Міністерство освіти і науки України
Запорізький національний технічний університет

МЕТОДИЧНІ ВКАЗІВКИ
до виконання лабораторних робіт з дисципліни
«ЕМПІРИЧНІ МЕТОДИ В ПРОЕКТУВАННІ ЕЛЕКТРОННИХ ЗАСОБІВ»
для студентів напряму підготовки
122 «Комп'ютерні науки та інформаційні технології»
(всіх форм навчання)

2017

Методичні вказівки до лабораторних робіт з дисципліни «Емпіричні методи в проектуванні електронних засобів» для студентів напряму підготовки 122 «Комп'ютерні науки та інформаційні технології» (всіх форм навчання) / А. В. Притула, Н.О. Миронова. – Запоріжжя: ЗНТУ, 2017. – 82 с.

Автори: А. В. Притула, к.т.н., професор
Н. О. Миронова, к.т.н., доцент

Рецензент: Т.В. Федорончак, к.т.н., доцент

Відповідальний
за випуск: А.В. Притула, к.т.н., професор

Затверджено
на засіданні кафедри
програмних засобів

Протокол №11
від 06 червня 2017

ЗМІСТ

ВСТУП

Дане видання призначене для вивчення та практичного освоєння студентами усіх форм навчання основ емпіричних методів програмної інженерії.

Відповідно до графіка студенти перед виконанням лабораторної роботи повинні ознайомитися з конспектом лекцій та рекомендованою літературою. Звичайно, в дані методичні вказівки неможливо було внести весь матеріал, необхідний для виконання та захисту лабораторних робіт. Тому тут містяться основні, базові теоретичні відомості, необхідні для виконання лабораторних робіт. Таким чином для виконання лабораторної роботи та при підготовці до її захисту необхідно ознайомитись з конспектом лекцій та проробити весь матеріал, наведений в переліку рекомендованої літератури. При цьому не варто обмежуватись лише наведеним списком.

Для одержання заліку з кожної роботи студент здає викладачу цілком 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).

Текст розміщують тільки з однієї сторони листа. Поля сторінки з усіх боків – 20 мм. Аркуші скріплюють за допомогою канцелярських скріпок або вміщують у канцелярський файл.

Під час співбесіди при захисті лабораторної роботи студент повинний виявити знання про мету роботи, по теоретичному матеріалу, про методи виконання кожного етапу роботи, по змісту основних розділів оформленого звіту з демонстрацією результатів 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.

ЛАБОРАТОРНА РОБОТА №1 СТАТИСТИЧНИЙ АНАЛІЗ І ПЕРВИННА ОБРОБКА ДАНИХ

Мета роботи

Ознайомитися з можливостями пакету статистичної обробки даних Statgraphics. Навчитися використовувати пакет статистичної обробки даних Statgraphics для первинного аналізу даних. Засвоїти основи мови програмування R для статистичних обчислень та аналізу та даних.

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 *x1,x2,…* що мають ймовірності *p1,p2,…* математичне очікування дорівнює:

$$M(X) = \sum x_i p_i \tag{1.1}$$

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

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

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 зазвичай позначають $\sigma^2 = D(X)$. Дисперсія D(X) випадкової величини – це математичне сподівання квадрату відхилень значення випадкової величини від її математичного очікування:

$$D(X) = M\left[(x - M(X))^2\right] \tag{1.3}$$

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 зветься 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 виміру 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 виміру вихідної випадкової величини. Частіше за все з нормування моментів використовуються асиметрія та ексцес – відповідно третій та четвертий нормовані центральні моменти:

$$асиметрія = \frac{M\left[x - M(X)\right]^3}{\left(\sqrt{D(X)}\right)^3} \tag{1.4}$$

$$ексцес = \frac{M\left[x - M(X)\right]^4}{\left(\sqrt{D(X)}\right)^4} \tag{1.5}$$

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 $\overline{x}$, 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 від’ємний, 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 $\varphi(x)$ зветься розв'язок x_p рівняння $\varphi(x) = p$. Величину x_p називають квантілем рівня р 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 (або 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

$$\nu_i = \frac{m'_i}{N} \tag{1.6}$$

де m'_i - число елементів, що потрапили в *i*-й інтервал;

N – загальна кількість елементів досліджуваної сукупності.

Ширину інтервалів δ_x звичайно приймають постійною (або необхідно кожний раз враховувати удільну вагу ширини інтервалу) для всіх інтервалів і розраховують по формулі:

$$\delta_x = \frac{x_{\max} - x_{\min}}{k} \tag{1.7}$$

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

$$k \approx 1 + 3.2\lg n \tag{1.8}$$

Число k, що розраховане за формулою (1.8), округлюють до найближчого цілого. Зазвичай $k = 6...12$. При невеликій кількості k інтервалів можна пропустити характерні особливості кривої розподілу, а при дуже великому значенню 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 по формулі (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 закриту (тобто включаючи те число, яким позначена границя), а круглою – відкриту 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.

$$\sum_{i=1}^{k} m'_i = N \tag{1.10}$$

або

$$\sum_{i=1}^{k} \nu_i = 1 \tag{1.11}$$

де i – номер інтервалів.

Гістограма – графічне уявлення ряду розподілу, де по осі х відкладаються границі інтервалів, а по осі у – частота m'_i.

Порядок виконання роботи

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 Analisys,** натиснути на передньому плані кнопку **Tabular Option** та вибрати перші 4 пункти із запропонованих (рис.1.1): **Analysis summary** – аналіз вибірки, **Summary Statistics for Col_1** – сумарна статистика для змінної, **Percentiles for Col_1** – відсоткова статистика, **Frequency Tabulation** – таблиця частот (рис. 1.2).

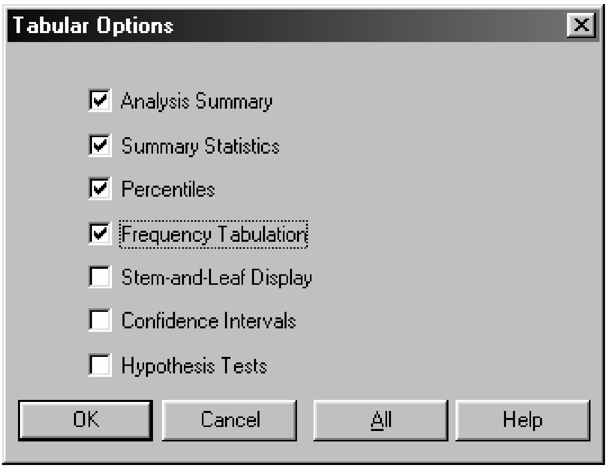


Рисунок 1.1 – Список опцій

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.

Frequency Tabulation for Col_1

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 1.2 – Таблиця 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. 1.3).

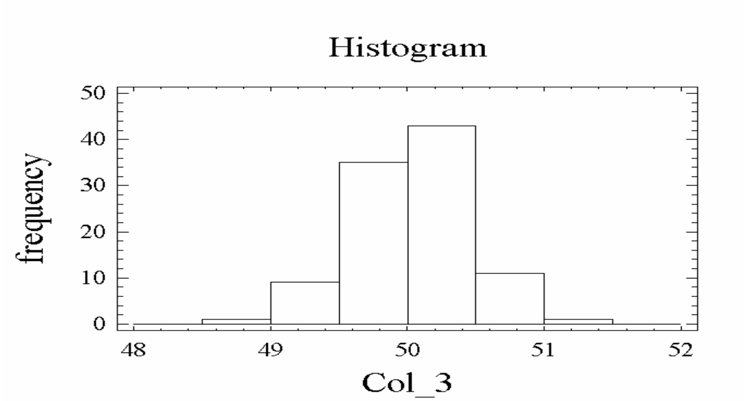


Рисунок 1.3 – Гістограма

Завдання на лабораторну роботу

1. Скориставшись рекомендованою літературою вивчити можливості пакету статистичної обробки даних Statgraphics та мови програмування R.

2. Вивчити загальні положення теорії ймовірностей.

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 $Rnormal(N, \mu, \sigma^2)$.

4. Зберегти отриману вибірку у форматі *.txt.

5. Побудувати гістограму, таблицю частот, отримати описові статистики ряду розподілу з використанням пакету статистичної обробки даних Statgraphics.

6. Зберегти результати аналізу у форматі *.rtf.

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

8. Зробити висновок.

9. Оформити звіт.

10. Відповісти на контрольні питання.

Зміст звіту

1 Назва і мета роботи.

2. Постановка завдання.

3. Вихідні дані.

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

5. Висновки по результатам аналізу.

Контрольні питання

1. Основні характеристики розподілу ймовірностей. Записати аналітичні вирази.

2. Що таке квантіль, мода, медіана?

3. Як визначити знак коефіцієнта асиметрії з вигляду графіка щільності ймовірності?

4. Як визначити знак коефіцієнта ексцесу з вигляду графіка щільності ймовірності?
5. Що таке ряди розподілу?
6. Які характеристики розподілу ви знаєте?
7. Як будується гістограма?
8. Що таке таблиця частот?
9. На що впливає ширина інтервалу?

ЛАБОРАТОРНА РОБОТА №2
СТАТИСТИЧНА ПЕРЕВІРКА ГІПОТЕЗ

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 R.

Короткі теоретичні відомості

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, тому що: по-перше, можна використовувати нормальний закон у якості першого наближення; по-друге, легко підібрати таке перетворення досліджуваної величини, яке видозмінить вихідний "не нормальний" закон розподілу, перетворюють його у нормальний.

Нормальний закон розподілу характеризується щільністю ймовірності виду:

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

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

Зміст цієї форми запису означає - 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 ймовірність того, що випадкова величина розташована у інтервалі між x_1 та x_2.

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

Теорема А.М. Ляпунова доводить, що сума достатньо великого числа незалежних або слабозалежних випадкових величин підкоряється нормальному закону. Тут ми маємо приклад з помилками вимірювань - якщо помилка 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 наведений на рис. 2.1.

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

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

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

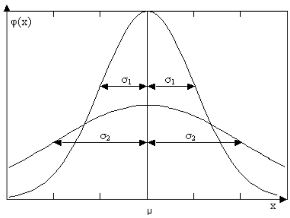


Рисунок 2.1– Щільність нормального розподілу з однаковим середнім μ та різними значеннями дисперсії σ_1, σ_2

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

$$P(\mu-\sigma_x \le x \le a+\sigma_x)=0.67$$
$$P(\mu-2\sigma_x \le x \le a+2\sigma_x)=0.95$$
$$P(\mu-3\sigma_x \le x \le a+3\sigma_x)=0.99$$

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 невідомий, але умова, вказана в наведеному прикладі виконується, то існує підстава вважати, що величина розподілена нормально; в іншому випадку вона не розподілена нормально.

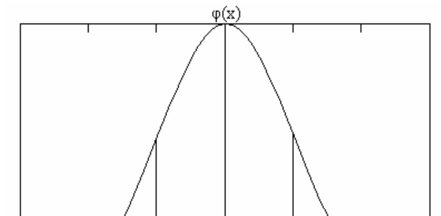


Рисунок 2.2 – Ілюстрація правила трьох сигм

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 справедливість 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 по формулам:

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

де $z_i = z_{i+1}$, S{x}-середнє квадратичне відхилення змінної x.

При нормуванні границь інтервалів найменше значення z_i, тобто z_1, визначаємо рівним $-\infty$, найбільше, тобто 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:

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

Теоретичні частоти нормального розподілу розрахуємо з

$$m_i = NP_i \tag{2.5}$$

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

В якості критерію перевірки нульової гіпотези приймається випадкова величина.

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

або

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

де k – кількість інтервалів ряду розподілу.

Ця величина випадкова, так як 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 розподілу. Зведення до квадрату різниць частот усувається можливістю взаємного погашення додатних і від'ємних різниць.

При необмеженому зростанні об'єму вибірки $(N \to \infty)$ закон розподілу випадкової величини (2.6), незалежно від того, якому закону розподілу підпорядкована генеральна сукупність, прагне до закону розподілу χ^2 з f ступенями свободи. Тому випадкова величина (2.6) позначена 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:

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

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) виявилося менше критичного $\chi^2_{кр}$, що знаходять по таблицям для відповідного рівня значимості q і числа ступенів свободи f, тобто якщо

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

то нема підстав відхилити нульову гіпотезу про нормальний розподіл. В протилежному випадку (при $\chi^2 > \chi^2_{кр}$) 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), повинне бути

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

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.

Порядок виконання роботи

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; μ ; σ^2), де N – кількість дослідів, μ – середнє квадратичне, σ^2 – дисперсія, відповідно до завдання.

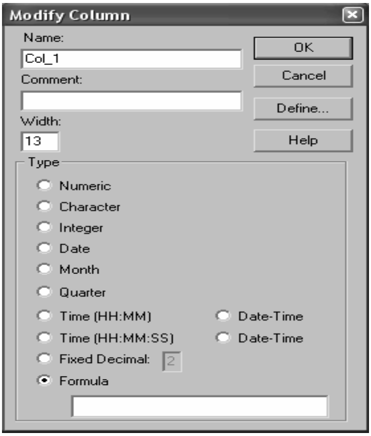


Рисунок 2.5 – Вікно з параметрами стовпця даних

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

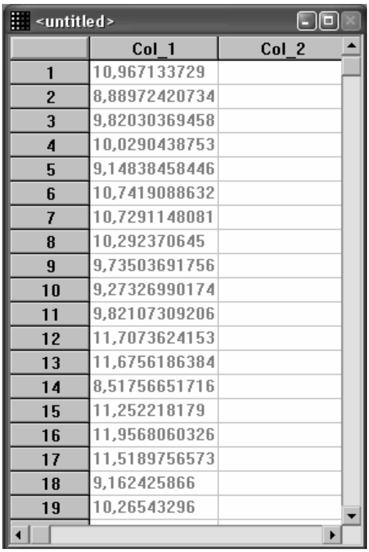


Рисунок 2.6 – Згенеровані значення для стовпця Col_1

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

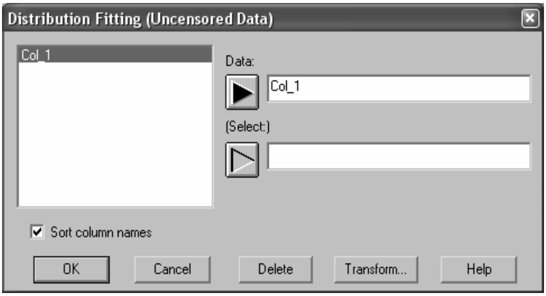


Рисунок 2.7 – Вікно Distribution Fitting

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

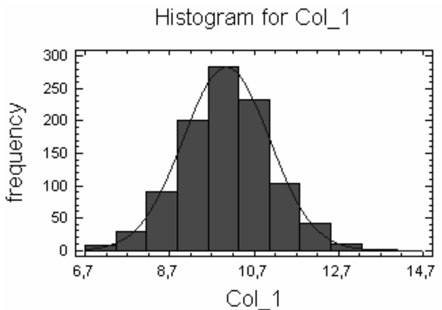


Рисунок 2.8 - Гістограма і перевірка її на нормальність

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, зображеній на рис. 2.9, **STATGRAPHICS** аналізує отриманий результат автоматично. Судячи з того, що значення **P-value** більше, ніж 0.1, (P-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	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

Завдання на лабораторну роботу

1. Вивчити загальні положення теорії статистичної перевірки гіпотез.
2. 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 пакета Statgraphics використати функцію $Rnormal(N,\mu,\sigma^2)$.
3. Зберегти отриману вибірку у форматі *.txt.
4. Перевірити гіпотезу про нормальний розподіл отриманої вибірки на основі критерію Пірсона χ^2 та критерію Колмогорова.
6. Зберегти результати аналізу у форматі *.rtf.
7. Перевірити гіпотезу про нормальний розподіл отриманої вибірки на основі критерію Пірсона χ^2 та критерію Колмогорова з використанням внутрішніх функцій мови R.
8. Зробити висновок.
9. Оформити звіт.

10. Відповісти на контрольні питання.

Зміст звіту

1. Назва і мета роботи.
2. Теоретичний аналіз щодо критеріїв перевірки статистичних гіпотез.
3. Індивідуальне завдання.
4. Отримані результати обчислень та графіки.
5. Аналіз отриманих результатів та висновки.

Контрольні питання

1. Які ви знаєте параметри нормального закону розподілу випадкових величин.
2. Наведіть аналітичне вираження щільності ймовірності для нормального закону розподілу.
3. Накресліть функцію щільності ймовірності для нормального закону.
4. Чому дорівнюють коефіцієнт асиметрії та коефіцієнт ексцесу для нормального закону?
5. Яким критерієм визначається закон розподілу випадкової величини? Записати аналітичний вираз.
6. Як змінюється вигляд гістограми при зміні величини інтервалу?
7. Як використовується правило трьох сігм для визначення закону розподілу випадкової величини?
8. Необхідність застосування статистичних методів обробки результатів спостережень.
9. У чому полягає основний принцип сатистичних гіпотез?
10. Які гіпотези можна перевірити за допомогою критерія Пірсона, Колмогорова, Фішера, Стьюдента, Кохрена?
11. Як визначити довірчі інтервали для математичного сподівання?

ЛАБОРАТОРНА РОБОТА №3
МЕТОДИ ВИВЧЕННЯ ВЗАЄМОЗВ'ЯЗКІВ

Мета роботи

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

Короткі теоретичні відомості

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

$$\mathrm{M}(y\,|\,x) = f(x;\Theta) \tag{3.1}$$

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.

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 і x.

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

$$r = \frac{M[(x - M(x))(y - M(y))]}{\sqrt{D(x) \cdot D(y)}} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y} \tag{3.2}$$

де коваріація *cov(x,y)* – другий центральний змішаний момент двомірної випадкової величини *(x,y)*, а σ_x і σ_y –середньоквадратичні (безумовні) відхилення відносно компонент *y* і *x*. Якщо коефіцієнт кореляції додатній, то це означає однаковий характер тенденції взаємопов'язаної зміни випадкових компонент *y* 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 і y (зі збільшенням x зменшується $M(y|x=x)$.

Вибіркове значення $\hat{r}$ **коефіцієнта кореляції** (тобто статистична оцінка $\hat{r}$ невідомого r) розраховується по вхідним даним за формулою:

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.3}$$

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

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

Але у випадку сукупного нормального розподілення досліджуємих випадкових величин x і y, що досліджуються, 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 і y відхиляється від нормального, одна із досліджуваних величин не є випадковою тощо) коефіцієнт кореляції можна використовувати лише в якості однієї із можливих характеристик ступеня тісноти зв'язку.

Кореляційне відношення. Найбільш зручною є ситуація, в якій характер вибіркових даних допускає їх групування по осі пояснювальної змінної і можливість підрахунку так званих "окремих" середніх ординат $\overline{y_i}$ всередині кожного (*i*-го) інтервалу групування. Нехай таке групування виконано. При цьому k – кількість інтервалів групування по осі абсцис; m_i *(i=1,2,…,k)* – число вибіркових точок, що потрапили в *i*-ий інтервал; $\overline{y_i} = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$ – середнє значення ординат точок, що потрапили в *i*-ий інтервал групування. Тоді вибірковою оцінкою дисперсії $\sigma_f{}^2$ буде величина

$$s-^2{}_{y(x)} = \frac{1}{n}\sum_{i=1}^{k} m_i(\overline{y_i} - \overline{\overline{y}}) \tag{3.4}$$

де загальне середнє $\overline{\overline{y}} = \frac{\sum_{i=1}^{k} m_i \overline{y_i}}{n}$.

У цьому випадку ми можемо отримати вираз для *кореляційне відношення* залежної змінної у по незалежній змінній х

$$\mathrm{R}_{y \cdot x}{}^2 = \frac{S{-}y(x)^2}{S_y{}^2} \tag{3.5}$$

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

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

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 більше 1.

Із $|R| = 1$ випливає наявність однозначного функціонального зв'язку між y і x, і, навпаки, однозначний функціональний зв'язок між x і y говорить про то, 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 $R_{y/x} = 0$. Навпаки, якщо $R_{y/x} = 0$, то $\overline{y_i} = \overline{y}$, і, як наслідок, часткові середні $\overline{y_i}$ не залежать від x.

Перевірка гіпотези про відсутність кореляційного зв'язку. Для побудови відповідного критерію перевірки даної гіпотези скористаємось фактом 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 розподілення залежної змінної y(x) при будь-якому фіксованому x описуються нормальним законом з постійною дисперсією σ^2.

Тому, якщо виявиться, що

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

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

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

Постановка задачі. Статистичний підхід до задачі побудови (точніше, відновлення) функціональної залежності y від x ґрунтується на припущенні, що нам відомі деякі експериментальні дані (xi, yi), де yi – значення відгуку при заданому значенні фактора xi, i змінюється від 1 до n. Пари значень (xi,yi) 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 x; інша частина – випадкова по відношенню до x. Позначимо першу через $f(x)$, другу через ε і представимо відгук у виді

$$y = f(x) + \varepsilon \qquad (3.6)$$

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 (3.7) приймає вид:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\ i = 1,...,n \qquad (3.7)$$

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

При вирішенні задачі лінійного регресійного аналізу використовується два основних 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 $N\left(0,\sigma^2\right)$ з деякою невідомою дисперсією σ^2.

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

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

$$\sum_{i=1}^{n}\left[y_i-\beta_0-\beta_1 x_i\right]^2 \to \min_{\beta_0,\beta_1} \qquad (3.8)$$

Геометричний зміст метода найменших квадратів проілюстровано на рис. 3.1.

Оцінки b_0 та b_1 обчислюємо за формулами:

$$b_1=\frac{\Sigma(x_i-\overline{x})(y_i-\overline{y})}{\Sigma(x_i-\overline{x})^2},\ (i=1..n) \qquad (3.9)$$

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

Рішення рівняння щодо вільного члена (відрізка на осі ординат при $x = 0$) b_0 дає

$$b_0=\overline{y}-b_1\overline{x} \tag{3.10}$$

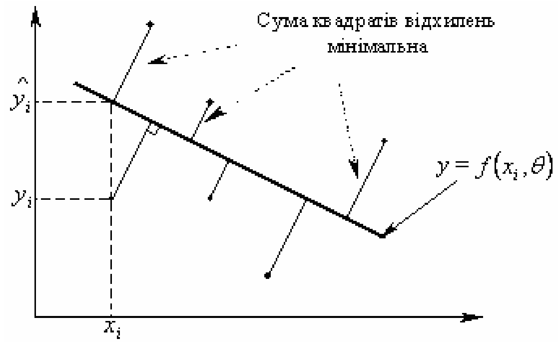


Рисунок 3.1 – Геометричний зміст методу найменших квадратів

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 Стьюдента, який в цьому випадку розраховується:

$$t=\frac{b_i}{s\sqrt{c_{ii}}}, \tag{3.11}$$

де b_i – *i-й* коефіцієнт рівняння регресії, s – середнє квадратичне відхилення, c_{ii} – діагональний коефіцієнт матриці дисперсійковаріацій, тобто матриці $(X^TX)^{-1}$, яка утворюється в процесі рішення системи рівнянь.

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

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.3.1).

Таблиця 3.1 – Таблиця дисперсійного аналізу

Джерело варіації	Число ступенів волі	Суми квадратів SS	Середні квадрати
Обумовлений регресією SS_3	1	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MS_{рег}$
Відносно регресії (залишок) SS_2	n – 2	$\sum_{i=1}^{n}(y_i - \hat{y})^2$	$s^2 = \frac{SS_2}{(n-2)}$
Загальний, скорегований 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 смислі під адекватністю розуміється відповідність моделі процесу чи об’єкту, який вона описує, раніше визначеним умовам.

Адекватність моделі перевіряється по критерію Фішера:

$$F = \frac{MS_{рег}}{s^2} \qquad (3.12)$$

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

Необхідно відзначити, що ця перевірка являється формальною, тому заключне рішення об адекватності моделі треба приймати, виходячи з придатності моделі к практичному використанню по всій сукупності параметрів.

Придатність лінії регресії для цілей передбачення залежить від того, яка частина SS_1 щодо середнього приходиться 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:

$$R^2 = \frac{\left(SS_3, \text{обумовлена регресією}\right)}{\left(\text{повна } SS, \text{скорегована на } \bar{y}\right)} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2},\ i = \overline{1,n} \qquad (3.13)$$

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 і $\hat{y}$ та його звичайно 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.

Порядок виконання роботи

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: градуювання, калібрування, тарировка і т.д. Розглянемо рішення подібної задачі за допомогою градуювального 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 3.2 – Дані каліброваного експерименту однієї компоненти тензовагів

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

У таблиці 3.2 приведені дані градуювального експерименту одного компоненту тензовагів, призначеного для виміру 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. Виміри повторювалися 6 разів. У таблиці приведені також середні відгуки $\overline{y}_i$, і стандартні відхилення s_i^2.

Потрібно побудувати рівняння регресії.

Рішення задачі. Методи регресійного аналізу в пакеті Statgraphics представлені в пункті Relate головного меню. У ньому мається 3 процедури (рис.3.2):

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

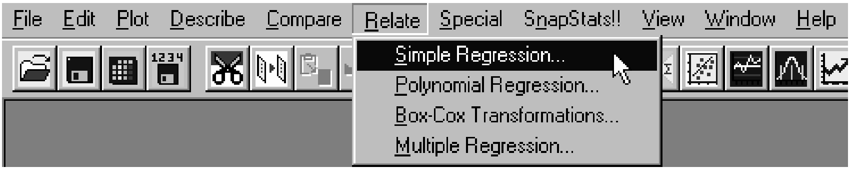
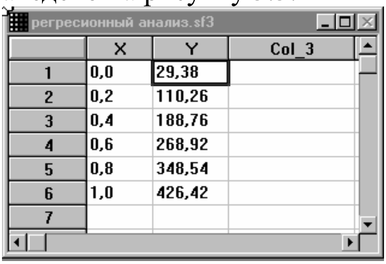


Рисунок 3.2 – Меню процедур регресійного аналізу

Для вирішення поставленої задачі ми припускаємо, що залежності буде мати вид прямої лінії і відповідно до цього вибираємо пункт **Simple Regression (**рис. 3.3**)**.

Вихідні дані наведено на рисунку 3.3.



	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 3.3 – Вихідні дані

В меню простої регресії потрібно вибрати спочатку поле Y (залежну змінну), а потім вибрати X – незалежну змінну(рис.3.4). Після натискання кнопки OK на екран будуть видані підсумки аналізу.

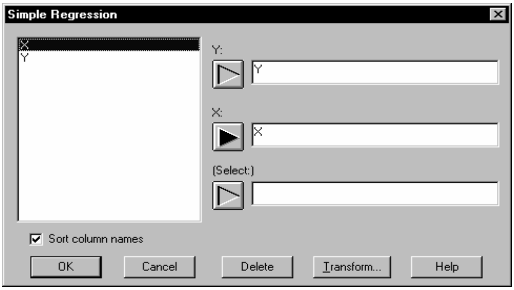


Рисунок 3.4 – Меню простої регресії

Результати. Вікно вибору текстових результатів можна викликати, натиснувши на другу ліворуч кнопку в панелі інструментів аналізу (Tabular Options). При цьому з'явиться вибір з пунктів (рис. 3.5):

– 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.

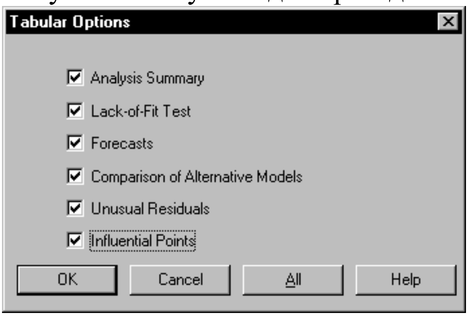


Рисунок 3.5 – Вікно вибору текстових результатів

Для аналізу результатів використовується таблиця з оцінкою коефіцієнтів регресійного рівняння та таблицею дисперсійного аналізу (рис.3.6).

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Y
Independent variable: X
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       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 3.6 – Результати регресійного аналізу

де *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 дисперсійного аналізу є базовою таблицею аналізу варіації і служить для оцінки адекватності запропонованої моделі даних. З таблиці дисперсійного аналізу видно, що отримане рівняння регресії адекватне і має вид $y = 30{,}13 + 397{,}2\,x$.

Відповідність моделі. Для аналізу відповідності моделі використовуємо: Correlation Coefficient – коефіцієнт парної кореляції Y від X; R-Squared – 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.

Графічні результати. Вікно вибору графічних результатів можна викликати, натиснувши на третю ліворуч кнопку в панелі інструментів аналізу (Graphical Options).

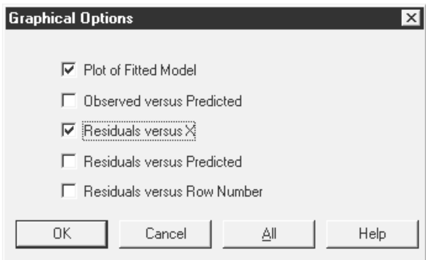


Рисунок 3.7 – Вікно вибору графічних результатів

При цьому з'явиться вибір з пунктів (рис. 3.7):

– 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 скористаємось графіком моделі (Plot of Fitted Model) (рис. 3.8):

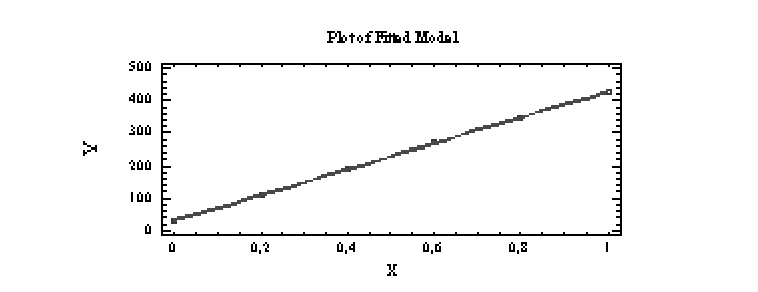


Рисунок 3.8 – Графік побудованої моделі

З графіку на рис. 3.8 бачимо, що модель підібрана добре, всі точки лежать на лінії.

Завдання на лабораторну роботу

1. Вивчити теоретичний матеріал по кореляційному та регресійному аналізу.
2. Отримати індивідуальне завдання (Var=Var mod 9 з додатку Б).
3. Використовуючи дані згідно індивідуального завдання, виконати кореляційний та регресійний аналіз в пакеті Statgraphics та написати відповідну програму на мові R.
4. Отримати таблицю кореляційного аналізу, перевірити коефіцієнти рівняння на значимість, отримати таблицю дисперсійного аналізу, перевірити адекватність моделі, побудувати рівняння регресії та графік, отримати таблицю множинного коефіцієнту регресії, зробити висновки за результатами аналізу.
5. Оформити звіт.

Зміст звіту

1. Назву і мету роботи.
2. Постановку задачі.
3. Таблицю вихідних даних.
4. Основні формулу розрахунку параметрів та критеріїв.
5. Таблиці кореляційного аналізу і графіки з коментаріями.
6. Висновки по результатам аналізу.
7. Вихідні дані.
8. Коефіцієнти рівнянь, таблиця ДА, рівняння регресії

9. Графічна інтерпретація задачі регресійного аналізу.

10. Прогнози та таблиця R^2.

11. Висновки.

Контрольні питання

1. Що значать поняття статистична залежність, функціональна залежність, кореляційна залежність?

2. Коли використовується кореляційний аналіз?

3. Як визначається індекс кореляції і що він показує?

4. Як визначається коваріація і коефіцієнт кореляції? Як визначити вибіркове значення коефіцієнта кореляції?

5. Як визначається кореляційне відношення? Який його фізичний зміст?

6. Яким чином перевіряється гіпотеза про відсутність кореляційного зв'язку?

7. Для чого призначений узагальнений засіб обчислення парних кореляційних характеристик?

8. Як розраховується узагальнений коефіцієнт кореляції?

9. В чому полягає мета регресійного аналізу?

10. Для чого використовується метод найменших квадратів?

11. Наведіть передумови регресійного аналізу.

12. Як виконується перевірка значущості оцінок коефіцієнтів рівняння регресії?

13. За допомогою якого критерію виконується перевірка адекватності рівняння регресії?

14. Як виконується вибір найкращої моделі рівняння регресії?

ЛАБОРАТОРНА РОБОТА №4
ПОВНИЙ ФАКТОРНИЙ ЕКСПЕРИМЕНТ

Мета роботи

Вивчити апарат математичного моделювання методом планування експерименту. Виконати статистичний аналіз рівнянь регресії, з використанням пакету Statgraphics та мови R, побудувати математичну модель за експериментальними даними, отриманими при проведенні повного факторного експерименту типу 2^k

Короткі теоретичні відомості

Постановка задачі. *Планування експерименту* – це вибір числа і умов проведення дослідів, необхідних і достатніх для розв'язку поставлених задач з необхідною точністю. При цьому важливим є:

а) мінімізація загальні кількості дослідів;

б) одночасне варіювання всіми 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 вибирають 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 $x_1, x_2,.., x_n$, 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 ", як на рис.4.1, де $y_1, y_2,.., y_m$ – відгуки, які залежать від параметрів впливу трьох типів, що позначені X,U,V; $x_1, x_2,.., x_n$ – це спостережувані і керовані в процесі експерименту незалежні між собою змінні, що називаються факторами;

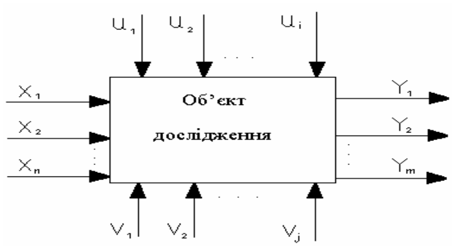


Рисунок 4.1 – Умовна схема «чорного ящика»

$u_1, u_2, .., u_i$ – це спостережувані, але не керовані параметри;

$v_1, v_2, .., v_j$ – 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 основних 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 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 не дуже велика.

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

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{\substack{i,l=1 \\ i \neq l}}^{n} a_{i,l} x_i x_l + \varepsilon \qquad (4.1)$$

де y – значення параметру оптимізації, яке прогнозується; $a_0, a_1, ..., a_n$ – 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.

Вибір параметру оптимізації. Залежні змінні (відгуки) мають відповідати наступним вимогам:

а) мати фізичний смисл і достатньо повно характеризувати досліджуємий об'єкт, процес чи явище;

б) бути відтворюємими, тобто при повторенні досліду 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.

Вибір факторів. Незалежні змінні (фактори) повинні відповідати наступним основним вимогам:

а) фактори мають бути керуємими: 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.

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

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 :

$$\left.\begin{array}{l} x_{i,0} + \lambda_{\mathrm{i}} \le x_{\mathrm{i}\max} \\ x_{i,0} - \lambda_{\mathrm{i}} \ge x_{i\min} \end{array}\right\} \qquad (4.2)$$

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 .

Отримання матриці планування.

Визначення координат точок для пробних експериментів

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

$$\mathrm{X_i} = \frac{x_i - x_{i,0}}{\lambda_i} \qquad (4.3)$$

де $\mathrm{X_i}$ -нормоване значення фактора, x_i -натуральне значення фактору.

Тоді нормовані значення факторів для базового $\mathrm{X_{i,0}}$, нижнього $\mathrm{X}_{\mathrm{i},н}$ та верхнього $\mathrm{X}_{\mathrm{i},в}$ рівней варіювання :

$$\mathrm{X_{i,0}} = \frac{x_{i,0} - x_{i,0}}{\lambda_i} = 0\,; \qquad (4.4)$$

$$\mathrm{X}_{i,н} = \frac{x_{i,н} - x_{i,0}}{\lambda_i} = \frac{x_{i,0} - \lambda_i - x_{i,0}}{\lambda_i} = -1\,; \qquad (4.5)$$

$$\mathrm{X}_{i,в} = \frac{x_{i,в} - x_{i,0}}{\lambda_i} = \frac{x_{i,0} + \lambda_i - x_{i,0}}{\lambda_i} = +1\,, \qquad (4.6)$$

де
$$\left.\begin{aligned} x_{i,в} &= x_{i,0} + \lambda_i \\ x_{i,н} &= x_{i,0} - \lambda_i \end{aligned}\right\}.$$

Таблиця 4.1 – 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

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. 4.1) наведені всі можливі комбінації рівнів, і виявилось чотири. У загальному випадку, коли кількість факторів дорівнює n, а рівнів варіювання k, кількість усіх можливих комбінацій визначається формулою :

$$N = k^n \tag{4.7}$$

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* (ПМП). Після заміни x_i на нормовані X_i за формулою (4.3) рівняння регресії (4.1) приймає вигляд:

$$y = a_0^* + \sum_{i=1}^{n} a_i^* X_i + \sum_{\substack{i,l=1 \\ i \neq l}}^{n} a_{i,l}^* X_i X_l + \varepsilon \,, \qquad (4.8)$$

де $a_0^*, a_i^*, a_{i,l}^*$ – оцінки коефіцієнтів нормованого рівняння регресії, які відрізняються від оцінок коефіцієнтів $\hat{a}_i$ рівняння у звичайному масштабі (4.1). У рівнянні (4.8) на відміну від рівняння (4.1) 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 :

$$\sum_{k=1}^{N} x_{i,k} = 0 \qquad (i = \overline{1,n}) \qquad (4.9)$$

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

$$\sum_{k=1}^{N} x_{i,k}^2 = N \qquad (i = \overline{1,n}) \qquad (4.10)$$

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

$$\sum_{k=1}^{N} x_{i,k} x_{j,k} = 0 \qquad (i, j = \overline{1, n-1}) \qquad (4.11)$$

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. Справа у тому, що установка заданих МП нижніх $x_{i,н}$ чи верхніх $x_{i,в}$ 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 дослідах (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.

Обчислення коефіцієнтів рівняння регресії. Завдячуючи властивостям ортогональності, нормування та симетричності МП формули для обчислення коефіцієнтів рівняння регресії виявляться доволі простими

$$a_i^* = \frac{1}{N} \sum_{k=1}^{N} \mathrm{X}_{\mathrm{i},k} y_k, \qquad (i = \overline{0, n-1}), \qquad (4.12)$$

де $\mathrm{X}_{i,k}$ – 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 (4.12), причому кількість незалежних оцінок має відповідати умові $(i = \overline{0, n-1})$, іншими словами кількість коефіцієнтів a_i^*,оцінених незалежно один від іншого, не повинно 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 у результаті перевірки адекватності отриманого рівняння регресії (4.8).

Статистична оцінка значущості коефіцієнтів рівняння регресії. Статистичну оцінку значущості коефіцієнтів $a_i^*; a_{i,l}^*$ проводять по t-критерію Стьюдента. Для кожного з коефіцієнтів обчислюють t-відношення

$$t_i = \frac{\left|a_i^* - a_i\right|}{S\{a_i^*\}} = \frac{\left|a_i^*\right|}{S\{a_i^*\}}, \tag{4.13}$$

де a_i – теоретичний генеральний коефіцієнт, який приймають дорівнюючим 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 (4.13) порівнюють з критичним (табличним) $t_{кр} = t_{табл}$, яке обирають для прийнятого рівня значущості q і числа супенів волі

$$f_в = N(\gamma - 1), \tag{4.14}$$

де γ – кількість паралельних дослідів.

Якщо виконується умова

$$t_i > t_{кр} = t_{табл}\,, \tag{4.15}$$

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 із рівняння регресії виключається.

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

$$F = \frac{S_{ад}^2\{y\}}{S_в^2\{y\}} \tag{4.16}$$

де $S_{ад}^2\{y\}$ – дисперсія неадекватності (іноді її називають дисперсією адекватності); $S_в^2\{y\}$ – 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:

$$\left.\begin{aligned} f_{числ} &= f_{ад} = f_1 = N - d \\ f_{знам} &= f_B = f_2 = N(\gamma - 1) \end{aligned}\right\} \tag{4.17}$$

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

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

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

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

Іноді може виявитись, що $F \le 1$. 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 $F_{кр}$, бо завжди $F_{кр}$ >1 за будь-яких кінцевих f_1, f_2.

Фізичний смисл перевірки адекватності складається у тому, що дисперсія адекватності $S^2_{ад}\{y\}$ не повинна значно перевищувати дисперсію відтворення $S^2_B\{y\}$, 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, по яким оцінюється адекватність рівняння регресії, збільшується на одну і стає N+1, тобто збільшується на 1 і кількість ступенів волі $f_{ад}$. Однак слід відмітити , що базова точка при ПФЕ не приймає участі у розрахунку коефіцієнтів рівняння.

Використання нормованого рівняння регресії для передбачення цільової функції. Рівняння регресії дозволяє передбачати математичне очікування цільової функції при даному 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 (4.8) можна перейти до рівняння у звичайному масштабі (4.1), через те що на практиці іноді зручніше використовувати рівняння, в яке підставляють фактори у реальних фізичних величинах. Перехід виконують за допомогою наступних співвідношень :

$$a_0 = a_0^* - \sum_{i=1}^{N} \frac{a_i^* x_{i,0}}{\lambda_i} + \sum_{i,l=1} a_{i,l}^* \frac{x_{i,0} x_{l,0}}{\lambda_i \lambda_l};$$
$$a_i = \frac{a_i^*}{\lambda_i} - 2\sum_{\substack{i,l=1 \\ i \neq l}}^{N} \frac{a_{i,l}^* x_{l,0}}{\lambda_i \lambda_l}; \quad a_{i,l} = \frac{a_{i,l}^*}{\lambda_i \lambda_l} \tag{4.19}$$

Якщо взаємодії відсутні, формули (4.19) суттєво спрощуються :

$$a_0 = a_0^* - \sum_{i=1}^{N} \frac{a_i^* x_{i,0}}{\lambda_i}; \quad a_i = \frac{a_i^*}{\lambda_i}; \quad a_{i,l} = 0 \tag{4.20}$$

Співвідношення (4.19), що дозволяють здійснити перехід від рівняння у нормованному вигляді (4.8) до рівняння у звичайному масштабі (4.1) , отримані підстановкою у (4.8) формули (4.3) .

Порядок виконання роботи

Реалізація ПФЕ у середовищі СППП Statgraphics Plus. Робота у середовищі СППП Statgraphics Plus складається з двох етапів:

а) побудова МП;

б) аналіз результатів експерименту.

Для виконання даної лабораторної роботи у середовищі СППП Statgraphics Plus спочатку потрібно із панелі меню обрати ***Special / Experimental Design /Create Design***. З'явиться вікно, що зображене на рис. 4.2., 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. 4.2.

Друге вікно, що з'явиться після натискання клавіші **ОК** (рис. 4.3) 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 x1, x2, x3, для кожного з факторів нижній рівень варіювання дорівнює –1, а верхній +1.

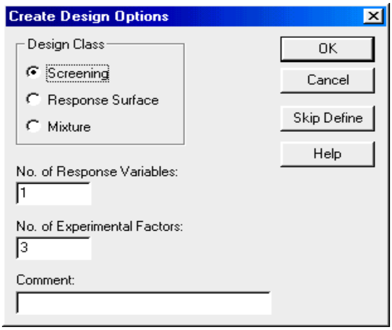


Рисунок 4.2 – Вікно вибору параметрів планування

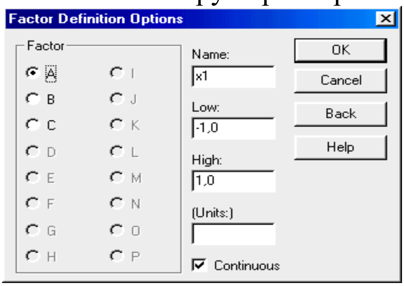


Рисунок 4.3 – Вікно визначення параметрів факторів

Вікно визначення параметрів відгуків (рис.4.4) дозволяє обрати назву змінної – відгуку. У нашому випадку назва Y.

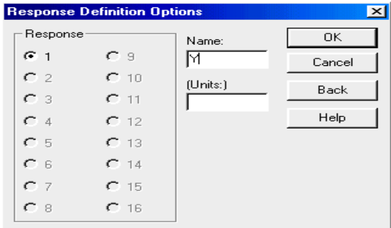


Рисунок 4.4 – Вікно визначення параметрів відгуку

Серед запропонованих МП (рис.4.5) обираємо план відповідний ПФЕ (2^3 , кількість дослідів 8)

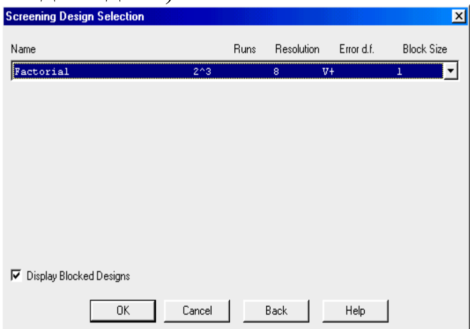
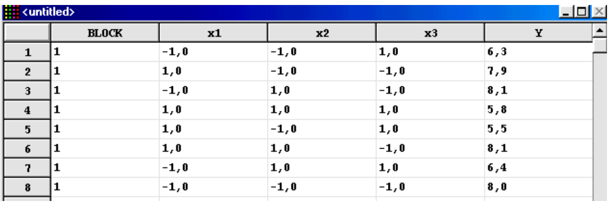


Рисунок 4.5 – Вікно вибору типу плану

В результаті цих дій отримуємо матрицю планування (рис.4.6) 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.


<untitled>

	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 4.6 – Вікно початкових даних для МП

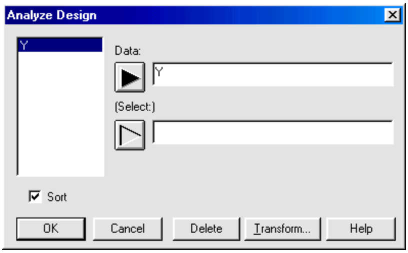


Рисунок 4.7 – Вікно вибору параметру оптимізації

Аналіз результатів починається з вибору параметру оптимізації (рис.4.7). Для наступного аналізу необхідна буде іконка  натиснувши на яку оберемо наступні пункти (рис.4.8)

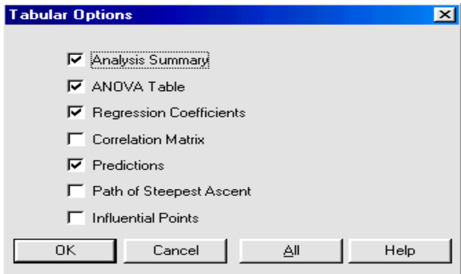


Рисунок 4.8 – Вікно вибору необхідних видів аналізу

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 4.9 – Вікно оцінки внесків факторів

Таблиця сумарного аналізу на рис. 4.9 показує кожний з оцінених внесків факторів, а також стандартну помилку кожного із внесків.

Наступна у списку таблиця дисперсійного аналізу (ANOVA Table, рис. 4.10) показує значущість кожного з факторів. Нагадаємо, що значимими вважаються фактори або їх взаємодії, якщо значення P-value менше 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.

Analysis of Variance for Y

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	0,0903
C:x3	8,20125	1	8,20125	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.)	8,76875	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
Durbin-Watson statistic = 2,0

Рисунок 4.10 – Таблиця дисперсійного аналізу

З таблиці коефіцієнтів регресії (рис. 4.11) ми бачимо підібрані для наших даних коефіцієнти рівняння регресії і саме рівняння регресії.

Regression coeffs. for Y

constant	= 7,0125
A:x1	= -0,1875
B:x2	= 0,0875
C:x3	= -1,0125
AB	= 0,0375
AC	= -0,1625
BC	= 0,0125

Y = 7,0125 - 0,1875*x1 + 0,0875*x2 - 1,0125*x3 + 0,0375*x1*x2 - 0,1625*x1*x3 + 0,0125*x2*x3

Рисунок 4.11 – Коефіцієнти рівняння регресії

Наступна таблиця прогнозувань (рис.4.12) дозволяє нам порівняти значення отриманих у результаті експерименту значеннь відгуку та значень відгуку для підібраної 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 4.12 – Порівняння експериментальних і отриманих даних

Завдання на лабораторну роботу

1. Ознайомитись із методикою ПФЕ.
2. Отримати індивідуальне завдання у викладача.
3. Згенерувати план експерименту 2^k, де

$$k = \begin{cases} Var, & Var \le 5 \\ Round(Var*2/5), & 5 < Var \le 15 \\ Round(Var/10), & Var > 15 \end{cases}.$$

4. Згенерувати дані експерименту. Для пакету Statgraphics використати функцію $Rnormal(N, \mu, \sigma^2)$.
5. Для кількості рівнобіжних дослідів 4, отримати матрицю планування, оцінки коефіцієнтів регресії та таблиці адекватності рівняння регресії в пакеті Statgraphics та написати відповідну програму з використанням мови R.
6. Зробити висновки.
7. Відповісти на контрольні питання.
8. Оформити звіт.

Зміст звіту

1. Мета та назва лабораторної роботи.
2. Початкові дані.
3. Матриця планування.
4. Оцінки коефіцієнтів регрессії та таблиця перевірки адекватності рівняння регресії.
5. Висновки до лабораторної роботи.

Контрольні питання

1. Дайте визначення теоріі планування експерименту.
2. Дайте визначення активного та пасивного експерименту.
3. Вимоги для вибору відгуків,факторів,координат базової точки та ступенів варіювання.
4. Призначення ПФЕ та етапи його проведення.
5. Що таке матриця планування. Як визначається кількість можливих комбінацій рівнів варіювання.
6. Властивості МП. Наведіть приклади кодування факторів.
7. З якою метою необхідна рандомізація дослідів.
8. Як впливає величина ступенів варіювання на значущість коефіцієнтів і на адекватність рівняння регресії.

ЛАБОРАТОРНА РОБОТА №5 ОРТОГОНАЛЬНЕ ЦЕНТРАЛЬНЕ КОМПОЗИЦІЙНЕ ПЛАНУВАННЯ

Мета роботи

Вивчити методику ортогонального центрального композиційного планування (ОЦКП) при обробці результатів активного експерименту та побудуванні нелінійних математичних моделей; побудувати математичні моделі другого порядку, оцінити їх адекватність, використовуючи для цього статистичний пакет Statgraphics та мову R.

Короткі теоретичні відомості

Центральне композиційне планування. Біля екстремуму поверхня відгуку має значну кривизну та не може бути адекватно описана неповним квадратичним рівнянням навіть у вузькій області. В цих випадках треба спробувати описати поверхню відгуку повним рівнянням другого ступеня. Для цього потрібно провести експеримент таким чином, щоб кожний фактор вар'їрував хоч би на трьох рівнях. Найпростішим рішенням можна назвати планування типа 3^n, проте повний факторний експеримент (ПФЕ) з числом рівней вар'їрування k=3 мав би більшу кількість рядків матриці планування (МП), який виражається формулою:

$$N = k^n = 3^n \tag{5.1}$$

Якщо ж використовувати ортогональні плани першого порядку (ПФЕ) в якості суті, на якій потім добудовується план другого порядку, то такі плани називаються центральні – композиційні (ЦКП). Потрапивши до області екстремуму, експериментатор реалізує ПФЕ, перевіряючи гіпотезу лінійної апроксимації. Пересвідчившись у неадекватності цієї гіпотези, він робить наступний крок, тобто добудовує ПФЕ до плану другого порядку, реалізує його і перевіряє гіпотезу об адекватності рівняння другого степеня.

Центральним планування називається тому, що передбачає постановку деякого числа N_0 експериментів у центрі планування, тобто в базовій точці, при нормованих рівнях $X_{i,0} = 0$.

Композиційним (або послідовним) таке планування називається тому, що воно складається з декілька частин, які здійснюються послідовно. Кожній частині відповідає група експериментальних точок.

Перша група – точки ПФЕ, в яких вже були поставлені досліди, але результати яких виявили неадекватну математичну модель. Експеримент у цих точках не потрібно ставити, використовується тільки його результати.

Друга група – зоряні точки, які лежать на усіх n вісях нормованого факторного простору по обидва боки від центру плану, тобто від базової точки B_0 на відстані – α (рис. 5.1).

Третя група – центральні точки з нормованим рівнем $X_{i,0} = 0$.

Загальна кількість точок ЦКП у факторному просторі:

$$N = N_\phi + N_\alpha + N_0 \tag{5.2}$$

N_ϕ – кількість точок факторного планування на двох рівнях (ПФЕ);

N_α – кількість зоряних точок;

N_0 – кількість точок у центрі плану.

Очевидно,

$$\left.\begin{aligned} N_\phi &= 2^n \\ N_\alpha &= 2 \cdot n \end{aligned}\right\} \tag{5.3}$$

Кількість центральних точок N_0 вибирають в залежності від вигляду ЦКП з метою забезпечення виконання прийнятого критерію оптимальності. Величину зоряного плеча α обирають так, щоб план залишався ортогональним або щоб задовольняв критерію ротатабельності. Таким чином – це планування на п'ятьох рівнях, які в нормованому вигляду можна представити так:

$$1/-\alpha\ ;\quad 2/-1;\quad 3/0;\quad 4/+1;\quad 5/+\alpha;$$

ЦКП буває двох видів: ортогональне (ОЦКП) другого та третього порядків та ротатабельне (РЦКП) другого та третього порядків. Види ЦКП відрізняються критерієм оптимальності.

Для ОЦКП другого порядку критерієм оптимальності є ортогональність усіх векторів – стовпців МП, який включає і вектор – стовпці для квадратичних членів. Для ортогонального ЦКП третього порядку МП складають так, щоб ортогональними з усімами іншими та один з одним були також і вектори – стовпці кубічних членів. Останнє ОЦКП складне і дуже рідко зустрічається.

Для ротатабельного (rotatable – здібний до обертання координат) ЦКП критерієм оптимальності є однакова точність прогнозування виходу рівняння регресії для будь-якої координати та взагалі для будь–якого напрямку у вивченій області факторного простору на рівній відстані від центру експерименту.

Отже, при плані, який має таку властивість, дисперсія прогнозованого значення відгуку $\hat{y}$ не змінюється при обертання плану навколо центра; звідси і виникає назва – ротатабельний план. Центральні композиційні плани при n=2 при n=3 наведені на рис. 5.1.

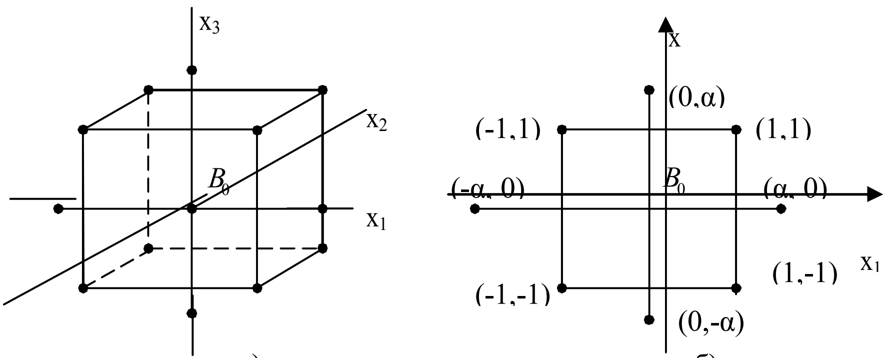


Рисунок 5.1 – Центральні композиційні плани при а) n=2 и n=3

Ротатабельність центрального композиційного плану забезпечується вибором α. Це значення α залежить від кількості точок у факторній частині плану.

Існує декілька інших планів, які іноді виявляються корисними в задачах з двома або трьома змінними.

Побудова матриці планування для ОЦКП. В ОЦКП другого порядку повинні бути ортогональні один з іншим усі векторі-стовпці, включаючи і векторі-стовпці квадратичних членів $X_i^2 (i = \overline{1,n})$ і нульового члена X_0. Ортогональність забезпечується вибором величини зоряного плеча α та кількості N_0 точок у центрі ОЦКП. Кількість точок факторного простору та розміри зоряних плеч при ОЦКП для різної кількості n факторів наведені у таблиці 5.1. Навіть мінімальна кількість центральних точок $N_0 = 1$ забезпечує ортогональність усіх векторів-стовпців, включаючи і квадратичні.

Таблиця 5.1 – Кількість точок факторного простору та розміри зоряних плеч при ОЦКП для різної кількості n факторів.

Кількість факторів	2	3	4
Кількість ПФЕ $N_ф$	2^2	2^3	2^4
Кількість зоряних точок N_α	4	6	8
Кількість центральних точок N_0	1	1	1
Загальне число точок N	9	15	25
Розмір плеча α	1.000	1.215	1.414

Величину $x_i(\alpha)$ в звичайному масштабі, потрібну для проведення реальних експериментів у зоряних точках, слід обчислювати за формулою:

$$x_i(\pm\alpha) = x_{i,0} \pm \alpha\lambda_i \tag{5.4}$$

де $x_{i,0}$ – базове значення фактора x_i.

Розглянемо приклад обчислення координат зоряних точок у реальному масштабі. Припустимо, центр ПФЕ, в результаті проведення якого була отримана неадекватна математична модель, був вибраний у точці трьохфакторного простору з координатами:

$$x_{1,0} = 40\,; \qquad x_{2,0} = 50\,, \qquad x_{3,0} = 60 \qquad (5.5)$$

а ступені варіювання вибрані в ПФЕ таким чином:

$$\lambda_1 = 15\,;\ \lambda_2 = 20\,; \qquad \lambda_3 = 10 \qquad (5.6)$$

Тоді координати зоряних точок будуть такі, як це показано у таблиці 5.2. Обчислення виконують за формулою (5.4):

$$x_1(-\alpha) = 40 - 1.215 \cdot 15 = 21.8 \approx 22\,; \qquad (5.7)$$

Таблиця 5.2 – Координати зоряних точок

Рівень	Нормована величина	Реальний масштаб		
		x_1	x_2	x_3
Базовий рівень $x_{i,0}$	0	40	50	60
Ступень варіювання λ_i	+1	15	20	10
Нижній рівень ПФЕ $x_{i,н}$	-1	25	30	50
Верхній рівень ПФЕ $x_{i,в}$	+1	55	70	70
Нижня зоряна точка ОЦКП -α	-1.215	22	26	48
Верхня зоряна точка ОЦКП +α	+1.215	58	74	72

Обчислення коефіцієнтів квадратичного рівняння. Статистична оцінка результатів ОЦКП. Для того, щоб зберегти ортогональність плану, потрібно при побудові матриці планування другого порядку забезпечити дві умови.

$$\sum_{k=1}^{N} x_{i,k} x_{j,k} = 0 \quad (i, j = \overline{1, n-1}) \qquad (5.8)$$

Перша умова – це умова ортогональності векторів-стовпців (5.8), якщо вона виконується, то матриця ЦКП ортогональна; та задовольняється шляхом перетворення квадратичних змінних.

$$\sum_{k=1}^{N} x_{i,k} = 0 \qquad (i = \overline{1,n}) \tag{5.9}$$

Друга умова – умова симетричності МП (5.9), що значить, що рівні, однакові по абсолютній величині, зустрічаються у будь-якому векторі-стовпці рівну кількість раз з різними знаками. Для векторів – стовпців з перетвореними квадратичними членами умова симетричності це значить, що сума значень рівней, які мають додатний знак, дорівнює по абсолютній величині сумі рівней з від’ємним знаком.

Забезпечивши ортогональність плану, таким чином одержуємо можливість визначити коефіцієнти регресії по дуже простій формулі та незалежно один від одного:

$$a_i^* = \frac{\sum_{j=1}^{k} X_{ij} \hat{y}_j}{\sum_{j=1}^{k} X_{ij}^2} \tag{5.10}$$

де і – номер стовпця в матриці планування;

X_{ij} – дорівнює +1 чи –1 у залежності від номеру j рядку МП;

$\hat{y}_j$ – значення цільової функції, отримане при значеннях факторів X_i, які містяться у j – тому рядку МП.

Рівняння регресії має такий вигляд:

$$\hat{y} = a_0^* + \sum_{i=1}^{k} a_i^* X_i + \sum_{i=1}^{k} a_{ii}^* X_i^2 + \sum_{\substack{i,l=1 \\ i \neq l}}^{k} a_{i,l}^* X_i X_l + \varepsilon \tag{5.11}$$

де $a_0^*, a_i^*, a_{i,l}^*$ – оцінки коефіцієнтів нормованого рівняння регресії;

ε – помилка одиничного прогнозування.

Дисперсії коефіцієнтів регресії оцінуються за формулою:

$$S^2_{a_i^*} = \frac{S_y^2}{\sum_{j=1}^{k} X_{ij}} \quad (5.12)$$

Залишкова сума квадратів

$$S_r = \sum_{j=1}^{k} \hat{y}_j^2 - \sum_{i=0}^{k1} a_i^* \sum_{j=1}^{k} X_{ij} \hat{y}_j \quad (5.13)$$

з числом ступеня волі $f_r = n - k_1$ де $k_1 = 0,5(k+2)(k+1)$

Методика обчислювання коефіцієнтів регресії для ротатабельних планів більш детально розглянута в літературі.

Вибір кількості паралельних дослідів, рандомізації порядку їх проводження, статистична оцінка відтворення дослідів здійснюється також як і при ПФЕ. Однак слід пам'ятати, що якщо вже був виконаний ПФЕ, то рандомізувати та проводити досліди треба тільки в зоряних та центральних точках. Але статистичну оцінку відтворення (або статистичну оцінку однорідності порядкових дисперсій) потрібно робити для усіх N порядкових дисперсій одразу ж, а не тільки для знову виконаних дослідів в зоряних та центральних точках.

Статистичну оцінку значущості одержаних коефіцієнтів рівняння регресії та перевірку адекватності квадратичного рівняння регресії, одержаного при ортогональному ЦКП, виконують також, як і при ПФЕ, однак з врахуванням числа ступеня волі, яке змінилося.

Якщо квадратичне рівняння виявилося адекватним з прийнятим рівнем значущості, то можна дослідити його з відомими аналітичними методами. Наприклад, можна з достатньою точністю знайти координати екстремальної точки, для чого слід узяти часткові похідні рівняння для усіх n факторів, прирівняти їх до нуля та вирішити одержану систему з n рівнянь.

Математичний опис області екстремуму можна використовувати для прогнозу цільової функції при заданих значеннях факторів, а також для оптимального керування процесом.

Аналіз моделей другого порядку. Вивчення поверхні відгуку. Аналіз поверхні другого порядку називається канонічним аналізом. Припустимо, що ми бажаємо знайти рівні $x_1, x_2 \ldots, x_k$, які максимізують прогнозований відгук. Стаціонарна точка може бути точкою максимуму відгуку, точкою мінімуму відгуку або седловою точкою, якщо вона існує, то описується таким набором координат $x_1, x_2\dots, x_k$ при умові, що часткові похідні $\partial\hat{y}/\partial x_1=\partial\hat{y}/\partial x_2=\dots=\partial\hat{y}/\partial x_k=0$. Ці три різновидності стаціонарних точок наведені на рис. 5.2.

Для стаціонарної точки можна одержати загальне рішення. Запишемо модель другого порядку в матричному виді.

$$\hat{y} = a_0 + x'a + x'Ax \tag{5.14}$$

де $x = (x_1, x_2 \dots x_k)$ – координати стаціонарної точки;

$a = (a_1, a_2 \dots a_k)$ – вектор коефіцієнтів регресії першого порядку;

A – симетрична матриця, на головній діагоналі якої стоять коефіцієнти при чисто квадратичних доданків a_{ii}, а недіагональні елементи дорівнюють половинам відповідних коефіцієнтів при змішаних добутках ($a_{ij}, i \neq j$).

Дорівнюючи 0 похідну $\hat{y}$ по вектору x, одержимо

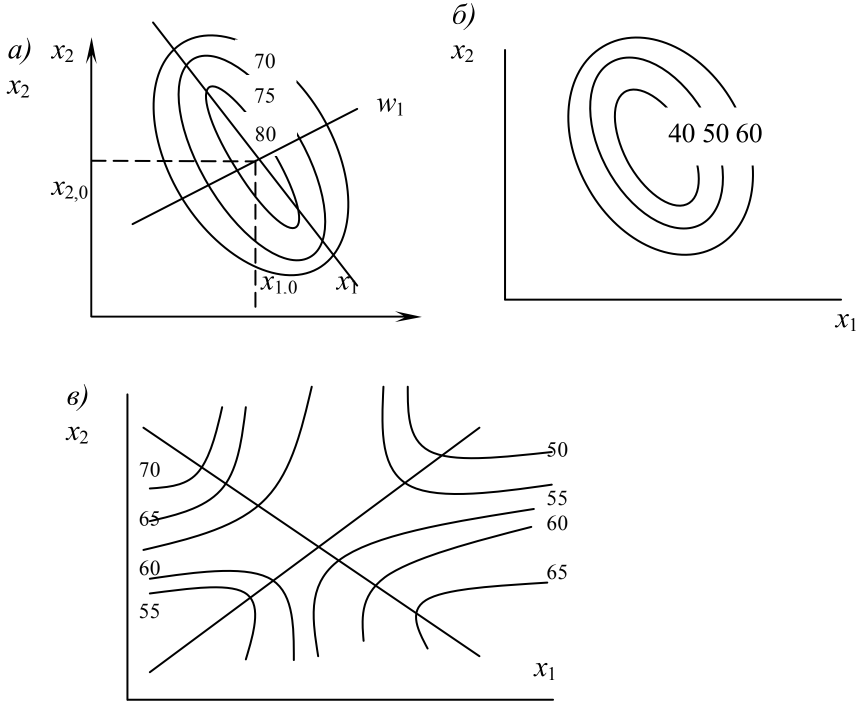
$$\frac{\partial \hat{y}}{\partial x} = a + 2Ax = 0 \tag{5.15}$$

Рішенням цього рівняння і є стаціонарна точка

$$x_0 = -\frac{1}{2}A^{-1}a \tag{5.16}$$

Крім того, підставивши одержане рішення у відношення (5.14), можна знайти прогнозоване значення відгуку в стаціонарній точці

$$\hat{y}_0 = a_0 + \frac{1}{2}x_0' a \tag{5.17}$$



а – максимум відгуку ;б – мінімум відгуку; в – седлова точка

Рисунок 5.2 – Стаціонарні точки поверхні відгуку другого порядку

Для визначення характеру стаціонарної точки корисно перетворити модель до нової системі координат з початком у точці x_0, а потім повернути вісі цієї системи так, щоб вони співпали з головними вісями підібраної поверхні відгуку. Можна показати, що при цьому підібрана модель матиме вид

$$\hat{y} = \hat{y}_0 + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \ldots + \lambda_k w_k^2, \qquad (5.18)$$

де w_i – перетворені незалежні змінні, а λ_i – постійні. Вираз (5.18) називається канонічним рівнянням моделі, λ_i є власними значеннями або характеристичними коренями матриці А.

По канонічному вигляду рівняння другого ступеня можливо класифікувати екстремальну точку та поверхню відгуку. Розглянемо задачу з двома факторами. Надаючи величині відгуку деякі фіксовані значення, отримують контурні криві рівного виходу.

На рисунках 5.2 – 5.3 наведені чотири можливих види контурних кривих.

Гіперболи. Характер поверхні відгуку можна визначити за стаціонарною точкою, знаку і величині λ_i. Припустимо спочатку, що стаціонарна точка знаходиться в області експерименту, яка досліджується, при підборі моделі другого порядку. Якщо ж знаки λ_i різні, то x_0 – центр фігури, який називається мінімаксом або седловою точкою. Поверхня відгуку уявляє собою гіперболіческий параболоїд (рис. 5.2, в).

Вивчення поверхні відгуку при віддалі за межі області, де виконувалися експерименти, зводиться до задачі пошуку умовного оптимуму при обмеженнях, які накладаються різними умовами. Поверхня відгуку, яка уявляє собою неповне квадратне рівняння завжди є гіперболіческим параболоїдом.

Еліпси. Якщо усі λ_i додатні, то x_0 – точка мінімуму відгуку , якщо усі λ_i від'ємні, то x_0 – точка максимуму відгуку. Крім цього, поверхня оказується найбільш крутою у напрямку w_i, для якого $|\lambda_i|$ найбільший(рис. 5.2, а,б).

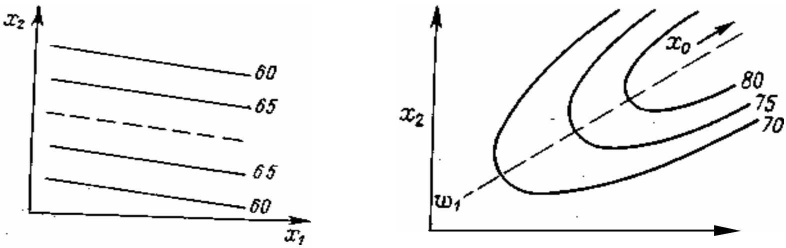
Наприклад, на рис. 5.2,а зображена система, для якої x_0—точка максимуму (λ_1 и λ_2 від'ємні), причому $|\lambda_1|>|\lambda_2|$. Якщо центр фігури знаходиться поблизу або в області експерименту, то вивчення поверхні відгуку закінчується канонічним перетворенням. Поверхня відгуку уявляє собою еліптичний параболоїд.

Параболи. Якщо стаціонарна точка знаходиться далеко за межами області, яка досліджується при підборі поверхні другого порядку, а одно (або більш) λ_i близько до нуля, то поверхня може бути зростаючим гребенем (на рис. 5.3, б наведений зростаючий гребень для k=2 змінних при λ_1 близьким до нуля та від'ємним λ_2).

Паралельні прямі. Коли одно або декілька λ_i дуже малі (наприклад, $\lambda_i \approx 0$), то система оказується невизначеною для змінної

w_i. Поверхня такого типу часто називається стаціонарним гребенем (рис. 5.3, а). Поверхня відгуку уявляє собою стаціонарне підвищення.

В системах з гребенем такого типу ми не можемо сказати нічого визначеного об істинній поверхні або стаціонарній точці, остільки x_0 не належить області експерименту, яка досліджується при підборі моделі. Однак розумно продовжити дослідження в напрямку w_i. Якщо λ_2 додатне, то ми назвали б систему низхідним гребенем. При кількості факторів $n > 2$ вивчення поверхні відгуку виконується таким же чином. Але на відміну від поверхонь при $n = 2$, які можна вивчати графічно, поверхні при $n > 2$, в загальному випадку вивчаються аналітично.



а – система з стаціонарним гребенем ;б – система з зростаючим гребенем

Рисунок 5.3 – Поверхні відгуку другого порядку

Порядок виконання роботи

Проілюструємо виконання ОЦКП в пакеті Statgraphics на основі наступного прикладу.

Постановка задачі. Проведемо дослідження операції створення внутрішніх з'єднань в мікросхемі за допомогою ультразвукового зварювання. В якості керуючих факторів були вибрані основні технологічні параметри процесу ультразвукового зварювання:

– тривалість зварювального імпульсу (фактор А),

– тиск зварювального інструменту (фактор В),

– амплітуда коливань зварювального інструменту (фактор С).

В якості параметра оптимізації (відгуку Var_1) приймаємо значення вихідної змінної – міцність зварювального з'єднання, яка залежить від трьох вище вказаних факторів.

Оптимізація процесу ультразвукового зварювання була почата з отримання математичної моделі траверса-траверса. Для цього на першому етапі був реалізований ПФЕ типу 2^3, в результаті якого було одержане рівняння , яке не має квадратичних коефіцієнтів і тому за його допомогою неможливо описати екстремальну область, в якій кривизна поверхні відгуку суттєва. Також були отримані дані о відсутності впливу деяких факторів на вихідну величину, хоч експериментатори вважали, що саме ці фактори впливають на відгук. Для продовження моделювання та одержання більш адекватної математичної моделі було використано ОЦКП другого порядку.

Підготовка даних. Побудова матриці планування.

Для побудови МП необхідно вибрати команду Special ▸ Experimental Design ▸ Create design у рядку меню. З'явиться діалогове вікно, у яке заносяться необхідні дані (рис. 5.4): вибираємо тип планування (у нашому випадку це опція Response Surface – дослідження поверхні відгуку), задаємо кількість впливаючих факторів на відгук (опція No. of Experimental Factors) та кількість змінних відгуку (опція No. of Response Variables).

Наступна частина побудови МП подібна виконанню аналізу ПФЕ, при якому здійснюється вибір рівнів варіювання для кожного фактора (опція Factor Definition Options), завдання параметрів змінної відгуку: назви, одиниць вимірювання (опція Response Definition Options). Після цих дій з'явиться вікно, в якому ми вибираємо тип плану – центральне композиційне планування (рис. 5.5).

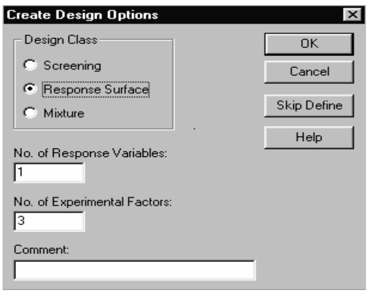


Рисунок 5.4 –Вибір типу планування, кількості факторів і змінних відгуку

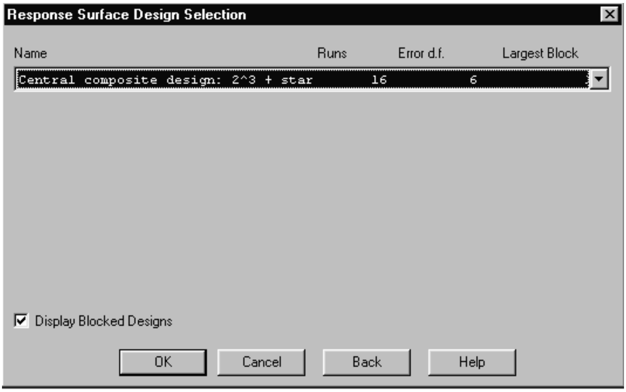


Рисунок 5.5 –Вибір типу плану

Після вибору типу плану потрібно вказати вид плану ротатабельний (Rotatable) чи ортогональний (Orthogonal) або ротатабельний і ортогональний (Rotatable and Orthogonal), розмір плеча (Axial Distance), який визначається автоматично і залежить від виду плану, кількість центральних точок (Centerpoints ▸ Number) та місця їх розташування в матриці планування: випадкове (Random), перед (First), після (Last) зоряних точок (рис. 5.6). В результаті був побудований план експерименту (рис. 5.7).

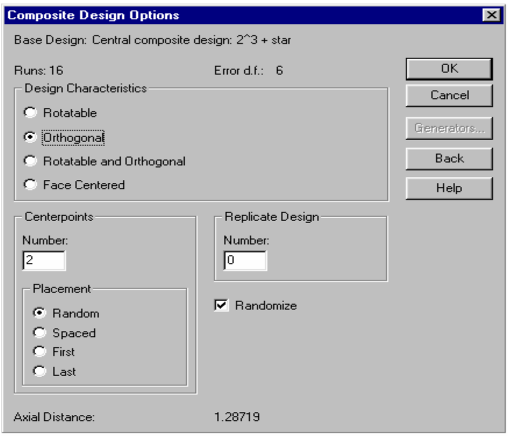


Рисунок 5.6 – Вибір виду плану, розміру плеча, кількості центральних точок та місця їх розташування

```
Design Summary
--------------
Design class: Response Surface
Design name: Central composite design: 2^3 + star
Design characteristic: Orthogonal
File name: D:\Ìîè äîêóìåíòû\METODS\Ìåòîäè÷êè\lab71.sfx

Base Design
-----------
Number of experimental factors: 3  Number of blocks: 1
Number of responses: 1             Number of centerpoints per block: 2
Number of runs: 16                 Error degrees of freedom: 6
Randomized: No

Factors       Low         High        Units       Continuous
-------------------------------------------------------------
Factor_A      -1,0        1,0                     Yes
Factor_B      -1,0        1,0                     Yes
Factor_C      -1,0        1,0                     Yes

Responses     Units
-------------------------
Var_1
```

Рисунок 5.7 – План експерименту

В стовпець змінної відгуку Var_1 (рис. 5.8) заносимо експериментальні значення.

	BLOCK	Factor_A	Factor_B	Factor_C	Var_1
1	1	0	0	0	8,18
2	1	0	0	0	8,18
3	1	-1	-1	-1	7,08
4	1	1	-1	-1	7,57
5	1	-1	1	-1	6,42
6	1	1	1	-1	6,91
7	1	-1	-1	1	8,19
8	1	1	-1	1	8,21
9	1	-1	1	1	7,88
10	1	1	1	1	7,9
11	1	-1,28719	0	0	7,88
12	1	1,28719	0	0	8,19
13	1	0	-1,28719	0	8,08
14	1	0	1,28719	0	7,44
15	1	0	0	-1,28719	7,14
16	1	0	0	1,28719	7,78

Рисунок 5.8 – Вихідна матриця планування експерименту

Виконання аналізу. Після того, як були введені вихідні дані (рис. 5.8), переходимо до аналізу результатів. Для цього необхідно вибрати команду Special ▸Experimental Design ▸Analyze design у рядку меню. З’явиться діалогове вікно, у яке заноситься назва змінної відгуку – Var_1 (рис. 5.9).

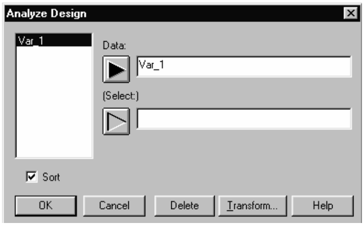


Рисунок 5.9 – Вибір назви змінної відгуку

Для того, щоб побачити результати проведеного аналізу: вихідні таблиці та графіки залежностей, потрібно натиснути правою кнопкою миші на наступних пунктах панелі інструментів: на опції таблиць Tabular Options і опції графіків Graphical Options.

При появі діалогового вікна Tabular Options (рис.5.10) вибираються чотири пункти: Analysis Summary (загальна таблиця аналізу), ANOVA Table (таблиця дисперсійного аналізу), Regression Coefficients (таблиця коефіцієнтів регресії), Predictions (таблиця визначення відповідності побудованої математичної моделі реальній).

При натисненні на пункт Graphical Options (рис.5.11), з'явиться вікно, в якому вибираються чотири пункти: Main Effects Plots (Графік впливу факторів), Interaction Plots (Графік сумісного впливу факторів), Response Plots (Графік поверхні відгуку в трьохвимірному просторі), Response Plots (Контурні криві поверхні відгуку).

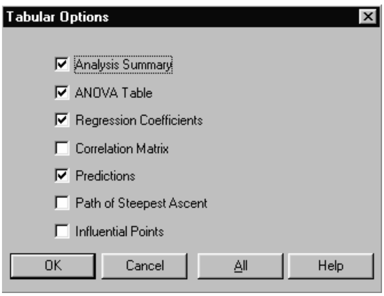


Рисунок 5.10 –Вікно таблиць -Tabular Options

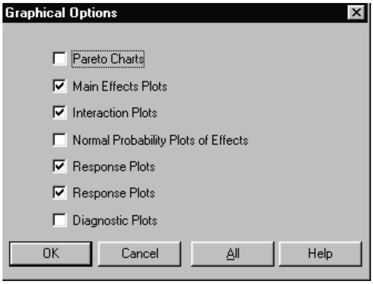


Рисунок 5.11 –Вікно графіків -Graphical Options

В результаті проведеного експерименту були отримані дані, на які аналітично дивлячись можна побачити вплив факторів на відгук та оцінити адекватність одержаної математичної моделі (рис. 5.12). Таблиця дисперсійного аналізу показує, що чотири фактора мають рівень значущості (P–Value) менший за 0.05, що свідчить о впливі або сумісного впливу цих факторів на змінну відгуку, яка досліджується. Критерій R–squared показує адекватність отриманої в результаті експерименту математичної моделі, чим він вище, тим більш підібрана модель адекватно описує істинну поверхню.

За результатами дослідів були отримані такі коефіцієнти рівняння регресії і саме рівняння регресії (рис. 5.13):

$$Var\_1 = 8.13 + 0.125 * Factor\_A - 0.24 * Factor\_B + 0.44 * Factor\_C -$$
$$- 0.125 * Factor\_A * Factor\_B - 0.20 * Factor\_B\text{^}2 - 0.39 * Factor\_C\text{^}2$$

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio
A:Factor_A	0,177982	1	0,177982	5,55
B:Factor_B	0,675163	1	0,675163	21,04
C:Factor_C	2,2308	1	2,2308	69,51
AA	0,0080096	1	0,0080096	0,25
AB	0,0	1	0,0	0,00
AC	0,11045	1	0,11045	3,44
BB	0,228871	1	0,228871	7,13
BC	0,06125	1	0,06125	1,91
CC	0,81481	1	0,81481	25,39
Total error	0,192565	6	0,0320942	
Total (corr.)	4,49989	15		

R-squared = 95,7207 percent
R-squared (adjusted for d.f.) = 89,3017 percent
Standard Error of Est. = 0,179148
Mean absolute error = 0,0864976
Durbin-Watson statistic = 1,94005

Рисунок 5.12 –Таблиця дисперсійного аналізу

Основною задачею планування було визначення оптимальних значень параметрів змінної відгуку, а саме це був пошук максимального значення міцності виводів у мікросхемі, тобто при яких значеннях параметрів одержується максимум функції відгуку (Var_1).

Для цього потрібно знайти умовний екстремум змінної відгуку при обмеженнях, які накладені різноманітними технологічними умовами. Отримаємо такі значення параметрів: $Factor\_A_{опт} = 1.0$; $Factor\_B_{опт} = -0.60$; $Factor\_C_{опт} = -2.23$.

Після переходу від нормованих значень до змінних у реальному масштабі отримаємо: $Factor\_A_{опт} = 23.84 мс$; $Factor\_B_{опт} = 43.84 г$; $Factor\_C_{опт} = 0.86 мкм.$

Підставивши отримані значення оптимальних параметрів у рівняння регресії, отримаємо розрахункове значення максимальної міцності виводів, яке дорівнює 8.33.

```
Regression coeffs. for Var_1
----------------------------------------------------------------------
constant     = 8,13316
A:Factor_A   = 0,125426
B:Factor_B   = -0,244288
C:Factor_C   = 0,444045
AA           = -0,0381949
AB           = 0,0
AC           = -0,1175
BB           = -0,204172
BC           = 0,0875
CC           = -0,385237
----------------------------------------------------------------------
```

Рисунок 4.13 – Коефіцієнти рівняння регресії

Використовуючи дані експерименту було одержано контурні криві поверхні відгуку (рис. 4.14).

Аналіз графіків показує, що поверхня відгуку уявляє собою гіперболічний параболоїд. На графіку контурних кривих лініями відображені значення відгуку при відповідних нормованих значеннях факторів A і B.

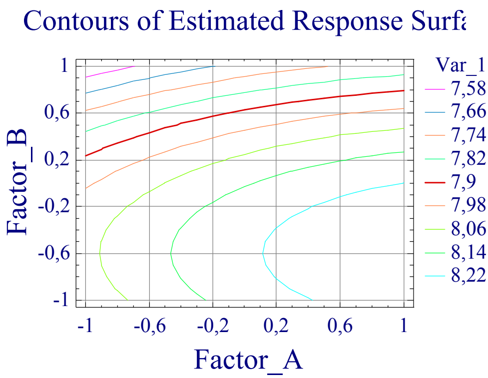


Рисунок 5.14 –Графік контурних кривих поверхні відгуку

Завдання на лабораторну роботу

1. Ознайомитись із методикою ОЦКП.
2. Отримати індивідуальне завдання у викладача.
3. Згенерувати план експерименту 2^k, де

$$k = \begin{cases} Var, & Var \le 5 \\ Round(Var * 2/5), & 5 < Var \le 15 \\ Round(Var/10), & Var > 15 \end{cases} .$$

4. Згенерувати дані експерименту. Для пакету Statgraphics використати функцію $Rnormal(N, \mu, \sigma^2)$.

5. Розрахувати кількість зоряних точок, кількість центральних точок, розмір плеча. Для кількості рівнобіжних дослідів 4, отримати матрицю планування, оцінки коефіцієнтів регресії та таблиці адекватності рівняння регресії в пакеті Statgraphics та написати відповідну програму з використанням мови R.

6. Зробити висновки.
7. Відповісти на контрольні питання.
8. Оформити звіт.

Зміст звіту

1. Назва і мета роботи.
2. Постановка задачі.
3. Таблиця вихідних даних.

4. Основні формули розрахунку параметрів.
5. Результати виконання ОЦКП з коментаріями.
6. Висновки по результатам.

Контрольні питання

1. Коли використовується ОЦКП?
2. З яких частин складається ЦКП?
3. Назвіть види ЦКП. Чим відрізняються види ЦКП?
4. Розкажіть о зоряних точках при ЦКП.
5. Що є критерієм оптимальності для ОЦКП?
6. Що є критерієм оптимальності для РЦКП?
7. Яким чином в ОЦКП забезпечується ортогональність векторів-стовпців, включаючи квадратичні вектори-стовпці?
8. Яким чином, знаючи величину зоряного плеча в нормованому масштабі, можна перейти до звичайного масштабу?
9. Як будується МП при ОЦКП?
10. Методика обчислювання формул ОЦКП.
11. Як виконується аналіз моделей другого порядку? Поняття стаціонарної точки.

ЛІТЕРАТУРА

1. Алексахина С.В. Прикладной статистический анализ данных. Теория. Компьютерная обработка. Области применения / С.В. Алексахина, А.В. Балдина, А.Б. Николаева, В.Ю. Строганова.– М.: ПРИОР, 2002. – 688 с.
2. Тюрин Ю. Н. Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров; под ред. В. И. Фигурнова. – М. : ИНФА-М., 2003. – 544 с.
3. Томашевський О. В. Комп'ютерні технології статистичної обробки даних: навч. посібник для студ. вищ. навч. закл. / О. В. Томашевський, В. П. Рисіков.–Запоріжжя: ЗНТУ, 2006. – 174 с.
4. Мастицкий С.Э. Статистический анализ и визуализация данных с помощью R / Мастицкий С.Э., Шитиков В.К. – М.: ДМК Пресс., 2014. – 401 с.
5. Gareth James An Introduction to Statistical Learning with Applications in R / Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. – Springer, New York, 2013. – 426 p.
6. Trevor Hastie The Elements of Statistical Learning Data Mining, Inference, and Prediction /Trevor Hastie Robert Tibshirani Jerome Friedman. – Springer, New York, 2013. – 745 p.
7. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL/ Э.А. Вуколов. – М.: ФОРУМ., 2008. – 464 с.
8. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. / А.И. Кобзарь. – М.: Физматлит, 2006. – 816 с.
9. Берк К. Анализ данных с помощью Microsoft Excel. / К.Берк, П. Кэйри. – М.: Вильямс, 2005. – 560 с.
10. Шипунов А.Б. Наглядная статистика. Используем R! / А.Б. Шипунов, Е.М. Балдин, П.А. Волкова, А.И. Коробейников, С.А.Назарова, С.В. Петров, В.Г. Суфиянов. – М.: ДМК Пресс, 2012. – 298 с.

ДОДАТОК А
ФОРМУЛИ ДЛЯ ВИКОНАННЯ ЛАБОРАТОРНИХ РОБІТ

1. Формула Стерджеса:

$$k = 1 + 3{,}322 \lg N \tag{А.1}$$

де N – обсяг вибірки.
2. Розмах R:

$$R = x_{\max} - x_{\min}, \tag{А.2}$$

де $x_{\max}$ – максимальне значення вибірки; $x_{\min}$ – мінімальне значення вибірки;
3. Математичне сподівання $M(x)$:

$$M(x) = \frac{1}{N} \sum_{i=1}^{N} x_i , \tag{А.3}$$

де x_i – елементи вибірки, i=1..N.
4. Середнє значення $\bar{x}$:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i , \tag{А.4}$$

де x_i – значення, які спостерігаються, відповідних вхідних даних в i-му досліді.
5. Середнє значення $\bar{y}$:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i , \tag{А.5}$$

де y_i – значення, які спостерігаються, відповідних вихідних даних в i-му досліді.

6. Незміщена дисперсія $S^2(x)$:

$$S^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 , \qquad (А.6)$$

7. Середньоквадратичне відхилення $S(x)$:

$$S(x) = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}{N-1}} , \qquad (А.7)$$

8. Табличне значення коефіцієнта Стьюдента:

$$t_{кр}(q, f), \qquad (А.8)$$

де q – рівень значимості, f=N-1– кількість ступенів волі.

9. Довірчий інтервал, розрахований за допомогою критерію Стьюдента:

$$\delta = \frac{1}{\sqrt{N}} t_{кр} S\{x\}, \qquad (А.9)$$

10. Критерій Кохрена:

$$G = \frac{S_j^{\,2}(x)\max}{\sum_{j=1}^{N} S_j^{\,2}(x)}, \qquad (А.10)$$

де $S_j^2(x)\max$ – значення дисперсії вибірки, яке є максимальним;

$\sum_{i=1}^{N} S_j^{\,2}(x)$ – сума всіх дисперсій.

11. Табличне значення критерію Кохрена:

$$G_{кр}\left(q, f_1, f_2\right), \tag{A.11}$$

де $f_1 = N_1 - 1,$ де N_1 – кількість екземплярів вибірок;

$f_2 = f_{знач}$ – це кількість дисперсій.

12. Табличне значення критерію Фішера:

$$F_{кр}\left(q, f_1, f_2\right), \tag{A.12}$$

де $f_1 = N_1 - 1,$ де N_1 – кількість екземплярів першої вибірки;

$f_2 = N_2 - 1,$ де N_2 – кількість екземплярів другої вибірки.

13. Критерій Фішера:

$$F = \frac{S_1{}^2(x)}{S_2{}^2(x)}, \tag{A.13}$$

де $S_1{}^2(x)$ – це значення першої дисперсії вибірки x;

$S_2{}^2(x)$ – це значення другої дисперсії вибірки x.

14. Коефіцієнт кореляції:

$$r_{xy} = \frac{\sum\limits_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)S\{x\}S\{y\}}, \tag{A.14}$$

15. Нульовий коефіцієнт регресії:

$$a_0 = \overline{y} - a_1\overline{x}, \tag{A.15}$$

16. Перший коефіцієнт регресії:

$$a_1 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}, \tag{A.16}$$

17. Множинний коефіцієнт кореляції:

$$R^2 = \frac{\sum\left(\hat{y}_i - \overline{y}\right)^2}{\sum\left(y_i - \overline{y}\right)^2}, \tag{A.17}$$

де $\overline{y}$ – загальне середнє відгуків; $\hat{y}_i$ – відхилення значення відгуку, що спостерігається, від загального середнього відгуків $\overline{y}$.

18. Критерій Фішера:

$$F = \frac{S_1}{S_2} = \frac{(N-2)\sum_{i=1}^{N}\left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \hat{y}\right)^2}, \tag{A.18}$$

19. Коефіцієнт a_0:

$$a_0 = \frac{\sum_{n=1}^{N} y^{експ}}{N}, \tag{A.19}$$

де $y^{експ}$ – це значення у, отримані в результаті експерименту.

20. Коефіцієнти a_1, a_2, a_3:

$$a_1 = \frac{\sum_{n=1}^{N} x_{1n} y_n^{експ}}{N}, a_2 = \frac{\sum_{n=1}^{N} x_{2n} y_n^{експ}}{N}, \; a_3 = \frac{\sum_{n=1}^{N} x_{1n} x_{2n} y_n^{експ}}{N} \tag{A.20}$$

21. Кількість точок ЦКП у факторному просторі:

$$N = N_{\phi} + N_{\alpha} + N_0, \quad \text{(A.21)}$$

де N_{ϕ} – кількість точок факторного планування на двох рівнях (ПФЕ); N_{α} – кількість зоряних точок; N_0 – кількість точок у центрі плану.

$$\left.\begin{aligned} N_{\phi} &= 2^n \\ N_{\alpha} &= 2n \end{aligned}\right\} \quad \text{(A.22)}$$

де n – кількість факторів.

ДОДАТОК Б ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

X1	Y1	X2	Y2	X3	Y3	X4	Y4	X5	Y5
460	0,300	4,500	619	3,000	3,100	4,000	27	0,971	3,000
450	0,300	4,500	1049	3,100	3,900	3,000	54	0,979	4,700
440	0,400	4,500	1033	3,000	3,400	5,000	86	0,982	8,300
430	0,400	4,000	495	3,600	4,000	8,000	136	0,971	9,300
420	0,600	4,000	723	3,800	3,600	4,000	65	0,957	9,900
410	0,500	4,000	681	2,700	3,600	3,000	109	0,961	11,0
450	0,500	5,000	890	3,100	3,100	3,000	28	0,956	12,3
440	0,600	5,000	1522	2,700	3,600	4,000	75	0,972	12,5
430	0,600	5,500	987	2,700	2,900	3,000	53	0,889	12,6
420	0,600	5,000	1194	3,300	3,600	5,000	33	0,961	15,9
410	0,700	0,500	163	3,200	4,100	7,000	168	0,982	16,7
400	0,600	0,500	182	2,100	2,600	3,000	47	0,975	18,8
420	0,600	6,000	764	3,000	3,100	8,000	52	0,942	18,8
400	0,600	6,000	1373	2,600	2,800			0,932	18,9
410	0,600	1,000	978					0,908	21,7
		1,000	466					0,970	21,9
		1,000	549					0,985	22,8
								0,933	24,2
								0,858	25,8
								0,987	30,6

X6	**Y6**	**X7**	**Y7**	**X8**	**Y8**	**X9**	**Y9**
0,958	36,2	100	-2	13,3	3,500	50	19
0,909	39,8	10	-4	16,9	5,100	60	20,21
0,859	14,3	20	1	19,9	4,800	70	17,22
0,863	46,8	20	3	23,2	6,700	80	25,28
0,811	46,8	30	2	26,3	6,000	90	21,25
0,877	58,1	30	5	30,1	9,500	95	25
0,798	62,3	40	0	42,6	8,100	100	30,29
0,855	70,6	40	1			105	35,32
0,788	71,1	40	2			110	30,28
0,821	71,3	50	-2			115	31,36
0,830	83,2	50	-3			120	36,25
0,718	83,6	50	-4				
0,642	99,5						
0,658	111,2						