звичайному масштабі (6.1). У рівнянні (6.8) на відміну від рівняння (6.1) відсутні квадратичні члени. Це пояснюється тим, що квадратичні коефіцієнти при плануванні на двох рівнях оцінити окремо неможливо, тому що вони змішані з вільним членом.

Через те що a_0^* – величина постійна і не залежна від інших факторів, вводять фіктивну змінну X_0 , якій у всіх рядках вектор-стовпця X_0 приписують рівні $X_0 = +1$. Це дозволяє не забувати обчислювати вільний член a_0^* .

МП для ПФЕ має такі властивості :

– *симетричність*. Алгебраїчна сума усіх факторів у стовпці дорівнює нулю, тобто усі пробні експериментальні точки при ПФЕ розташовані симетрично відносно базової точки:

$$\sum_{k=1}^N x_{i,k} = 0 \quad (i = \overline{1, n}); \quad (6.9)$$

– *нормування*. Сума квадратів кожного стовпця дорівнює кількості дослідів, тобто у МП передбачені або нижні рівні -1 , або верхні рівні $+1$:

$$\sum_{k=1}^N x_{i,k}^2 = N \quad (i = \overline{1, n}); \quad (6.10)$$

– *ортогональність*. Скалярний добуток будь-яких двох вектор-стовпців (для факторів) дорівнює нулю:

$$\sum_{k=1}^N x_{i,k} x_{j,k} = 0 \quad (i, j = \overline{1, n-1}). \quad (6.11)$$

6.2.7 Проведення експериментів у запланованих точках. Рандомізація дослідів

Необхідно забезпечити випадковість помилок, які накладаються на вхідні та вихідні змінні. Однак, якщо проводити досліди у тому порядку, в якому йдуть рядки МП, то обов'язково виникнуть систематичні похибки. Це викликано тим, що установка заданих МП нижніх $x_{i,n}$ чи верхніх $x_{i,1}$ рівнів варіювання не може проводитись абсолютно точно, без помилок, і якщо одна і та ж установка, з тією ж помилкою у визначенні рівня не змінюється підряд у декількох дослідях (тобто в декількох рядках МП підряд), то виникає систематична помилка. Щоб залишити тільки випадкову помилку в установці рівнів факторів з нульовим математичним очікуванням, всі досліди виконують у випадковому порядку.

Проведення експерименту повинно строго відповідати обраному випадковому порядку, причому установка рівнів факторів x_i повинна бути як можливо більш точною.

6.2.8 Обчислення коефіцієнтів рівняння регресії

Завдячуючи властивостям ортогональності, нормування та симетричності МП формули для обчислення коефіцієнтів рівняння регресії виявляються доволі простими:

$$a_i^* = \frac{1}{N} \sum_{k=1}^N X_{i,k} y_k, \quad (i = \overline{0, n-1}), \quad (6.12)$$

де $X_{i,k}$ – дорівнює $+1$ чи -1 у залежності від номеру k рядку МП; y_k – значення цільової функції, отримане при значеннях факторів X_i , які містяться у k -тому рядку МП.

Коефіцієнти обчислюють за формулою (6.12), причому кількість незалежних оцінок має відповідати умові ($i = \overline{0, n-1}$), іншими словами кількість коефіцієнтів a_i^* , оцінених незалежно один від одного, не повинна перевищувати кількість N рядків МП. Та чи будь-які із N коефіцієнтів можна оцінити? Виявляється, не будь-які, а лише N коефіцієнтів тих факторів X_i чи їх взаємодій, для яких комбінації знаків у вектор-стовпцях не повторюються.

Таким чином, при плануванні типу $N = 2^n$ отримати незалежну, незмішану оцінку можна тільки для a_0^* та й то лише у випадку, коли ділянка поверхні відгуку в районі базової точки відрізняється відносно невеликою кривизною – у цьому випадку квадратичні члени та члени більш високих порядків виявляються незначними. Як можна бачити, при організації планування $N = 2^n$ виходять із гіпотези про невелику кривизну поверхні відгуку в районі базової точки, а у справедливості цієї гіпотези переконуються тільки в результаті перевірки адекватності отриманого рівняння регресії (6.8).

6.2.9 Статистична оцінка значущості коефіцієнтів рівняння регресії

Статистичну оцінку значущості коефіцієнтів a_i^* ; $a_{i,l}^*$ проводять за t -критерієм Стьюдента.

Для кожного з коефіцієнтів обчислюють t -відношення (6.13):

$$t_i = \frac{|a_i^* - a_i|}{S\{a_i^*\}} = \frac{|a_i^*|}{S\{a_i^*\}}, \quad (6.13)$$

де a_i – теоретичний генеральний коефіцієнт, який приймають рівним нулю (основна гіпотеза); $S\{a_i^*\}$ – оцінка середньоквадратичної помилки у визначенні коефіцієнтів a_i^* , яку знаходять через оцінку дисперсії коефіцієнтів $S^2\{a_i^*\}$.

Розраховане значення t_i (6.13) порівнюють з критичним (табличним) $t_{кр} = t_{табл}$, яке обирають для прийнятого рівня значущості q і кількості ступенів волі

$$f_g = N(\gamma - 1), \quad (6.14)$$

де γ – кількість паралельних дослідів.

Якщо виконується умова

$$t_i > t_{кр} = t_{табл}, \quad (6.15)$$

то основна гіпотеза ($a_i = 0$) відкидається, і коефіцієнт a_i^* визнається статистично значущим. У протилежному випадку основна гіпотеза приймається, a_i^* визнається статистично незначущим та з рівняння регресії виключається.

6.2.10 Статистична перевірка адекватності рівняння регресії

Статистична перевірка адекватності рівняння регресії заключається у перевірці того, наскільки добре воно апроксимує отримані експериментальні точки \bar{y}_k ($k = \overline{1, N}$).

Перевірка адекватності виконується за F -критерієм Фішера

$$F = \frac{S_{ao}^2\{y\}}{S_{\varepsilon}^2\{y\}}, \quad (6.16)$$

де $S_{ao}^2\{y\}$ – дисперсія неадекватності (іноді її називають дисперсією адекватності); $S_{\varepsilon}^2\{y\}$ – дисперсія відтворення.

Отримане дисперсійне F -відношення порівнюємо з критичним (табличним) значенням F -статистики, яке обираємо для прийнятого рівня значущості q та чисел ступенів волі відповідно чисельника й знаменника:

$$\left. \begin{aligned} f_{\text{числ}} &= f_{ao} = f_1 = N - d \\ f_{\text{знам}} &= f_B = f_2 = N(\gamma - 1) \end{aligned} \right\}, \quad (6.17)$$

причому f_1 обираємо у горизонтальному заголовку зверху таблиці, а f_2 – у вертикальному заголовку зліва. Якщо розраховане F -відношення виявилось менше критичного $F_{\text{кр}}$, тобто якщо

$$F < F_{\text{кр}} = F_{\text{табл}}, \quad (6.18)$$

то отримане рівняння регресії з прийнятим рівнем значущості q вважається адекватним експериментальним даним, у протилежному випадку гіпотеза про адекватність відкидається.

Перевірка адекватності можлива при $f_{ao} = N - d > 0$. Якщо кількість N варіантів варіювання плану ПФЕ дорівнює кількості усіх значущих оцінок коефіцієнтів рівняння регресії ($N - d$), то для перевірки гіпотези про адекватність математичного опису ступенів волі не

залишається ($f_{a0} = 0$). Якщо деякі оцінки коефіцієнтів регресії виявились незначущими, то кількість d членів перевіряемого рівняння у цьому випадку менше кількості N варіантів варіювання ($N > d$), і для перевірки гіпотези про адекватність залишається один чи декілька ступенів волі ($f_{a0} > 0$).

Іноді може виявитись, що $F \leq 1$. У такому випадку рівняння регресії адекватно експериментальним даним при будь-якому рівні значущості q і при будь-якій кількості ступенів волі f_1, f_2 , тобто немає необхідності звертатися до статистичних таблиць та вибирати $F_{кр}$, оскільки завжди $F_{кр} > 1$ за будь-яких кінцевих f_1, f_2 .

Фізичний зміст перевірки адекватності складається у тому, що дисперсія адекватності $S_{a0}^2\{y\}$ не повинна значно перевищувати дисперсію відтворення $S_B^2\{y\}$, яка характеризує помилку експерименту.

Для підвищення надійності перевірки адекватності часто проводять додатково γ паралельних експериментів у базовій точці $x_i = 0$ ($i = \overline{1, N}$) або в реальному масштабі при $x_i = x_{i,0}$, тоді кількість точок, за якими оцінюється адекватність рівняння регресії, збільшується на одну і стає рівним $N + 1$, тобто збільшується на 1 і кількість ступенів волі f_{a0} . Однак слід відзначити, що базова точка при ПФЕ не приймає участі у розрахунку коефіцієнтів рівняння.

6.2.11 Використання нормованого рівняння регресії для передбачення цільової функції

Рівняння регресії дозволяє передбачати математичне очікування цільової функції при заданому наборі значень вхідних факторів. Однак слід враховувати особливості використання рівнянь регресії у нормованому вигляді і у звичайному масштабі. Розглянемо, як використовується нормоване рівняння регресії для передбачення математичного очікування вихідного показника об'єкта.

Від рівняння у нормованому вигляді (6.8) можна перейти до рівняння у звичайному масштабі (6.1) через те, що на практиці іноді зручніше використовувати рівняння, в яке підставляють фактори у

реальних фізичних величинах. Перехід виконують за допомогою наступних співвідношень:

$$\begin{aligned} a_0 &= a_0^* - \sum_{i=1}^N \frac{a_i^* x_{i,0}}{\lambda_i} + \sum_{i,l=1} a_{i,l}^* \frac{x_{i,0} x_{l,0}}{\lambda_i \lambda_l}; \\ a_i &= \frac{a_i^*}{\lambda_i} - 2 \sum_{\substack{i,l=1 \\ i \neq l}}^N \frac{a_{i,l}^* x_{l,0}}{\lambda_i \lambda_l}; \quad a_{i,l} = \frac{a_{i,l}^*}{\lambda_i \lambda_l} \end{aligned} \quad (6.19)$$

Якщо взаємодії відсутні, формули (6.19) суттєво спрощуються:

$$a_0 = a_0^* - \sum_{i=1}^N \frac{a_i^* x_{i,0}}{\lambda_i}; \quad a_i = \frac{a_i^*}{\lambda_i}; \quad a_{i,l} = 0. \quad (6.20)$$

Співвідношення (6.19), що дозволяють здійснити перехід від рівняння у нормованому вигляді (6.8) до рівняння у звичайному масштабі (6.1), отримані шляхом підстановки у (6.8) формули (6.3).

6.2.12 Приклад

Залежність між відгуком Y і факторами X_1 та X_2 описується двовірним нелінійним рівнянням $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2$. Експериментальні дані наведені у таблиці 6.2.

Побудувати план-матрицю проведення експерименту, обчислити коефіцієнти a_1 , a_2 , a_3 і знайти аналітичний запис моделі залежності $y = f(x_1, x_2)$.

Таблиця 6.2 – Експериментальні дані

№ п/п	1	2	3	4
X1	6	5	6	5
X2	80	70	70	80
Y	61,5	46,6	55,7	51,5

Розв'язання

Матрицю планування наведемо у таблиці 6.3.

Таблиця 6.3 – Матриця планування

№ п/п	X0	X1	X2	X1X2	$y^{експ}$	$y^{роз}$
1	1	1	1	1	61,5	61,48
2	1	-1	-1	1	46,6	46,34
3	1	1	-1	-1	55,7	55,7
4	1	-1	1	-1	51,5	51,6

У табл. 6.3 використовуються такі позначення: $y^{експ}$ – значення y , отримані в результаті експерименту; $y^{роз}$ – розраховані значення y .

Обчислимо коефіцієнти за формулами:

$$a_0 = \frac{\sum_{n=1}^N y^{експ}}{N} = \frac{61,5 + 46,6 + 55,7 + 51,5}{4} = \frac{215,3}{4} = 53,82,$$

$$a_1 = \frac{\sum_{n=1}^N x_{1n} y_n^{експ}}{N} = \frac{1 \cdot 61,5 + (-1) \cdot 46,6 + 1 \cdot 55,7 + (-1) \cdot 51,5}{4} = \frac{19,1}{4} = 4,77,$$

$$a_2 = \frac{\sum_{n=1}^N x_{2n} y_n^{експ}}{N} = \frac{1 \cdot 61,5 + (-1) \cdot 46,6 + (-1) \cdot 55,7 + 1 \cdot 51,5}{4} = \frac{10 \cdot 9}{4} = 2,72,$$

$$a_3 = \frac{\sum_{n=1}^N x_{2n} x_{2n} y_n^{експ}}{N} = \frac{1 \cdot 61,5 + 1 \cdot 46,6 + (-1) \cdot 55,7 + (-1) \cdot 51,5}{4} = \frac{0,7}{4} = 0,17.$$

Таким чином, математична модель має вигляд:

$$y_{роз} = 53,82 + 4,77x_1 + 2,72x_2 + 0,17x_1x_2.$$

6.3 Порядок виконання роботи

6.3.1 Реалізація ПФЕ у середовищі Statgraphics

Робота у середовищі Statgraphics складається з двох етапів:

- а) побудова МП;
- б) аналіз результатів експерименту.

Для виконання даної лабораторної роботи у середовищі Statgraphics Plus спочатку потрібно із панелі меню обрати ***Special / Experimental Design / Create Design***. З'явиться вікно, що зображене на рис. 6.2, де можна обрати клас плану, кількість вихідних змінних та факторів. Для нашого завдання обираємо параметр Screening та кількість залежних змінних та незалежних факторів згідно рис. 6.2.

Друге вікно, що з'явиться після натискання клавіші **ОК** (рис.6.3) дозволяє обрати параметри обраних факторів, такі як назву, нижній та верхній рівні кодування. Задаємо назви факторів та рівні варіювання. Для нашої задачі ці назви будуть x_1 , x_2 , x_3 , для кожного з факторів нижній рівень варіювання дорівнює -1 , а верхній $+1$.

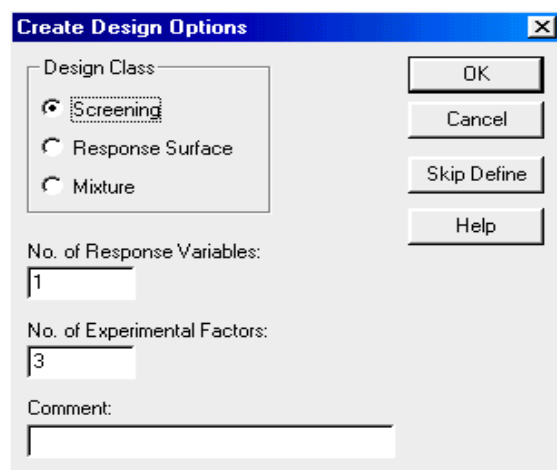


Рисунок 6.2 – Вікно вибору параметрів планування

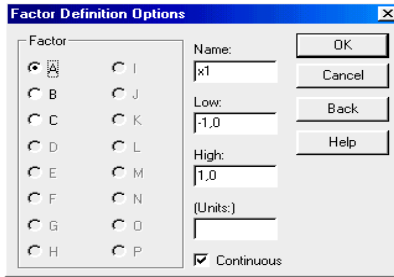


Рисунок 6.3 – Вікно визначення параметрів факторів

Вікно визначення параметрів відгуків (рис. 6.4) дозволяє обрати назву змінної – відгуку. У нашому випадку назва Y.

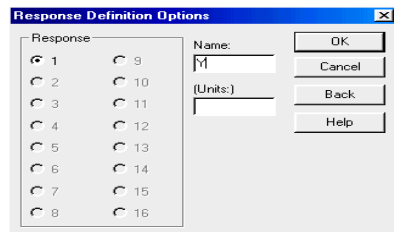


Рисунок 6.4 – Вікно визначення параметрів відгуку

Серед запропонованих МП (рис. 6.5) обираємо план відповідний ПФЕ (2^3 , кількість дослідів 8).

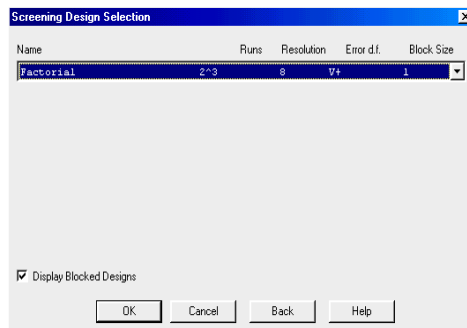
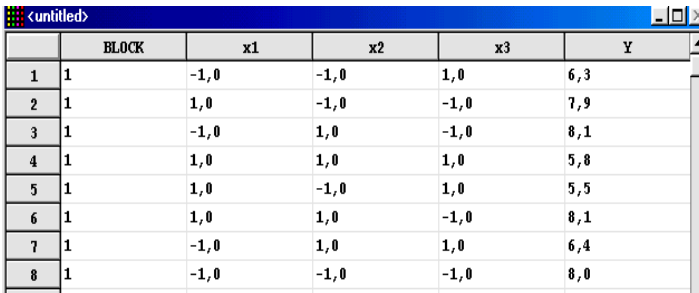


Рисунок 6.5 – Вікно вибору типу плану


В результаті цих дій отримуємо матрицю планування (рис. 6.6) для нашого експерименту, де у якості параметру a_0 використано змінну Block. У стовпець зі змінною відгуку (Y) занесемо результати експерименту і приступимо до другого етапу, тобто до аналізу результатів.



	BLOCK	x1	x2	x3	Y
1	1	-1,0	-1,0	1,0	6,3
2	1	1,0	-1,0	-1,0	7,9
3	1	-1,0	1,0	-1,0	8,1
4	1	1,0	1,0	1,0	5,8
5	1	1,0	-1,0	1,0	5,5
6	1	1,0	1,0	-1,0	8,1
7	1	-1,0	1,0	1,0	6,4
8	1	-1,0	-1,0	-1,0	8,0

Рисунок 6.6 – Вікно початкових даних для МП

Аналіз результатів починається з вибору параметру оптимізації (рис. 6.7).

Для наступного аналізу необхідною буде іконка , натиснувши на яку оберемо наступні пункти (рис. 6.8).

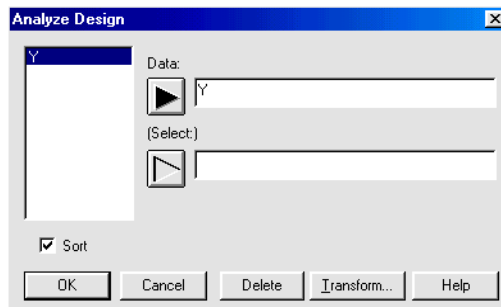


Рисунок 6.7 – Вікно вибору параметру оптимізації

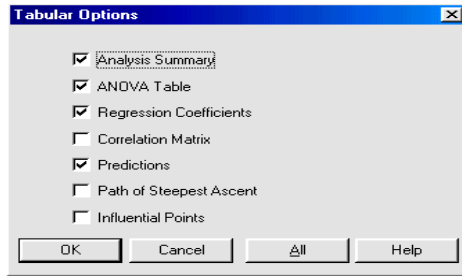


Рисунок 6.8 – Вікно вибору необхідних видів аналізу

На основі обраних даних отримуємо наступні результати :

```

Analysis Summary
-----
File name: <Untitled>

Estimated effects for Y
-----
average = 7,0125 +/- 0,0125
A:x1     = -0,375 +/- 0,025
B:x2     = 0,175 +/- 0,025
C:x3     = -2,025 +/- 0,025
AB       = 0,075 +/- 0,025
AC       = -0,325 +/- 0,025
BC       = 0,025 +/- 0,025
-----
Standard errors are based on total error with 1 d.f.

```

Рисунок 6.9 – Вікно оцінки внесків факторів

Таблиця сумарного аналізу на рис. 6.9 показує кожний з оцінених внесків факторів, а також стандартну помилку кожного із внесків.

Наступна у списку таблиця дисперсійного аналізу (ANOVA Table, рис. 6.10) показує значущість кожного з факторів. Нагадаємо, що значимими вважаються фактори або їх взаємодії, якщо значення P-value є меншим за 0,05. У нашому випадку значимими є фактори x1, x3 та їх взаємодія x1x3. Але незалежно від цього експериментатор має право враховувати ці фактори у рівнянні регресії при наступному плануванні для отримання оптимального результату. Критерій R-squared показує наскільки добре підібрана модель. У нашому випадку модель є адекватною.

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:x1	0,28125	1	0,28125	225,00	0,0424
B:x2	0,06125	1	0,06125	49,00	0,0003
C:x3	2,20125	1	2,20125	6561,00	0,0079
AB	0,01125	1	0,01125	9,00	0,2048
AC	0,21125	1	0,21125	169,00	0,0489
BC	0,00125	1	0,00125	1,00	0,5000
Total error	0,00125	1	0,00125		
Total (corr.)	2,76875	7			

R-squared = 99,9857 percent
 R-squared (adjusted for d.f.) = 99,9002 percent
 Standard Error of Est. = 0,0353553
 Mean absolute error = 0,0125
 Duchin-Watson statistic = 2,0

Рисунок 6.10 – Таблиця дисперсійного аналізу

З таблиці коефіцієнтів регресії (рис. 6.11) ми бачимо підібрані для наших даних коефіцієнти рівняння регресії і саме рівняння регресії.

constant	= 7,0125
A:x1	= -0,1875
B:x2	= 0,0875
C:x3	= -1,0125
AB	= 0,0375
AC	= -0,1625
BC	= 0,0125

$$\begin{aligned}
 Y = & 7,0125 - 0,1875 \cdot x_1 + 0,0875 \cdot x_2 - 1,0125 \cdot x_3 + 0,0375 \cdot x_1 \cdot x_2 - \\
 & 0,1625 \cdot x_1 \cdot x_3 + 0,0125 \cdot x_2 \cdot x_3
 \end{aligned}$$

Рисунок 6.11 – Коефіцієнти рівняння регресії

Наступна таблиця прогнозувань (рис. 6.12) дозволяє нам порівняти значення отриманих у результаті експерименту значень відгуку та значень відгуку для підбраної математичної моделі. Отримана модель (як ми з'ясували раніше) відповідає отриманим у результаті експерименту даним, тобто за її допомогою можливо прогнозувати цільову функцію.

Estimation Results for Y

Row	Observed Value	Fitted Value	Lower 95.0% CL for Mean	Upper 95.0% CL for Mean
1	6,3	6,2875	5,86728	6,70772
2	8,0	8,0125	7,59228	8,43272
3	8,1	8,1125	7,69228	8,53272
4	5,5	5,5125	5,09228	5,93272
5	8,1	8,0875	7,66728	8,50772
6	6,4	6,4125	5,99228	6,83272
7	7,9	7,8875	7,46728	8,30772
8	5,8	5,7875	5,36728	6,20772

Рисунок 6.12 – Порівняння експериментальних і отриманих даних

6.3.2 Програмна реалізація повного факторного експерименту засобами мови R

Наведемо приклад програмної реалізації побудови математичної моделі за експериментальними даними для кількості факторів $N = 6$, де x_1, x_2, \dots, x_6 – відповідні фактори, y – відгук.

```

x1 <- c(-1,...    1)
x2 <- c(-1,...    1)
x3 <- c(1,...     1)
x4 <- c(-1,...   -1)
x5 <- c(-1,...   -1)
x6 <- c(1,...     1)
y<- c(19.0832376437    ,...    17.4004932366)
x12<-x1*x2 # в залежності від кількості x скласти всі можливі
комбінації
x13<-x1*x3
x14<-x1*x4
x15<-x1*x5
x16<-x1*x6
x23<-x2*x3
x24<-x2*x4
x25<-x2*x5
x26<-x2*x6
x34<-x3*x4

```

```

x35<-x3*x5
x36<-x3-x6
x45<-x4*x5
x46<-x4*x6
x56<-x5*x6
x<-rbind(x1, x2, x3, x4, x5, x6, x12,x13,x14,x15,
          x16,          x23,          x24,          x25,
          x26,          x34,          x35,
          x36,          x45,          x46,
          x56           )
N=length(y)
a0<-sum(y)/N
a<-c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) #    кіль-
кість нулів відповідає кількості комбінацій з x
M=length(a)
for(j in 1:N)
  for(i in 1:M)
    a[i]<-a[i]+(y[j]*x[i, j])
for(i in 1:M)
  a[i]<-a[i]/N
s<-c("x1", "x2", "x3", "x4", "x5", "x6", "x1*x2", "x1*x3",
"x1*x4", "x1*x5", "x1*x6", "x2*x3", "x2*x4", "x2*x5",
"x2*x6", "x3*x4", "x3*x5", "x3*x6", "x4*x5", "x4*x6",
"x5*x6")
s1=""
for(i in 1:M)
  s1<-paste0(s1, sprintf("% + f*s", a[i], s[i]))
s2<-paste0("y = ", a0, s1)
s2

```

Результати роботи RStudio

```

[1] "y = 15.0781017293312-0.233105*x1-0.308488*x2+
0.261353*x3-0.187244*x4-0.084091*x5+0.048129*x6-
0.123289*x1*x2+0.156115*x1*x3-0.080164*x1*x4-
0.281490*x1*x5+0.033936*x1*x6-0.276591*x2*x3-0.067796*x2*x4-
0.196198*x2*x5+0.059099*x2*x6-0.235224*x3*x4-
0.160939*x3*x5+0.213224*x3-x6-0.125783*x4*x5+0.551356*x4*x6-
0.262511*x5*x6"

```

6.4 Завдання на лабораторну роботу

6.4.1 Ознайомитись з методикою ПФЕ.

6.4.2 Отримати індивідуальне завдання у викладача.

6.4.3 Згенерувати план експерименту 2^k , де

$$k = \begin{cases} Var, & Var \leq 5 \\ Round(Var * 2 / 5), & 5 < Var \leq 15 \\ Round(Var / 10), & Var > 15 \end{cases}$$

6.4.4 Кількість рівнобіжних дослідів 4.

6.4.5 Дані експерименту згенерувати за формулою:

для пакета Statgraphics – $Rnormal(N, \mu, \sigma^2)$, для пакета Statistica згенерувати випадкову вибірку.

6.4.6 Виконати аналіз плану експерименту.

6.4.7 Згенерувати випадкову вибірку для відгуку.

6.4.8 Побудувати матрицю планування, отримати оцінки коефіцієнтів регресії та рівняння регресії з використанням мови R.

6.4.9 Зробити висновки.

6.4.10 Оформити звіт.

6.4.11 Відповісти на контрольні запитання.

6.5 Зміст звіту

6.5.1 Мета та назва лабораторної роботи.

6.5.2 Дані для аналізу.

6.5.3 Матриця планування.

6.5.4 Оцінки коефіцієнтів регресії та таблиця перевірки адекватності рівняння регресії.

6.5.5 Відповіді на контрольні запитання.

6.5.6 Висновки до лабораторної роботи.

6.6 Контрольні запитання

6.6.1 Дайте визначення теорії планування експерименту.

6.6.2 Дайте визначення активного та пасивного експерименту.

6.6.3 Вимоги для вибору відгуків, факторів, координат базової точки та ступенів варіювання.

6.6.4 Призначення ПФЕ та етапи його проведення.

6.6.5 Що таке матриця планування? Як визначається кількість можливих комбінацій рівнів варіювання?

6.6.6 Властивості МП. Наведіть приклади кодування факторів.

6.6.7 З якою метою необхідна рандомізація дослідів?

6.6.8 Як впливає величина ступенів варіювання на значущість коефіцієнтів і на адекватність рівняння регресії?

ЖИТЕПАТҮҮҮ

1. Gelman A. Regression and Other Stories (Analytical Methods for Social Research) / A. Gelman, J. Hill, A. Vehtari. – Cambridge : Cambridge University Press, 2020. – 548 p.
2. Taleb N.N. Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications (Technical Incerto) / N. N. Taleb. – California : STEM Academic Press, 2020. – 446 p.
3. Bruce P. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python / P. Bruce, A. Bruce, P. Gedeck. – California : O'Reilly Media, 2020. – 368 p.
4. Ramachandran K.M. Mathematical Statistics with Applications in R / K. M. Ramachandran, C. P. Tsokos. – Cambridge : Academic Press, 2020. – 704 p.
5. Berk R.A. Statistical Learning from a Regression Perspective / R. A. Berk. – Berlin : Springer, 2020. – 480 p.
6. Weinberg S.L. Statistics Using R: An Integrative Approach / S. L. Weinberg, D. Harel, S. K. Abramowitz. – Cambridge : Cambridge University Press, 2020. – 650 p.
7. Rosling H. Factfulness: Ten Reasons We're Wrong About the World - and Why Things Are Better Than You Think / H. Rosling. – California : SCEPTRE, 2018. – 341 p.
8. Spiegelhalter D. The Art of Statistics: How to Learn from Data / D. Spiegelhalter. – New York : Basic Books, 2019. – 448 p.
9. Pearl J. The Book of Why: The New Science of Cause and Effect / J. Pearl, D. Mackenzie. – New York : Basic Books, 2018. – 432 p.
10. Rago G. Applied Mathematics: Principles and Techniques / G. Rago. – New York : NY RESEARCH PRESS, 2020. – 212 p.
11. Handbook of Meta-Analysis / Ed. C. H. Schmid, T. Stijnen, I. White // Chapman & Hall/CRC Handbooks of Modern Statistical Methods. – London : Chapman and Hall/CRC, 2020. – 570 p.
12. Solaiman B. Possibility Theory for the Design of Information Fusion Systems (Information Fusion and Data Science) / B. Solaiman, É. Bossé. – Berlin : Springer, 2019. – 288 p.
13. Mishra S.K. Introduction to Unconstrained Optimization with R / S. K. Mishra, B. Ram. – Berlin : Springer, 2019. – 304 p.

14. Advanced Studies in Multi-Criteria Decision Making / Ed. S. B. Amor, A. T.a de Almeida, J. L. de Miranda, E. Aktas // Chapman & Hall/CRC Series in Operations Research. – London : Chapman and Hall/CRC, 2019. – 274 p.

15. Ismay C. Statistical Inference via Data Science: A Modern Dive into R and the Tidyverse / C. Ismay, A. Y. Kim // Chapman & Hall/CRC The R Series. – London : Chapman and Hall/CRC, 2019. – 460 p.

16. Мова програмування R // Режим доступу: https://bookdown.org/geka/applied_analytics/VPZ.html#rlang.

17. Mathematical Analysis and Analytic Number Theory 2019 – <https://www.mdpi.com/books/book/3568-mathematical-analysis-and-analytic-number-theory-2019>.

18. Mathematical Analysis and Applications II – <https://www.mdpi.com/books/book/2110-mathematical-analysis-and-applications-ii>.

19. Statistical Analysis and Stochastic Modelling of Hydrological Extremes – <https://www.mdpi.com/books/book/1751-statistical-analysis-and-stochastic-modelling-of-hydrological-extremes>.

20. Probability and Stochastic Processes with Applications to Communications, Systems and Networks – <https://www.mdpi.com/books/book/6742-probability-and-stochastic-processes-with-applications-to-communications-systems-and-networks>.

21. R Programming for Data Science – <https://bookdown.org/rdpeng/rprogdatascience/>.