ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«ПРИКАРПАТСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
імені ВАСИЛЯ СТЕФАНИКА»

Кафедра інформаційних технологій

Гарпуль О.З., Незамай Б.С.

ЛАБОРАТОРНИЙ ПРАКТИКУМ

«Емпіричні методи програмної інженерії»

Івано-Франківськ

2018

УДК 004.42: 519.25
ББК 32.973.2 – 018
Г – 105

Рекомендовано до друку вченою радою факультету математики та інформатики ДВНЗ «Прикарпатський національний університет імені Василя Стефаника»
(протокол № 10 від 27.06. 2018 р.)

Рецензенти:

Когут І.Т. – доктор технічних наук, завідувач кафедри комп'ютерної інженерії та електроніки фізико-технічного факультету ДВНЗ «Прикарпатський національний університет імені Василя Стефаника»;

Власій О.О. – кандидат технічних наук, доцент кафедри інформатики факультету математики та інформатики ДВНЗ «Прикарпатський національний університет імені Василя Стефаника».

Гарпуль О.З., Незамай Б.С.
Г – 105 Лабораторний практикум «Емпіричні методи програмної інженерії» / О.З. Гарпуль, Б.С. Незамай. – Івано-Франківськ: Видавництво Прикарпатського національного університету, 2018. – 105 с.

Лабораторний практикум «Емпіричні методи програмної інженерії» містить цикл лабораторних робіт, в яких коротко описані найбільш важливі методи обробки даних, основні технічні прийоми, необхідні при роботі та приклади розв'язку задач, а також методичні рекомендації щодо виконання самостійної роботи та програмові вимоги до іспиту.

Лабораторний практикум розрахований для студентів технічних спеціальностей, зокрема спеціальності «Інженерія програмного забезпечення» вищих навчальних закладів III-IV рівнів акредитації і містить методичний матеріал про основні методи обробки емпіричних даних за допомогою математичної статистики у середовищі програмування.

Даний практикум допоможе студентам засвоїти основні методи при обробці емпіричних даних та дозволить легко і швидко знайти відповідь на вирішення конкретної проблеми.

УДК 004.42: 519.25
ББК 32.973.2 – 018

ЗМІСТ

ВСТУП

Початковим етапом пізнання людиною навколишнього світу завжди було і лишається живе спостереження, яких би складних і технічних форм воно не набувало. Кінцевим результатом процесу пізнання є формулювання законів, вираженням яких врешті є певні зв'язки між явищами та процесами, які описуються математичними співвідношеннями між величинами, що характеризують спостережувані об'єкти. В ідеальному випадку вдається отримати функціональну залежність між спостережуваними величинами, але зазвичай залежності спотворюються величезною кількістю випадкових впливів, наявністю взаємозв'язків між факторами, нестаціонарністю їх (тобто змінюваністю характеристик у часі) тощо. В результаті залежності набувають випадкового характеру, що робить надзвичайно важким їх виявлення і, тим більше, математичний опис. Для кожного окремого випадку у певний момент чи проміжок часу залежність виступає як випадковий багатовимірний розподіл кількох величин. Виявлення залежностей за цих умов можливе тільки на основі масових досліджень, коли ефекти усереднення дозволяють виявити залежність і побудувати її математичну модель. Величини, що виступають як незалежні називаються факторами чи факторними ознаками, а величина, яка виступає як залежна змінна – результативним фактором. Тому емпіричні дослідження використовуються для відповіді на емпіричні питання, які повинні бути точно визначені згідно з даними.

Предметом вивчення є засоби і методи прикладної статистики, науки про те, як обробляти дані. Методи прикладної статистики активно застосовуються в технічних дослідженнях економіці теорії і практиці управління. З результатами спостережень, вимірювань, випробувань, експериментів, з їх аналізом мають справу фахівці у всіх галузях практичної діяльності, майже в усіх областях теоретичних досліджень. Завданням курсу є вивчення сучасних методів прикладної статистики на рівні, достатньому для використання цих методів у науковій та практичній діяльності, оволодіння методами побудови математичних моделей з використанням статистичних методів, розвиток логічного й алгоритмічного мислення студентів. Математична статистика - це сучасна галузь математичної науки, яка займається статистичним описом результатів експериментів і спостережень, а також побудовою математичних моделей, що містять поняття ймовірності. Теоретичною базою математичної статистики служить теорія ймовірностей.

Метою практикуму є засвоєння принципів застосування емпіричних методів у галузі програмної інженерії, формування практичних навичок при обробці великої кількості експериментальних даних і розв'язку задач оптимізації для багатомірних об'єктів. Потрібно показати також, що розвиток технічних програмних засобів обчислювальної техніки дає можливість говорити про нову

концепцію в організації наукових досліджень - автоматизації експерименту. Розглядаються основи описової статистики, дискретні і безперервні поділи ймовірностей, методи оцінювання параметрів регресійніх залежностей, кореляції, статистичні тести, які найчастіше вживаються в галузі програмної інженерії, а також методи планування експерименту і перевірка гіпотез, застосування емпіричних методів для аналізу продуктивності і надійності програмних систем тощо.

Виконання лабораторних робіт даного практикуму спрямовано на поглиблення засвоєння лекційного і опрацьованого самостійно матеріалу і набуття студентами практичних навичок обробки емпіричних даних. На лабораторних заняттях студенти розв'язують задачі щодо обробки даних з використанням загальних можливостей і базових функцій електронних таблиць Excel і вбудованого в Excel статистичного пакету "Аналіз даних", а також розробляють алгоритми і програми обробки даних статистичними методами, реалізовуючи їх у програмних середовищах.

Основними завданнями циклу лабораторних робіт є набуття студентами знань та умінь:

знати:

- основні характеристики випадкових величин, закони розподілу;
- параметри функції розподілу;
- методи кореляційного і регресійного аналізу;
- методи планування експерименту і перевірка гіпотез;
- застосування емпіричних методів для аналізу продуктивності і надійності програмних систем;
- методи аналізу і підвищення надійності елементів і систем на етапах проектування;

вміти:

- визначати характеристики надійності елементів та виробів;
- визначати кількісні оцінки ступеня ризику на виробництві;
- розробляти алгоритм рішення задачі оптимізації об'єкта дослідження;
- по заданому алгоритму складати програму;
- налагоджувати і виконувати програму на ЕОМ;

вирішувати задачі з використанням прикладного програмного забезпечення;

здійснювати обробку емпіричних даних за допомогою базових функцій електронних таблиць Excel і статистичного пакету "Аналіз даних";

розробляти алгоритми і створювати програмні засоби обробки емпіричних даних.

Виконання лабораторних робіт сприяє формуванню самостійності у аналізі проведених обчислень, дослідженні практичних задач, які є необхідною складовою підвищення технічного рівня підготовки студента. Зміст і структура методичних вказівок відображають новітні тенденції у питаннях навчання та підготовки кваліфікованих спеціалістів.

Лабораторні роботи (комп'ютерний практикум) дисципліни містять короткі теоретичні відомості, приклади розв'язування задач, також індивідуальні відповідні завдання та перелік основних теоретичних питань вказаної теми.

ЛАБОРАТОРНА РОБОТА № 1

Тема: Варіаційні ряди та статистичні розподіли. Обчислення емпіричних даних статистичними засобами первинної обробки.

Мета: закріпити, поглибити та узагальнити теоретичні знання, набуті студентами під час вивчення теми: «Емпіричні методи в наукових дослідженнях. Первинний статистичний аналіз», розвивати навички їх практичного застосування.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

У лабораторних роботах ми не маємо можливості працювати з будь-якими дійсними даними, тому необхідною складовою частиною кожної роботи буде генерація вибірки. В основі даної процедури лежить робота з генератором базової випадкової величини.

Базовою випадковою величиною (БВВ) у статистичному моделюванні називають неперервну випадкову величину z, рівномірно розподілену на інтервалі [0, 1].

БВВ моделюється на ЕОМ за допомогою генераторів БВВ. Генератор БВВ – це пристрій або програма, що видає за запитом одне або кілька незалежних значень $z_1,\ldots,z_n$ БВВ.

Генератори БВВ бувають трьох типів: табличні, фізичні та програмні.

Табличний генератор БВВ – це таблиця випадкових чисел. Основний недолік такого генератора полягає в обмеженій кількості випадкових чисел у таблицях.

Фізичний генератор БВВ – це спеціальний радіоелектронний пристрій в ЕОМ, що містить джерело шуму. Шум перетворюється на випадкові числа з рівномірним розподілом. Недоліками є складність тиражування (необхідна додаткова плата) та схемна нестабільність такого генератора.

Програмний генератор БВВ, зазвичай, обчислює значення $z_1,\ldots,z_n$ за рекурентною формулою типу $z_i = f(z_n)$ у разі заданого стандартного значення z_0. Оскільки z_0 повністю визначає всю послідовність, її називають псевдовипадковою. Але її статистичні властивості повністю відповідають властивостям ”справді випадкової”.

Найпростіший метод програмної генерації випадкових чисел – мультиплікативний; у його основі лежить таке рекурентне співвідношення:

$$z_i = \left(A \cdot z_{i-1} + C\right) \bmod M ,$$

де z_i, z_{i-1} – чергове і попереднє випадкові числа відповідно; A, C – константи; M – велике ціле позитивне число (чим більше M, тим довша неповторювана послідовність).

Дискретна випадкова величина d задається множиною можливих значень $\{d_1, d_2, \ldots, d_n\}$ та їх ймовірністю $p_1, \ldots p_n$. У випадку, коли результати рівноймовірні ($p_1 = \ldots = p_n = 1/n$) алгоритм наступний:

$$d = [nz] + 1, \qquad (1.1)$$

де [ ] – операція обчислення цілої частини числа;

z – базова випадкова величина.

Емпіричні дані, що отримані шляхом вимірювання властивостей вибіркових об'єктів – чи отримані методом імітаційного моделювання, як описано вище – повинні пройти первинну обробку. Під первинною обробкою найчастіше розуміють внесення у табличні форми (табуляцію), впорядкування у варіаційні послідовності (або ряди), групування (при побудові інтервального варіаційного ряду), побудова статистичного розподілу, обчислення окремих простих статистичних параметрів (статистик).

Варіаційний ряд - це упорядкована за збільшенням (або за зменшенням) послідовність значень досліджуваної змінної X (у табл. 1 значення x^*_j). Варіаційний ряд дає можливість наочно і швидко сприйняти структуру даних: варіанти значень (x_i,), які може приймати і приймає змінна X, а також кількість відповідних варіант (m_i), їхні мінімальне і максимальне значення. Варіаційний ряд дозволяє безпосередньо оцінити деякі важливі показники вибірки, наприклад, моду і медіану. Систематизація даних у варіаційний ряд є підготовчим етапом до розрахунків і побудови статистичних розподілів досліджуваної змінної.

Статистичний розподіл - це математична модель об'єктів реальності у вигляді співвідношення значень змінної X, що характеризує властивості вибірки, до частот їх появи. Наприклад, стовпчики значень (x_i), (варіанти X) і значень m_i (кількість варіант) у табл. 1 по суті утворюють статистичний розподіл, який розкриває залежність частоти появи (f_i) від значень (x_i,) змінної, тобто $f_i \sim x_i$. Отже, під поняттям «статистичний розподіл» $f(x)$ слід розуміти емпіричний розподіли частот появи певних значень досліджуваної змінної (слово «частота» нерідко опускають, маючи на увазі його присутність). Частота f_i - це функція, де аргументом виступає варіанта x_i.

Статистичні розподіли можна класифікувати за ознакою типів вимірювань змінної на: варіаційні, ранжировані та атрибутивні.

Рис. 1.1. Класифікація статистичних розподілів за типами вимірювань.

Варіаційні розподіли базуються на даних, які виміряні за шкалою відношень або інтервалів. Ранжировані розподіли застосовуються у разі порядкових (рангових) типів вимірювання. Атрибутивні розподіли характеризують дані, які виміряні за номінальними шкалами або шкалами «найменувань». Основні види статистичних розподілів такі: диференціальні та інтегральні, які можуть складатися з абсолютних і відносних частот (рис. 1.2).

Рис. 1.2. Основні види статистичних розподілів.

Диференціальні розподіли представляють значення частот окремо (тобто диференційовано) для кожної варіанти x_i, змінної X.

Диференціальні абсолютні частоти - це кількості об'єктів m_i, з однаковими значенням x_i, змінної X (або кількість однакових значень).

Диференціальні відносні частоти - це відношення диференціальних абсолютних частот m_i, до загальної кількості n поб'єктів, тобто $f_i = m_i/n$.

Інтегральні розподіли («накопичені» або «кумулятивні») формуються як доданки попередніх диференціальних частот. Вони визначають сумарні частоти для варіанти, що не перевищує значення x_i змінної X.

Інтегральні абсолютні частоти $\sum_{i=1}^{j} m_i = m_i^{нак}$ це накопичена сума диференціальних абсолютних частот від 1-ї до j-ї варіанти.

Інтегральні відносні частоти $F_j = \sum_{i=1}^{j} f_i$ - це накопичена сума диференціальних відносних частот від 1-ї до j-ї варіанти.

Варіаційні розподіли у разі інтервальних або відносних типів вимірювань залежать від:

• характеру досліджуваної змінної - дискретна змінна, чи неперервна;

• діапазону значень змінної - вузький і невеликий, чи широкий і різноманітний.

Тому за технологією побудови варіаційні розподіли поділяють на розподіли не згрупованих і згрупованих варіант. З метою лаконічності домовимося їх називати не згрупованими і згрупованими розподілами. Для незгрупованих розподілів частоти мають відношення до безпосередніх значень варіант з варіативного ряду; для згрупованих розподілів - до груп (або інтервалів) значень варіант.

Тобто, стовпець Варіанти та Кількість варіантів утворюють статистичний розподіл – див. таблиця 1.1. Практично використовують також накопичені, або інтегральні частоти. Крім того, часто використовують не частоту варіантів, а долю в сумі всіх частот, яка рівна відношенню частоти варіанта до загального числа спостережень:

$$\omega_x = \frac{m_x}{n}, \qquad (1.2)$$

де ω_x – частість; у більшості випадків також називають частотою;

m_x – кількість варіанту x;

n – кількість елементів у вибірці.

Використовується також накопичена (інтегральна частість).

Таблиця 1.1. Обробка емпіричних даних.

Дані		Варіаційний ряд	Варіанти	Частота	Накопичена частота
j	x_j	x_j^*	x_i	m_i	$m_i^{нак}$
1	3	1	1	1	1
2	4	2	2	1	2
3	4	3	3	2	4
4	4	3			
5	3	4	4	5	9
6	2	4			
7	4	4			
8	4	4			
9	5	4			
10	1	5	5	1	10

Приклади розв'язування задач

Приклад. Записати у вигляді варіаційного ряду вибірку:

3; 9; 9; 5; 3; 6; 9; 5; 5; 5; 6; 3; 6; 5; 6; 3; 3; 5; 6; 5.

Визначити розмах вибірки, побудувати полігон частот і полігон відносних частот.

Розв'язування. Обсяг вибірки (число її елементів) дорівнює $\boldsymbol{n} = 20$. Упорядкуємо варіанти за зростанням і підрахуємо кількість повторень значень x_i у кожному варіанті, одержимо:

$$x_1 = 3; \quad x_2 = 5; \quad x_3 = 6; \quad x_4 = 9;$$
$$\boldsymbol{n}_1 = 5; \quad \boldsymbol{n}_2 = 7; \quad \boldsymbol{n}_3 = 5; \quad \boldsymbol{n}_4 = 3.$$

Контроль обчислень виконаємо шляхом підсумовування частот варіантів:

$$\Sigma\, \boldsymbol{n}_i = 20.$$

Для кожного варіанту $\boldsymbol{x}_i$ знаходимо відносну частоту: $\boldsymbol{n}_i^* = \boldsymbol{n}_i / \boldsymbol{n}$.

Отриманий варіаційний ряд набуде вигляду (див. табл. 2).

Таблиця 1.2.

Дискретний варіаційний ряд

x_i	3	5	6	9
$\boldsymbol{n}_i$	5	7	5	3
$\boldsymbol{n}_i^*$	0,25	0,35	0,25	0,15

Контроль обчислень виконаємо шляхом підсумовування відносних частот варіантів:

$$\Sigma\, \boldsymbol{n}_i^* = 1.$$

Розмах вибірки одержимо:

$$\boldsymbol{R} = \boldsymbol{x}_{max} - \boldsymbol{x}_{min} = 9 - 3 = 6.$$

Для побудови полігону частот нанесемо отримані точки $(\boldsymbol{x}_i, \boldsymbol{n}_i^*)$ із табл. 2 на графік (див. рис. 2).

Якщо обсяг вибірки великий (сотні варіантів), то дискретний варіаційний ряд стає незручною формою запису. Тоді увесь діапазон значень ознаки $\boldsymbol{\xi}$ розбивають на $\boldsymbol{k}$ часткових інтервалів, які не перетинаються, (синонім – розрядів) $[a_{i-1}; a_i)$, $(\boldsymbol{i} = 1, 2, \ldots, \boldsymbol{k})$.

Для кожного $\boldsymbol{i}$-го інтервалу підраховують кількість значень x_i ознаки $\boldsymbol{\xi}$, що потрапили в цей інтервал, – частоту інтервалу $(\boldsymbol{n}_i)$. Елемент x_i, який *співпав* із границею інтервалу, відносять до наступного інтервалу, а не до попереднього.

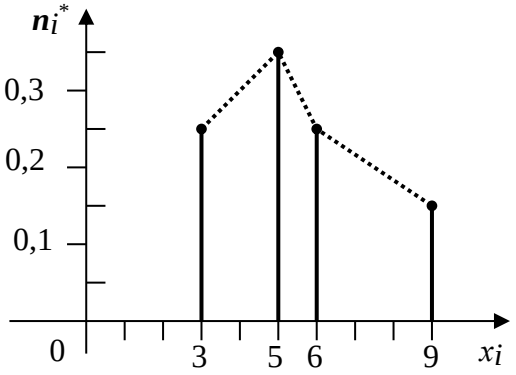


Рис. 1.3. Полігон відносних частот (див. табл. 1.2).

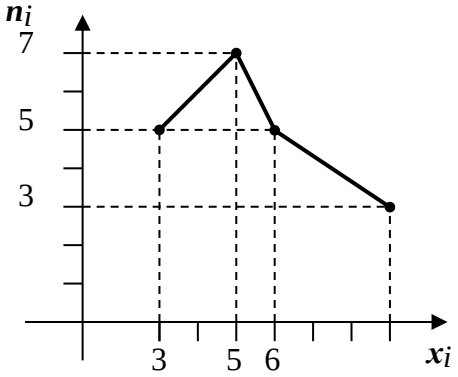


Рис. 1.4. Полігон частот (див. табл. 1.2).

Довжина розрядів може бути як однаковою, так і різною, що визначається необхідністю наявності в кожному інтервалі (розряді) ***не менше*** 5–10 значень випадкової ознаки ξ. Для ділянок із найбільшою щільністю значень ознаки довжина інтервалу (розряду) може бути меншою, із малою щільністю – більшою.

Раціональне число інтервалів (розрядів) складає 10–20.

Якщо до точності розрахунків немає дуже високих вимог, то частіше використовують розряди рівної довжини. Результати групування значень ознаки ξ за інтервалами записують у вигляді таблиці.

Інтервальним варіаційним рядом називається таблиця відповідності інтервалів значень (розрядів) випадкової величини ознаки ξ і частот n_i або/і відносних частот n_i^* входження значень ознаки в ці розряди, яка (таблиця) отримана за результатами спостережень або спроб (див. табл. 1.3). Такий ряд іноді називають *безперервним* або *статистичним* рядом.

Таблиця 1.3.

Інтервальний варіаційний ряд

Інтервал (розряд)	$[a_1; a_2)$	$[a_2; a_3)$	$[a_3; a_4)$	…	$[a_{m-1}; a_m]$
Частота n_i	n_1	n_2	n_3		n_m
Відносна частота n_i^*	n_1^*	n_2^*	n_3^*		n_m^*

Якщо в кожному інтервалі як представницьке значення ознаки ξ узяти середнє значення інтервалу: $x_i = \dfrac{a_{i-1} + a_i}{2}$, ($i$=1, 2, …, k), то *інтервальний варіаційний* ряд можна умовно подати вже розглянутим *дискретним варіаційним* рядом. У цьому випадку в першому рядку табл. 4.4 указуються *середні* значення ознак для кожного інтервалу.

Графічно варіаційний ряд зображають у вигляді полігону частот або полігону відносних частот.

Приклад. Використати 6 інтервалів рівної довжини і побудувати інтервальний варіаційний ряд розподілу за даною вибіркою:

15	18	22	26	15	10	23	27	19	18	14	15	28	6	29	7	11
26	24	19	14	15	7	27	14	19	8	20	5	29	16	10	16	5
18	20	12	16	22	23	20	21	11	16	22	22	6	18	14	11	

Потім перейти до дискретного варіаційного ряду.

Розв'язування. Обсяг вибірки $n = 50$.

Знайдемо розмах вибірки: $R = x_{max} - x_{min} = 29 - 5 = 24$.

За умовою кількість інтервалів $\boldsymbol{m} = 6$, тому довжина кожного часткового інтервалу $\boldsymbol{h} = \boldsymbol{R} / \boldsymbol{m} = 24 / 6 = 4$. Перший інтервал починається з $x_{min} = 5$ і закінчується точкою $x_{min} + \boldsymbol{h} = 5 + 4 = 9$, значення якої до складу інтервалу *не входить*. Ця ж точка є початком другого інтервалу, до складу якого вона *входить*. Виконуючи послідовність таких же розрахунків, одержимо границі 6 інтервалів: [5;9), [9;13), [13;17), [17; 21), [21; 25) [25; 29).

Підраховуємо кількість елементів вибірки, що потрапили в кожний із знайдених інтервалів. Елемент вибірки 21 є границею інтервалів [17; 21) і [21; 25). При підрахунку частоти відносимо його до інтервалу [21; 25). У результаті одержимо такий інтервальний варіаційний ряд (див. табл. 1.4).

Таблиця 1.4.

Інтервальний варіаційний ряд

$a_{i-1} \div a_i$	[5÷9)	[9÷13)	[13÷17)	[17÷21)	[21÷25)	[25÷29]
$\boldsymbol{n}_i$	7	6	12	10	7	8
$\boldsymbol{n}_i^*$	0,14	0,12	0,24	0,20	0,14	0,16
$n_i^{(n)}$	7	13	25	35	42	50
$F_n(x)$	0,00	0,14	0,26	0,50	0,70	0,84 ($F_n(x>29)=1$

Контроль обчислень виконаємо шляхом підсумовування частот варіантів, тоді одержимо: $\Sigma\, \boldsymbol{n}_i = 50$.

Для кожного часткового інтервалу обчисляємо відносну частоту за формулою: $\boldsymbol{n}_i^* = \boldsymbol{n}_i / \boldsymbol{n}$ і результат заносимо в нижній рядок таблиці.

Контроль обчислень виконаємо шляхом підсумовування відносних частот варіантів, одержимо: $\Sigma\, \boldsymbol{n}_i^* = 1$.

Для переходу від інтервального варіаційного ряду до дискретного варіаційного ряду візьмемо як варіанти середнє значення в кожному інтервалі: $x_i = (a_{i-1} + a_i) / 2$. Одержимо наступний дискретний варіаційний ряд:

x_i	7	11	15	19	23	27
$\boldsymbol{n}_i$	7	6	12	10	8	7
$\boldsymbol{n}_i^*$	0,14	0,12	0,24	0,20	0,16	0,14

Хід роботи

Написати програму, що генерує випадкову послідовність даних з $(N+10)$ елементів, які набувають значень із набору (1, 2, 3, 4, 5). Тут N – номер студента у журналі старости. Одержати та вивести на екран вихідні дані, варіаційний ряд, статистичний розподіл, інтегральну частоту та частість.

Рекомендовано максимально використовувати стандартні бібліотеки та алгоритми.

Звіт повинен містити схему алгоритму сортування (що відповідає використаному алгоритму), код програми з необхідними коментарями, зображення результату роботи програми з екрану, висновки.

Контрольні питання.

1. Що таке статистичний розподіл?
2. Що таке варіаційний ряд?
3. Методи побудови генераторів базової випадкової величини.
4. Що таке генеральна сукупність?
5. Які існують методи вибірок?
6. Що таке базова випадкова величина?
7. Методи одержання базової випадкової величини.

ЛАБОРАТОРНА РОБОТА № 2

Тема: Знаходження основних показників вибірки. **Міри центральної тенденції (МЦТ). Міри мінливості (ММ).**

Мета: ознайомитись з основними поняттями статистичних показників вибірки, начитись правильно застосовувати їх для розв'язування задач, встановлювати приховані взаємозв'язки та закономірності явищ, спрогнозовувати розвиток досліджуваних процесів.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

Статистичні показники, які розкривають властивості вибірки, можна представити такими основними групами:

– емпіричними розподілами, що характеризують структуру досліджуваної області;

– вибірковими показниками (мірами центральної тенденції та мінливості);

– кореляційно-регресійними показниками, які дають можливість встановити приховані взаємозв'язки та закономірності явищ, спрогнозувати розвиток досліджуваних процесів.

Мірами центральної тенденції (МЦТ) називають чисельні показники типових властивостей емпіричних даних. Існує порівняно невелика кількість таких показників-мір і в першу чергу: мода, медіана, середнє арифметичне. Кожна конкретна МЦТ має свої особливості, що роблять її цінною для характеристики об'єкта дослідження в певних умовах.

Мода Mo – це значення, яке найчастіше трапляється серед емпіричних даних. Так, для ряду значень 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5 мода дорівнює 3.

При визначенні моди дотримуються наступних домовленостей (матеріал, наведений нижче, не є вичерпним, пропонується для виконання роботи обмежитись даними варіантами):

– мода може бути відсутня, наприклад, для даних 2, 2, 3, 3, 4, 4, 5, 5;

– якщо варіанти суміжні і мають однакову частоту, мода визначається як середнє значення сусідніх варіант; наприклад, для ряду 2, 2, 3, 4, 4, 4, 5, 5, 5 мода рівна 4.5;

– якщо варіанти несуміжні, може існувати декілька мод. Так, для даних 2, 2, 3, 3, 3, 4, 5, 5, 5 характерна бімодальність, тобто дві моди M_{o1}=3, M_{o2}=5;

– емпіричні дані можуть мати великі та малі моди. Наприклад, дані 2, 2, 3. 3. 3 4. 4. 4. 5, 6. 6. 6. 6. 6. 6. 6. 7, 7, 7, 8, 9. 9. 9. 9 мають одну велику моду Mo_1= 6 та дві малі моди Mo_2 = 3,5 і Mo_3 = 9.

Медіана Md – це значення, яке проходить через середину упорядкованої послідовності емпіричних даних. Для непарної кількості даних медіана

визначається середнім елементом. Наприклад, для 11-ти значень 4, 4, 4, 5, 5, 5, 5, 5, 6 , 6, 7 медіана Md=5. Якщо кількість значень даних є парною, то медіаною є середнє значення центральних сусідніх елементів $Md = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$. Наприклад, для 12 значень 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 7, 7, медіана Md=(5+6)/2=5.5.

Середнє арифметичне сукупності n значень дорівнює

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n}. \qquad (2.1)$$

Середнім гармонічним чисел називають число, обернене середньому арифметичному їх обернених, тобто

$$A_{-1}(x_1, \ldots x_n) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_n}}. \qquad (2.2)$$

Середнім квадратичним двох чисел називають число, рівне квадратному кореню з середнього арифметичного квадратів двох чисел

$$S = \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}. \qquad (2.3)$$

Середнім геометричним додатних чисел називають таке число, яким можна замінити кожне з цих чисел так, щоб їх добуток не змінився

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \qquad (2.4)$$

Середнє геометричне двох чисел $\sqrt{a_1 \cdot a_2}$ називають їх середнім пропорційним.

Особливості мір центральної тенденції:

• мода вибірки обчислюється просто, її можна визначити «на око». Для дуже великих груп даних мода є досить стабільною мірою центру розподілу;

• медіана займає проміжне положення між модою і середнім з погляду її підрахунку. Ця міра особливо легко визначається у разі ранжированих даних;

• середнє арифметичне передбачає використовування всіх значень вибірки, причому всі вони впливають на значення цієї міри.

Обмеженість мір центральної тенденції для характеристики сукупностей можна продемонструвати на прикладі двох вибірок (рис. 2.1), які мають різні розподіли, проте однакові (і це не складно перевірити) МЦТ (значення моди Mo, медіани Md і середнього X дорівнюють 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Емпіричні дані												МЦТ			ММ	
2	i	1	2	3	4	5	6	7	8	9	10	11	Mo	Md	$\bar{X}$	s_x^2	s_x
3	Вибірка 1	3	3	3	4	4	4	4	4	5	5	5	4	4	4	0,6	0,77
4	Вибірка 2	2	2	3	4	4	4	4	5	5	5	6	4	4	4	1,6	1,26

Рис. 2.1. Властивості ММ.

Проте вибірки мають істотну різницю значень основних ММ: дисперсій s^2_x і стандартних відхилень s_x (див. два останні стовпчики рис. 2.1). Можна відзначити своєрідну «чутливість» показників ММ щодо властивостей сукупності.

Дисперсія вибірки обсягом пвизначається як:

$$s_x^2 = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots + (x_n - \overline{X})^2}{n-1},$$

$$\text{або } s_x^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}, \text{ де } \overline{X} - \text{середнє арифметичне вибірки}$$

Дисперсія вибірки s^2_x, що розрахована за цією формулою, є незміщеною оцінкою свого генерального параметра σ^2_x, завдяки внесенню поправки Бесселя n/(n - 1), тобто:

$$s_x^2 = \frac{\sum (x_i - \overline{X})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum (x_i - \overline{X})^2}{n-1}. \quad (2.5)$$

Різницю n- 1 називають числом степенів вільності к - кількість об’єктів або значень у складі обмеженої статистичної сукупності, які можуть вільно варіювати. Якщо обмежень вільності варіації існує декілька (ν), то число степенів вільності дорівнюватиме k= n-ν.

Тоді формула дисперсії має такий вигляд:

$$s_x^2 = \frac{1}{n-1}(\sum x_i^2 - n\overline{X}^2). \quad (2.6)$$

Якщо дані представлено розподілами частот, дисперсія визначається як

$$s_x^2 = \frac{1}{n-1} \cdot \sum f_i (x_i - \overline{X})^2, \quad (2.7)$$

де x_i, - варіанти незгрупованих частот або центральні значення класових інтервалів у разі згрупованих частот; f_i - диференціальні частоти, X - середнє.

Дисперсія служить мірою однорідності сукупностей емпіричних даних. Чим вища однорідність, тим нижче значення дисперсії. Для повністю однорідних сукупностей дисперсія дорівнює нулю.

Дисперсія генеральної сукупності обсягом N визначається як:

$$\sigma_x^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}, \quad (2.8)$$

$$\text{або } \sigma_x^2 = \frac{\sum (x_i - \mu)^2}{N}, \text{ де } \mu = \frac{1}{N}\sum x_i -$$ середнє арифметичне генеральної сукупності.

Стандартне відхилення вибірки визначається як $s_x = \sqrt{s_x^2}$. (2.9)

Стандартне відхилення генеральної сукупності $\sigma_x = \sqrt{\sigma_x^2}$. (2.10).

Коефіцієнт варіації V_x використовується у разі порівняльної оцінки різноякісних середніх величин і визначається (у тому числі у %) як відношення стандартного відхилення до середнього арифметичного:

$$V_x = s_x / \overline{X} \cdot 100\% .$$ (2.11).

Приклад розв'язування задачі

Завдання. Побудувати статистичний розподіл вибірки, записати емпіричну функцію розподілу та обчислити такі числові характеристики:

- вибіркове середнє;
- вибіркову дисперсію;
- підправлену дисперсію;
- вибіркове середнє квадратичне відхилення;
- підправлене середнє квадратичне відхилення;
- розмах вибірки;
- медіану;
- моду.

Вибірка задана рядом *11, 9, 8, 7, 8, 11, 10, 9, 12, 7, 6, 11, 8, 7, 10, 9, 11, 8, 13, 8.*

Розв'язування: Запишемо вибірку у вигляді варіаційного ряду (у порядку зростання): 6; 7; 7; 7; 8; 8; 8; 8; 8; 9; 9; 9; 10; 10; 11; 11; 11; 11; 12; 13. Далі записуємо статистичний розподіл вибірки у вигляді дискретного статистичного розподілу частот:

x_i	6	7	8	9	10	11	12	13
n_i	1	3	5	3	2	4	1	1

Емпіричну функцію розподілу визначатимемо за формулою $F_n(x) = \frac{n_x}{n}$ де n_x – кількість елементів вибірки, що менші за x.

Використовуючи таблицю і враховуючи, що обсяг вибірки рівний n=20, запишемо емпіричну функцію розподілу:

$$F_n(x) = 0, \quad x \le 6;$$
$$F_n(x) = \frac{1}{20}, \quad 6 < x \le 7;$$
$$F_n(x) = \frac{4}{20}, \quad 7 < x \le 8;$$
$$F_n(x) = \frac{9}{20}, \quad 8 < x \le 9;$$
$$F_n(x) = \frac{12}{20}, \quad 9 < x \le 10;$$
$$F_n(x) = \frac{14}{20}, \quad 10 < x \le 11;$$
$$F_n(x) = \frac{18}{20}, \quad 11 < x \le 12;$$
$$F_n(x) = \frac{19}{20}, \quad 12 < x \le 13;$$
$$F_n(x) = 1, \quad x > 13.$$

Далі обчислимо числові характеристики статистичного розподілу вибірки.

Вибіркове середнє обчислюємо за формулою:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} =$$

$$= \frac{6\cdot 1 + 7\cdot 3 + 8\cdot 5 + 9\cdot 3 + 10\cdot 2 + 11\cdot 4 + 12\cdot 1 + 13\cdot 1}{20} =$$

$$= \frac{183}{20} = 9,15.$$

Вибіркову дисперсію знаходимо за формулою

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\bar{x})^2.$$

$$\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_k x_k^2}{n} =$$

$$= \frac{36\cdot 1 + 49\cdot 3 + 64\cdot 5 + 81\cdot 3 + 100\cdot 2 + 121\cdot 4 + 144\cdot 1 + 169\cdot 1}{20}$$

$$= \frac{1743}{20} = 87,15.$$

Вибіркове середнє, що фігурує в формулі у квадраті знайдено вище. Залишається все підставити у формулу

$$D_B = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\bar{x})^2 = 87,15 - (9,15)^2 = 3,43.$$

Підправлену дисперсію обчислюємо за формулою:

$$s^2 = \frac{n}{n-1} D_B = \frac{20}{19}\cdot 3,43 = 3,61.$$

Вибіркове середнє квадратичне відхилення обчислюємо за формулою:

$$\sigma_B = \sqrt{D_B} = \sqrt{3,43} = 1,85.$$

Підправлене середнє квадратичне відхилення обчислюємо за формулою:

$$s = \sqrt{s^2} = \sqrt{3,61} = 1,9.$$

Розмах вибірки обчислюємо як різницю між найбільшим і найменшим значеннями її варіант, тобто: $R = x_k - x_1 = 13 - 6 = 7.$

Медіану обчислюють за формулами:

$$Me(X) = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$$ якщо число n – парне;

$$Me(X) = x_{\frac{n+1}{2}}$$ якщо число n – непарне.

Тут беремо індекси в хі згідно з нумерацією варіант у варіаційному ряді.

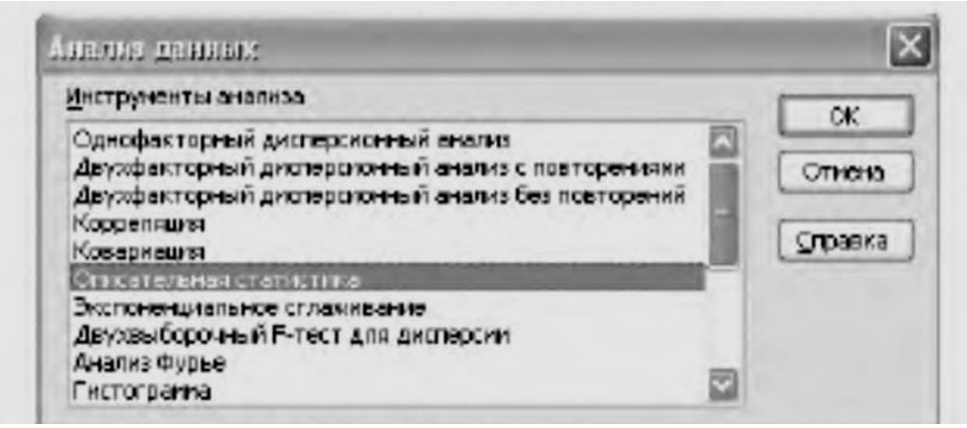
У нашому випадку n=20, тому $Me(X) = \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(9+9) = 9.$

Мода – це варіанта, яка у варіаційному ряді трапляється найчастіше, тобто $Mo(X) = 8.$

Розрахунки показників МЦТ Та ММ можна здійснити в MS Excel трьома способами:

- математичних операцій за відповідними формулами МЦТ і ММ;
- вбудованих статистичних функцій:

Обсяг вибірки	=СЧЁТ()	Дисперсія	=ДИСП()
Середнє	=СРЗНАЧ()	Ст. відхилення	=СТАНДОТКЛОН()
Мода	=МОДА()	Асиметрія	=СКОС()
Медіана	=МЕДИАНА()	Ексцес	=ЭКСЦЕСС()



- спеціального розділу «Описова статистика» пакету «Аналіз даних»: виконати команди головного меню Excel Аналіз даних із закладки Дані та викликати діалогове вікно «Описова статистика». Встановити в описовому вікні «Описова статистика» вхідні дані та параметри виводу, отримати результат.

Хід роботи

Написати програму, що генерує випадкову послідовність даних з $(N+10)$ елементів, які набувають значень із набору (1, 2, 3, …., (N+1)). Тут N – номер студента у журналі старости. Розподіл рівномірний.

Вивести початкову та впорядковану послідовність.

Обчислити моду, медіану, середнє арифметичне сукупності. Обчислення кожної з цих величин має бути реалізоване у вигляді окремого блоку програми (функції/процедури/класу/методу класу залежно від використовуваного середовища для виконання робіт).

Написати програму, яка перевіряє розроблений раніше елемент для обчислення моди. Для цього сформувати таблицю прикладів, які відображають різні варіанти визначення моди та результати при цьому. Здійснити перевірку.

Протокол перевірки в зрозумілій формі має відображатись на екрані (і в звіті). До звіту внести як результат першої перевірки (з можливими наявними помилками), так і виправлений варіант.

Зробити висновки.

Завдання для самостійного виконання:

Завдання 1. Нижче наведено дані про наслідки 150 аналізів відносно вмісту триокису сірки в суміші (у відсотках), проведених на протязі місяця.

15,8	16,0	15,7	16,0	15,7	15,9	16,0	15,7	15,8	15,7
15,4	15,7	15,8	15,7	15,9	16,0	15,7	15,7	15,7	15,8
15,8	15,6	15,9	15,8	15,5	16,0	15,7	15,7	15,7	15,8
15,9	15,7	15,8	16,0	15,8	15,9	16,2	15,7	15,5	15,9
15,7	15,7	15,3	15,6	16,1	15,7	16,1	15,9	15,8	16,0
16,1	15,7	15,5	15,6	15,8	15,6	15,8	15,8	15,6	15,7
15,6	15,9	15.8	15,8	15,8	15,9	15,6	15,8	15,8	15,9
15,5	15,8	15,4	15,5	15,5	15,7	15,6	15,9	15,8	15,5
15,9	15,8	15,5	15,9	15,6	15,8	15,6	15,7	15,7	15,7
15,7	15,7	16,0	16,1	15,6	15,5	15,6	15,5	16,0	15,5
15,8	15,8	15,9	16,1	15,5	15,7	16,0	15,9	15,7	15,5
16,1	15,7	15,7	15,5	16,2	15,7	15,6	16,0	15,6	15,7
15,3	15,5	15,4	16,0	15,7	15,5	15,8	15,4	15,7	16,3
15,9	15,6	15,7	15,4	15,9	15,6	16,0	15,7	15,8	15,9
16,0	16,0	15,8	15,9	15,7	15,6	15,6	15,9	15,6	15,5

Побудувати гістограму та емпіричну функцію розподілу. Обчислити вибіркове середнє значення відсоткового вмісту сірки в суміші, вибіркове середнє квадратичне відхилення,визначити моду, медіану.

Завдання 2. Наведені нижче дані являють собою кількості виробів, виготуваних на протязі однієї години деяким дрібним виробником.

Побудувати груповану вибірку, визначити медіану, моду вибірки. Обчислити середню кількість виробів за годину, вибіркове середнє квадратичне відхилення.

136	122	132	128	123	133	130	131
134	149	138	127	119	137	133	130
143	134	128	131	118	133	131	132
118	128	122	130	139	145	122	130
128	136	132	126	124	117	139	132
141	144	138	133	127	150	144	133
134	125	140	135	129	138	138	147
150	126	135	136	150	135	138	140
122	142	127	127	132	145	140	133
127	142	144	125	132	145	137	132

Контрольні питання

1. Мода.
2. Медіана.
3. Поняття про середнє степеневе.
4. Середнє гармонічне.
5. Середнє арифметичне.
6. Середнє квадратичне.
7. Середнє геометричне.
8. Що таке контроль якості?
9. Визначити поняття емпіричний.
10. Методи моделювання дискретних випадкових величин.

ЛАБОРАТОРНА РОБОТА № 3

Тема: Основи статистичної обробки неперервних даних. Побудова гістограм для нормального та показникового розподілів вибірки.

Мета: дослідити вплив обсягів експериментальних даних на емпіричні закони розподілу ймовірностей випадкових величин.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

Емпіричним (статистичним) розподілом випадкової величини називається перелік варіант і відповідних їм частот або відносних частот, які спостерігались у вибірці експериментальних даних.

Емпіричний розподіл можна визначити також у вигляді послідовності інтервалів та суми частот, що опинилися в інтервалах. Для цього виконується групування даних спостережень у вигляді інтервального варіаційного ряду.

Використовуючи вибіркові дані, можна визначити емпіричну функцію розподілу, а також гістограму та полігон, які в певній мірі відображають функцію щільності розподілу. Позначимо через N_x кількість спостережень (варіант), при яких величини ознак (тобто значень X) були меньшими за деяке x.

Емпіричною функцією розподілу називають функцію F*(x), яка для кожного значення x визначає відносну частоту події “X<x”:

F^* (x) N_x / N.

Різниця між емпіричною F*(x) та теоретичною F(x) функціями розподілу полягає в тому, що F(x) визначає імовірність події “X<x”, а F*(x) характерезує відносну частоту цієї події. Згідно з теоремою Бернуллі, у великій кількості дослідів відносна частота наближається до імовірності F(x) цієї події. Емпірична функція F*(x) має ті ж самі властивості, що і теоретична функція F(x).

Під час моделювання неперервних випадкових величин із заданим законом розподілу можна використовувати три методи:

- метод нелінійних перетворень;
- метод композицій;
- табличний метод (метод “звернення”).

Перші два методи не розглядатимемо, оскільки вони потребують достатньо серйозної математичної підготовки. Третій метод заснований на заміні закону розподілу неперервної випадкової величини спеціальним розрахунковим співвідношенням, яке дає змогу обчислювати значення випадкової величини за значенням випадкового числа, рівномірно розподіленого на інтервалі $(0,1)$. Наведемо співвідношення для показникового та нормального законів розподілу.

Показниковий (експотенційний) розподіл:

$$x = -\frac{1}{\lambda} \cdot \ln(z), \qquad (3.1)$$

де $\lambda > 0$ – параметр показникового розподілу,

z – рівномірно розподілене випадкове число.

Розрахункове співвідношення для випадкового числа нормального розподілу :

$$x = m + s\left[\sum_{i=1}^{12} z_i - 6\right], \qquad (3.2)$$

де m, s – параметри нормального закону розподілу (математичне очікування і середньоквадратичне відхилення); z_i – рівномірно розподілене випадкове число.

Математичне сподівання.

Однією з часто використовуваних на практиці характеристик при аналізі випадкових величин є математичне сподівання. Під даним терміном часто вживають "середнє значення" випадкової величини X. Розраховувати його не так важко, особливо якщо маємо дискретну величину з невеликою кількістю точок. **Математичним сподіванням** випадкової величини X визначеної на дискретній множині значень називається величина, яка рівна сумі попарних добутків величин X на їх ймовірності появи $M(X) = \sum_{i=1}^{\infty} x_i p_i$ Якщо множина обмежена, то потрібно шукати суму скінченного числа доданків $M(X) = \sum_{i=1}^{n} x_i p_i$ Якщо множина X є неперервною, то математичне сподіванням випадкової величини X визначається інтегруванням за формулою $M(X) = \int_{\Omega} x f(x) dx$ Якщо $\Omega = (-\infty; \infty)$, то $M(X) = \int_{-\infty}^{\infty} x f(x) dx$ Якщо $\Omega = [a; b]$, то $M(X) = \int_{a}^{b} x f(x) dx$.

Ймовірність випадкової події A дорівнює відношенню кількості випадків m, що сприяють появі події A до кількості всіх можливих випадків n:

$$p(A) = \frac{m}{n}.$$

Зауважимо, що ймовірність вірогідної події p(U)= 1, а ймовірність неможливої події p(V) = 0.

Приклади розв'язування задач.

Приклад 1. Закон розподілу дискретної випадкової величини задано таблично:

x_i	– 5	– 3	-1	2	4	5
p_i	0,1	0,1	0,3	0,25	0,15	0,1

Обчислити математичне сподівання.

Розв'язування. Згідно наведеної вище формули, обчислюємо

$$M(X)=\sum_{i=1}^{6}x_ip_i=x_1p_1+x_2p_2+x_3p_3+x_4p_4+x_5p_5+x_6p_6=$$
$$=-5\cdot 0,1-3\cdot 0,1-1\cdot 0,3+2\cdot 0,25+4\cdot 0,15+5\cdot 0,1=$$
$$=-0,5-0,3-0,3+0,5+0,6+0,5=0,5.$$

Таким чином, знайдене математичне сподівання рівне M(x)=0,5.

Середнє квадратичне відхилення – корінь квадратний із дисперсії. Її позначають грецькою літерою «сігма». $\sigma(X)=\sqrt{D(X)}.$

Розглянемо приклади для ознайомлення з практичною стороною визначення цих величин.

Приклад 2. Закон розподілу дискретної випадкової величини X задано таблицею:

x_i	– 2	– 1	1	3	5	6
p_i	0,2	0,1	0,3	0,1	0,2	0,1

Обчислити дисперсію D(X) та середнє квадратичне відхилення $\sigma(X)$.

Розв'язування. Згідно з властивостями дисперсії знаходимо

$$D(X)=M(X^2)-M^2(X)=\sum_{i=1}^{6}x_i^2p_i-\left(\sum_{i=1}^{6}x_ip_i\right)^2;$$

.

Математичне сподівання обчислюємо за формулою

$$M(X)=\sum_{i=1}^{6}x_ip_i=-2\cdot 0,2-1\cdot 0,1+1\cdot 0,3+3\cdot 0,1+5\cdot 0,2+6\cdot 0,1=$$

$$=-0,4-0,1+0,3+0,3+1+0,6=1,7;$$

.

Далі знаходимо $M(X^2)$

$$M(X^2)=\sum_{i=1}^{6}x_i^2p_i=(-2)^2\cdot 0,2+(-1)^2\cdot 0,1+1\cdot 0,3+3^2\cdot 0,1+5^2\cdot 0,2+6^2\cdot 0,1=$$

$$=0,8+0,1+0,3+0,9+5+3,6=10,7;$$ та дисперсію $$D(X)=10,7-(1,7)^2=10,7-2,89=7,81;$$

Середнє квадратичне відхилення рівне кореню з дисперсії $\sigma(X)=\sqrt{D(X)}=\sqrt{7,81}\approx 2,79.$.

На цьому обчислення завершені і Ви можете переконатися, що знаходження імовірнісних характеристик на практиці доволі просто реалізувати.

Отже, значення, які приймає досліджувана ознака, можуть відрізнятись на як завгодно малу величину. В такому випадку для аналізу значення ознаки групують по інтервалах.

Таблицю, що дозволяє судити про розподіл частот між інтервалами варіювання значень ознаки, називають інтервальним варіаційним рядом. Для побудови інтервального варіаційного ряду необхідно визначити величину інтервалу, встановити шкалу інтервалів, згрупувати результати спостережень.

Для побудови оптимальної величини інтервалу використовують емпіричну формулу Стерджеса:

$$h=\frac{X_{\max}-X_{\min}}{1+3.322\cdot\lg(n)}. \qquad (3.3)$$

Під оптимальною величиною інтервалу розуміють таку величину, при якій побудований варіаційний ряд не буде надто громіздким, але дозволить виявити особливості вибірки. Якщо в результаті обчислень h – дробове число, то рекомендується використовувати або близький нескладний дріб, або для величин, що описуються значеннями набагато більшими за 1 близьке ціле число.

Існують загально прийняті способи графічного відображення варіаційних рядів. Використовуються наступні варіанти графічного зображення: полігон та гістограма.

Полігоном частот називають ламану лінію, відрізки якої з'єднують точки (x1,N1), (x2,N2), ... , (xk ,Nk), де k – кількість різних варіант. Для побудови полігону частот на осі абсцис відкладають варіанти xi , а на осі ординат - відповідні частоти Ni, точки (xi ,Ni) з'єднують відрізками прямих ліній.

Полігоном відносних частот називають ламану лінію, відрізки якої з'єднують точки (x1,W1), (x2,W2), ..., (xk ,Wk). Для побудови полігона відносних частот на осі абсцис відкладають варіанти xi , а на осі ординат - відповідні відносни частоти Wi . Якщо результати вимірювань відносяться до неперервної випадкової величини X, то доцільно будувати гістограму. Полігон – як правило використовується лише для зображення дискретного варіаційного ряду. Для його

побудови у прямокутній системі координат наносять точки з координатами: (Xi;mX) і з'єднують їх прямими відрізками;

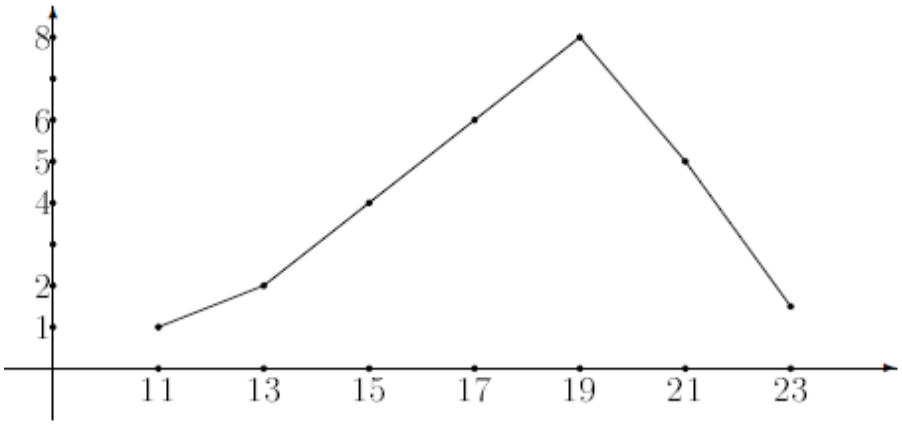


Рис. 3.1. Приклад полігона.

Гістограмою частот називають східчасту фігуру, яка складається з прямокутників, основами яких є часткові інтервали довжиною Δ, а висоти дорівнюють відношенню Ni / Δ , яке називається щільністю частоти.

Гістограмою відносних частот називають східчасту фігуру, яка складається з прямокутників, основами яких є часткові інтервали довжиною Δ , а висоти дорівнюють Wi / Δ (щільність відносної частоти). Гістограма – використовується для зображення інтервального варіаційного ряду. Для її побудови у прямокутній системі координат по осі абсцис (X) відкладають відрізки, що відображають інтервали. На цих відрізках будують прямокутники з висотами рівними частотам або частотностям інтервалів. Іноді інтервальний ряд зображають полігоном. В цьому випадку інтервали замінюють їх серединами.

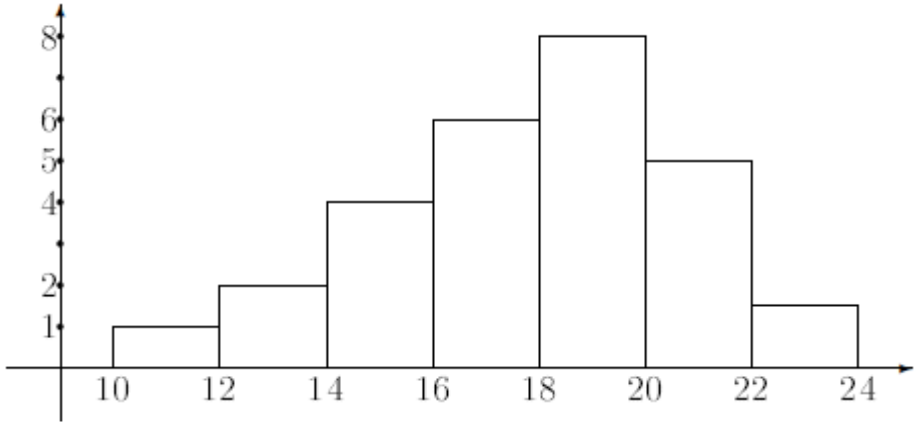


Рис. 3.2. Приклад гістограми.

- кумулятивна крива – крива накопичення частот або частотностей.
- огіва – якщо в кумулятивної кривої змінити місцями осі координат, то одержана крива називається огівою.

Хід роботи

Задатись двома вибірками із 20N елементів неперервних випадкових величин нормального та показникового розподілу в діапазоні від 0 до N включно. Побудувати гістограми. Для побудови графіків рекомендовано використовувати стандартні бібліотеки.

Контрольні питання

1. Який разподіл називається рівномірним?
2. Який розподіл називається нормальним?
3. Густина розподілу імовірності випалкової величини з нормальним розподілом.
4. Методи графічного зображення варіаційних рядів.
5. Емпірична формула Стерджеса.
6. Які бібліотеки для побудови графіків використовуються з мовою (середовищем) програмування?
7. Як розуміти поняття оптимальності для задачі визначення величини інтервалу?
8. Що таке integration tests?
9. Методи моделювання неперервних випадкових величин.

ЛАБОРАТОРНА РОБОТА № 4

Тема: Обчислення дисперсії, середнього квадратичного відхилення, коефіцієнтів асиметрії та ексцесу.

Мета: навчитися розраховувати точкові оцінки числових характеристик, побудувати надійні інтервали для математичного сподівання у випадку відомої та невідомої дисперсії.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

Для оцінювання розміру варіації використовується система абсолютних показників, які розглядаються як абсолютна міра варіації.

Розмах варіації (R) характеризує максимальну амплітуду коливань значень ознаки в сукупності:

$$R = X_{\max} - X_{\min}, \qquad (4.1)$$

де: $X_{\max}, X_{\min}$ – відповідно найбільше та найменше значення ознаки.

В інтервальних рядах розподілу розмах варіації визначається як різниця між верхньою межею останнього та нижньою межею першого інтервалу.

Середнє лінійне відхилення ($\bar{l}$), що характеризує середній розмір відхилень значень ознаки від середнього рівня. Для розрахунку за індивідуальними даними використовують середнє лінійне відхилення просте:

$$\bar{l} = \frac{\sum \left| X - \overline{X} \right|}{n}, \qquad (4.2)$$

де X – індивідуальні значення ознаки; $\overline{X}$ – середнє значення ознаки; n – кількість одиниць у сукупності.

Вибіркове середнє для дискретного статистичного ряду обчислюють за

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \text{ де } \sum_{i=1}^{k} n_i = n.$$

формулою: .

Вибіркове середнє для *інтервального статистичного ряду* обчислюють за формулою

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i z_i, \text{ де } z_i - \text{середина } i - \text{того інтервалу, } \sum_{i=1}^{k} n_i = n.$$

.

Дисперсія (σ^2) – це середній квадрат відхилень значень ознаки від її середнього рівня. Для розрахунку за індивідуальними даними використовують дисперсію просту:

$$\sigma^2 = \frac{\sum (X - \overline{X})^2}{n}. \qquad (4.3)$$

Оцінкою дисперсії σ 2 випадкової величини *X* є ***вибіркова дисперсія:***

$$\overline{s}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

яка є ***зміщеною*** оцінкою для σ^2.

Величина

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, s^2 = \frac{n}{n-1}\overline{s}^2$$

є ***незміщеною*** оцінкою дисперсії σ^2 випадкової величини *X*. Якщо математичне сподівання *a* відоме, то ***незміщеною сильно слушною оцінкою дисперсії*** σ^2 випадкової величини *X* є оцінка:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - a\right)^2.$$

Вибіркову дисперсію для *дискретного статистичного ряду* обчислюють

$$\overline{s}^2 = \frac{1}{n}\sum_{i=1}^{k}\left(n_i x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i {x_i}^2 - \overline{x}^2,$$

за формулою:
відповідно

$$s^2 = \frac{n}{n-1}\overline{s}^2;\ s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(n_i x_i - \overline{x}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i {x_i}^2 - \overline{x}^2.$$

Вибіркову дисперсію для *інтервального статистичного ряду* обчислюють за формулою:

$$\overline{s}^2 = \frac{1}{n}\sum_{i=1}^{k}\left(n_i z_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i {z_i}^2 - \overline{x}^2,$$

відповідно:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\left(n_i z_i - \overline{x}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i {z_i}^2 - \overline{x}^2.$$

Величина $\bar{s} = \sqrt{\overline{s^2}}$ називається ***вибірковим середнім квадратичним відхиленням***, $s = \sqrt{s^2}$ – ***вибірковим виправленим середнім квадратичним відхиленням*** (s^2 – вибіркова виправлена дисперсія).

Середнє квадратичне відхилення (σ) — показує середній розмір відхилень значень ознаки від середнього рівня.

Середнє квадратичне відхилення найчастіше використовують у статистичному аналізі, тому його також називають стандартним відхиленням.

Середнє арифметичне та дисперсія варіаційного ряду є частковими випадками більш загального поняття про моменти варіаційного ряду. Початковим моментом порядку q називають середнє арифметичне q степені варіантів:

$$\tilde{\nu}_q = \bar{x}^q = \frac{\left(\sum_x x^q \cdot m_x\right)}{\sum_x m_x}, \qquad (4.4)$$

де m_x – частота варіанта.

Відповідно до цієї формули початковим моментом нульового порядку буде 1. Початковий момент першого порядку буде рівний середньому арифметичному.

Центральним моментом порядку q називають середнє арифметичне q степенем відхилень варіантів від їх середнього арифметичного:

$$\tilde{m}_q = \frac{\sum_x \left(x - \bar{x}^2\right)^q \cdot m_x}{\sum_x m_x}. \qquad (4.5)$$

Центральний момент 1 порядку завжди рівний нулю, центр. момент 2 порядку відповідає дисперсії. Коефіцієнтом асиметрії називають відношення центрального моменту третього порядку до куба середнього квадратичного відхилення.

$$\tilde{A} = \frac{\widetilde{\mu^3}}{S^3} = \frac{\sum_x (x-\bar{x})^3 * m_x}{\sum_x m_x * S^3}. \qquad (4.6)$$

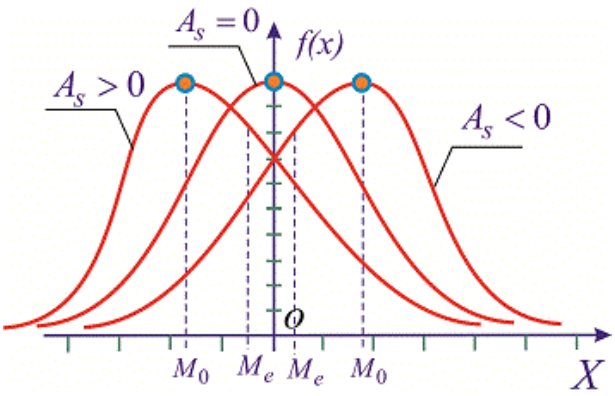
Встановлення асиметрії та ексцесу дозволяє встановити симетричність розподілу випадкової величини *X* відносно *M(X)=1*. Для цього знаходять третій центральний момент, що характеризує асиметрію закону розподілу випадкової величини. Якщо він рівний нулю $\mu_3 = 0$, то випадкова величина *X* симетрично розподілена відносно математичного сподівання *M(X)*. Оскільки момент μ_3 має розмірність випадкової величини в кубі, то вводять безрозмірну величину — коефіцієнт асиметрії: $As = \frac{\mu_3}{\sigma^3}$.

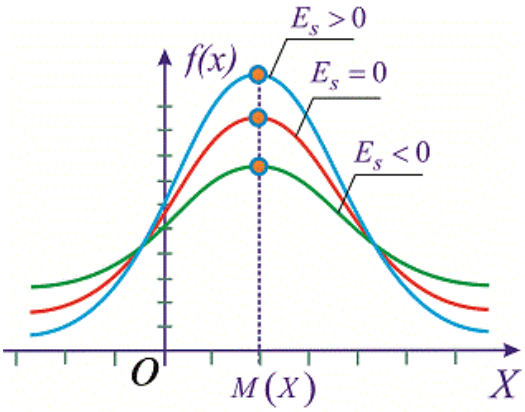
Центральний момент четвертого порядку використовується для визначення ексцесу, що характеризує плосковершинність, або гостровершинність щільності ймовірності $f(x)$. Ексцесом (коефіцієнтом крутості) називають зменшене на 3 відношення центрального моменту 4 порядку до 4 степені середнього квадратичного відхилення: $\tilde{E} = \mu^4 / s^4 - 3$. Для кривої, яка представляє собою нормальний розподіл величина ексцесу рівна нулю.

Число 3 віднімається для порівняння відхилення від центрального закону розподілу (нормального закону), для якого справджується рівність: $\frac{\mu_4}{\sigma_4} = 3$.

Отже, ексцес рівний нулю *Es=0* для нормального закону розподілу. Якщо ексцес додатній *Es>0* то на графіку функція розподілу має гостру вершину і для від'ємних значень *Es<0* більш пологу. В такий спосіб можна встановити відхилення заданого закону від нормального. Для наочності при різних значеннях асиметрії і ексцесу *Es<0* графіки щільності ймовірностей *f(x)* зображені на рисунках нижче.

Якщо полігон варіаційного ряду скошений, то такий ряд називають асиметричним. Якщо у варіаційному ряді більше варіантів, менших за середнє арифметичне, то говорять, що є лівостороння асиметрія (відповідно, правостороння асиметрія). Для помірно асиметричних рядів коефіцієнт асиметрії менший одиниці.





Хід роботи

Задатись вибіркою із 5N елементів (нормального розподілу) в діапазоні від 0 до N включно з довільними параметрами.

Обчислити дисперсію, середньє квадратичне відхилення, коефіцієнти асиметрії та ексцесу на основі змодельованих даних.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

ЗАДАЧА

Побудувати статистичний розподіл вибірки, записати емпіричну функцію розподілу та обчислити такі числові характеристики: 1) вибіркове середнє, 2) вибіркову дисперсію, 3) підправлену дисперсію, 4) вибіркове середнє квадратичне відхилення, 5) підправлене середнє квадратичне відхилення, 6) розмах вибірки, 7) медіану, 8) моду, 9) квартильне відхилення, 10) коефіцієнт варіації, 11) коефіцієнт асиметрії та 12) ексцес для вибірки: (далі – по варіантах, № варіанту = № студента у списку групи).

Варіант-1
3, 2, 3, 6, 5, 4, 7, 2, 1, 6, 4, 3, 2, 5, 4, 6, 3, 6, 8, 3.
Варіант-2
8, 7, 8, 11, 10, 9, 12, 7, 6, 11, 9, 8, 7, 10, 9, 11, 8, 11, 13, 8.
Варіант-3
4, 3, 4, 7, 6, 5, 8, 3, 2, 7, 5, 4, 3, 6, 5, 7, 4, 7, 9, 4.
Варіант-4
6, 5, 6, 9, 8, 7, 10, 5, 4, 9, 7, 6, 5, 8, 7, 9, 6, 9, 11, 6.
Варіант-5
5, 4, 5, 8, 7, 6, 9, 4, 3, 8, 6, 5, 4, 7, 6, 8, 5, 8, 10, 5.
Варіант-6
3, 8, 6, 3, 6, 4, 5, 2, 3, 4, 6, 1, 2, 7, 4, 5, 6, 3, 2, 3.
Варіант-7
8, 13, 11, 8, 11, 9, 10, 7, 8, 9, 11, 6, 7, 12, 9, 10, 11, 8, 7, 8.

Контрольні питання

1. Розмах варіації.
2. Стандартне лінійне відхилення.
3. Середнє квадратичне відхилення. Дисперсія. Зміщені і незміщені оцінки.
4 Поняття про моменти.
5. Асиметрія та ексцес.
6. Поняття життєвого циклу програмного забезпечення.

ЛАБОРАТОРНА РОБОТА № 5

Тема: Основи програмування перевірки статистичних гіпотез.

Мета: перевірити гіпотезу про згоду закону розподілу статистичних даних із нормальним законом, дослідити властивості показників статистичного зв'язку між випадковими величинами.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

У математичній статистиці часто виділяють особливий розділ, в якому розглядається перевірка гіпотез. Статистична перевірка гіпотез застосовується для того, щоб використовувати отриману за вибіркою інформацію для судження про закон розподілу генеральної сукупності. Зазвичай статистична гіпотеза перевіряється за допомогою критеріїв згоди, які дозволяють оцінити відповідність того чи іншого теоретичного закону розподілу деякого емпіричного ряду розподілу.

Статистичною гіпотезою називається будь-яке припущення щодо властивостей генеральної сукупності на основі оцінок вибірки, припущення щодо виду або параметрів невідомого закону розподілу. Статичну гіпотезу прийнято позначати Н.

Розрізняють прості і складні статистичні гіпотези.

Проста гіпотеза – повністю визначає теоретичну функцію розподілу випадкової величини. Ті що не визначають – називаються складними. Статистичні гіпотези підрозділяються на нульові та альтернативні.

Нульова гіпотеза - позначається H_0 – це гіпотеза про відсутність відмінностей у значеннях ознак.

Альтернативна гіпотеза – позначається H_1 – це гіпотеза є логічним запереченням H_0, тобто – це гіпотеза про існування відмінностей.

Статистичні гіпотези можуть бути також:

- *спрямованими* – висувають про те, що значення показника в одній сукупності нижче ніж значення показника в іншій. Також називають однобічними;

- *неспрямованими* – формулюють якщо необхідно довести відмінності форми розподілу або значень показників відхилень. Також називають двобічними.

Перевірка гіпотез здійснюється на основі статистичних критеріїв.

Критерії згоди повинні дати відповідь на питання, чи можна прийняти для даного емпіричного розподілу модель, відображену деяким теоретичним законам розподілу. У математичній статистиці близькість емпіричних і теоретичних розподілів оцінюють за допомогою критеріїв згоди, які розроблені

багатьма вченими. Одні з них оцінюють вірогідність розбіжності між емпіричними і теоретичними даними (критерії згоди Пірсона та Колмогорова), інші конкретно відповідають на питання про можливість збігу даного емпіричного розподілу і вибраного теоретичного закону (критерії згоди Романовського і Ястремського).

Статистичний критерій – це правило, що забезпечує математично обґрунтоване прийняття істинної і відхилення помилкової гіпотези. Статистичні критерії – практично являють собою метод розрахунку певного числа, яке позначається як емпіричне значення критерію. Дане значення порівнюється з деким критичним значенням для даного критерію. Співідношення між ними є підставою для підтвердження чи спростування гіпотези.

Критерії поділяються на:

- *параметричні* – використовуються в завданнях перевірки параметричних гіпотез і включають в свій розрахунок конкретні показники розподілу. Дозволяють безпосередньо оцінити параметри сукупностей чи вибірок. Оцінити середні відмінності в дисперсіях. Такі критерії дають можливість виявити тенденції зміни ознак, оцінити впливи факторів на ознаку;

- *непараметричні* – оперують частотами, рангами тощо.

Застосування таких критеріїв для прийняття або відхилення статистичних гіпотез завжди здійснюється з певною довірчою ймовірністю, інакше кажучи на певному рівні значущості.

Рівень значущості – це ймовірність того, що ми в результаті застосування критеріїв визнали відмінності істотними, а насправді вони випадкові. Рівень статистичної значущості у більшості випадків прийнятий за 5%. Існує значна кількість різних типів статистичних гіпотез. Ці типи визначаються сукупністю завдань та методів їх Розв’язування.

Основні групи статистичних гіпотез за прикладними задачами, яких вони стосуються:

- гіпотези стосовно закону розподілу;

- гіпотези стосовно чисельних показників параметрів розподілів;

- гіпотези стосовно однорідності вибірок;

- гіпотези стосовно рівня ознак досліджуваного явища або процесу.

Незважаючи на різноманітність типів гіпотез і критеріїв загальна схема перевірки статистичних гіпотез така:

1) формулювання нульової та альтернативної гіпотези на основі задачі дослідження;

2) перевірка припущень щодо відповідності розподілам, перевірка параметрів вибірки, та іншої додаткової інформації;

3) прийняття рівня значущості;

4) вибір статистичного критерію;

5) розрахунки емпіричного критерію;

6) визначення області критичних значень критерію;

7) прийняття статистичного рішення;

8) формулювання статистичних висновків;

9) формулювання змістовних висновків.

В статистиці існують 2 підходи стосовно методів перевірки гіпотез. За одним із них обов'язково формулюють і 0, і альтеративну гіпотезу, перевірки яких відбуваються незалежно і повноцінно. При іншому підході формулювання альтернативних гіпотез не відбувається взагалі.

Задачі що вирішуються:

- перевірка гіпотез щодо однорідності вибірок;
- перевірка гіпотез про чисельні значення параметрів;
- гіпотези про виявлення відмінностей та зсувів в ознаках;
- перевірка значущості коефіцієнтів кореляції.

Методи параметричної оцінки законів розподілу.

Для визначення чисельних значень параметрів передбачуваних розподілів застосовуються методи: лінійного оцінювання, імовірнісних сіток, квантилів, максимального правдоподібності і моментів.

Найбільш простим і поширеним є метод моментів, який полягає в тому, що початкові і центральні моменти теоретичного розподілу, що залежать від невідомих параметрів цього розподілу, прирівнюються до статистичних моментів. При цьому статистичні початкові і центральні моменти відповідно 1-го і 2-го порядків визначаються за формулами

$$V_1 = \frac{1}{n}\sum_{j=1}^{n} X_j^1;\ M_2 = \frac{1}{n}\sum_{j=1}^{n}(X_i - \overline{X})^2 . \qquad (5.1)$$

Формули взаємозв'язку аналітичних моментів з невідомими параметрами теоретичних розподілів наведені в таблиці (Додаток Б). У результаті вирішується система рівнянь, що зв'язує параметри з моментами, звідки визначаються оцінки відповідних параметрів.

Критерій асиметрії та ексцесу застосовують для приблизної перевірки гіпотези про нормальність емпіричного розподілу. Для нормального розподілу коефіцієнти асиметрії та ексцесу рівні 0. Практично щоб одержати оцінку за даним методом обчислюються так звані дисперсії асиметрії та ексцесу:

$$D(A) = \frac{6(n-2)}{(n+1)(n+3)}, \qquad (5.1)$$

$$D(E) = \frac{24n(n-2)(n-3)}{(n+1)(n+3)(n+5)}. \qquad (5.2)$$

Вважають, що при нормальному розподілі вибіркові показники асиметрії та ексцесу дорівнюватимуть нулю, але реально таке майже не спостерігається. Тому емпіричний розподіл вважають близьким до нормального (приймають нульову гіпотезу), якщо виконуються умови: $|A_x| \le 3\sqrt{D(A)}$ та $|E_x| \le 5\sqrt{D(E)}$.

Технологічно при цьому розраховують показники $t_A = \dfrac{|A_x|}{\sqrt{D(A)}}$ і $t_E = \dfrac{|E_x|}{\sqrt{D(E)}}$. Про достовірну відмінність емпіричного розподілу від нормального свідчать показники t_A і t_E, якщо приймають значення 3 і більше.

Критерій Романовського.

Романовський В.І. запропонував використовувати критерій хі-квадрат в іншому вигляді. Значення критерію обчислюється за формулою:

$$\beta = \frac{|\chi - K|}{\sqrt{2K}}, \qquad (5.4)$$

де K - число ступенів свободи.

У тому випадку, якщо β за абсолютним значенням менше 3, то розбіжність між емпіричним і теоретичним розподілами вважається несуттєвою і прийнятий закон розподілу можна прийняти в якості моделі емпіричного розподілу. Якщо ж вираз β більше 3, то розбіжність між розподілами істотна.

Ставлення Романовського ґрунтується на тому, що математичне сподівання χ дорівнює числу, а дисперсія – подвоєному числу ступенів свободи ($2K$). В цьому випадку імовірність відхилення величини хі-квадрат на $3\sigma^2 = 3\sqrt{2K}$ близька до одиниці.

Хід роботи

Задатись вибіркою з (2А+5) елементів із показниковим (експотенційним) розподілом. Задатись вибіркою з 100*(2А+5) елементів із нормальним розподілом (А – номер у журналі старости). Написати програму, яка формує протокол статистичного дослідження для довільних даних на основі критерію асиметрії та ексцесу, результати представити для одержаних вибірок. Для одержання нормального розподілу рекомендується використати спеціальні алгоритми (бібліотеки).

Контрольні питання

1. Що таке статистична гіпотеза?
2. Що таке статистичний критерій?
3. Які є статистичні критерії?
4. Методи генерації вибірок з нормальним розподілом.
5. Критерій асиметрії та ексцесу.
6. Для чого проводиться статистична перевірка гіпотез про закон розподілу генеральної сукупності?
7. Які Ви знаєте критерії згоди і для чого вони застосовуються?
8. Яке призначення критерію Романовського і для чого застосовується?
9. Які теоретичні закони розподілу Ви знаєте?
10. Які ви знаєте підходи до організації промислового розроблення програмного забезпечення?
11. Назвіть основні положення екстремального програмування.
12. У чому полягає концепція парного програмування?

ЛАБОРАТОРНА РОБОТА № 6

Тема: Програмна реалізація критерію узгодженості Пірсона.

Мета: оцінити параметри і перевірити гіпотези про відповідність емпіричних даних передбачуваному теоретичному розподілу.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості.

Після виконання оцінок статистичних параметрів може виникнути **задача ідентифікації** отриманого в спробі розподілу $F^*(x)$ випадкової величини ξ з *відомими* теоретичними функціями розподілу. Таку задачу називають *задачею вирівнювання*.

Задачею вирівнювання статистичних рядів називається задача добору підхожої теоретичної функції розподілу ймовірностей або щільності ймовірностей при обробці статистичних даних. Задача може бути виконана як за методом найменших квадратів (див. п. 4.10), так і з інших міркувань, наприклад із використанням *методу моментів*.

Метод моментів передбачає виконання вимоги збігу статистичних моментів статистичного ряду випадкової величини ξ, наприклад математичного сподівання і дисперсії, із їх значеннями в теоретичній кривій розподілу ймовірностей $f(x)$. Якщо крива лінія $f(x)$ залежить від двох параметрів, то можна підібрати їх так, щоб співпали перші два моменти і так далі – до чотирьох моментів.

Точність розрахунків моментів вище четвертого порядку різко спадає, тому їх застосування уважається недоцільним.

Після вирівнювання статистичного розподілу $F^*(x)$ за допомогою теоретичної кривої $F(x)$ можна помітити, що між цими кривими є розбіжності.

Тому виникає таке питання: чи є отримані розбіжності випадковими, пов'язаними з недостатньою кількістю спроб, або вони вказують на невідповідність підібраної теоретичної кривої лінії $F(x)$ реальному розподілу $F^*(x)$

Формалізуємо цю ситуацію у вигляді такої гіпотези H_0: випадкова величина ξ ***має*** інтегральний закон розподілу $F(x)$.

Для відповіді на питання "Дану гіпотезу H_0 варто прийняти або спростувати?" служать так називані "критерії згоди".

Критерій узгодженості Пірсона слугує для перевірки гіпотези про те, що дійсний розподіл ВВ і гіпотетичний розподіл є однаковими.

Нехай у результаті n незалежних спостережень за випадковою змінною y має функцію розподілу $F(x)$, що також відповідає певному гіпотетичному

розподілу. На основі цієї вибірки ми хочемо перевірити правдоподібність гіпотези H_o: обидві випадкові величини належать до одного закону розподілу.

Функція розподілу випадкової змінної y повністю описує її, зокрема і простір її можливих значень. Даний простір P розіб'ємо довільно на $r+1$ $(r=1,2,\ldots)$ частину $S_1, S_2, \ldots S_{r+1}$ так, що $S_i \cdot S_j = 0 (i \neq j)$.

Нехай у разі такого розбиття в комірку S_i попадає m_i $(i=1,\ldots,r+1)$ елементів вибірки x, де $\sum_{i=1}^{r+1} m_i = n$. Отже, відносна частота трапляння вибіркових значень у комірку, $S_i = \frac{m_i}{n}$. Згідно з гіпотетичною функцією розподілу $F(x)$ імовірність попадання значень випадкової змінної y в цю саму комірку $P(y \in S_i) = p_i$ $(i=1,\ldots,r+1)$, де $\sum_{i=1}^{r+1} p_i = 1$.

Якщо обидві вибірки випадкової величини дійсно належать до одного закону розподілу і керуються функцією розподілу $F(x)$, то за великого n майже напевно (про це існують теореми, доказані Я. Бернуллі та Е. Бореля) $\frac{m_i}{n}$ як завгодно мало відрізняється від p_i. Тому за міру узгодженості висунутої гіпотези природно вибирати величину

$$\chi^2(r,n,F) = \sum_{i=1}^{r+1} C_i \left(\frac{m_i}{n} - p_i \right)^2, \qquad (6.1)$$

де C_i – деякі додатні константи.

К. Пірсон показав, що при $C_i = \frac{n}{p_i}$ вибірковий розподіл величини:

$$\chi^2(r,n,F) = \sum_{i=1}^{r+1} \frac{(m_i - np_i)^2}{np_i} \qquad (6.2)$$

при $n \to \infty$ прямує до розподілу, який не залежить від виду гіпотетичної функції розподілу $F(x)$ генеральної сукупності.

Отже, алгоритм критерію χ^2 перевірки приналежності вибірки x з гіпотезою H_0 про функцію розподілу $F(x)$ і приналежності двох послідовностей до одного закону розподілу, такий.

Вибираємо рівень значущості α. Фіксуємо розбиття S_i гіпотетичного простору можливих значень змінної y на $r+1$ частину. Визначаємо кількості m_j попадань елементів цієї вибірки в кожну комірку S_i розбиття. Якщо деякі $m_j < 5$ то відповідні комірки об'єднуються з сусідніми так, щоб $m_j > 5$. При цьому змінюється відповідне число ступенів свободи r.

Обчислюємо емпіричне значення $\chi^2_{емп}$ статистики К. Пірсона. За вибраного рівня значущості α та числа ступенів свободи r знаходимо з таблиці "Квантилі

статистики ” критичне значення $\chi^2_{кр}$ статистики χ^2. При написанні програм використовують апроксимації даної таблиці, наприклад апроксимація Голдштейна, формули якої наведено нижче. Якщо $\chi^2_{емп} > \chi^2_{кр}$, то гіпотезу H_o відкидаємо, а якщо $\chi^2_{емп} < \chi^2_{кр}$, то кажемо, що дані вибірки за рівня значущості α не суперечать висунутій гіпотезі H_o, тобто належать до одного закону розподілу.

Апроксимація Голдштейна полягає в такому алгоритмі:

$$\chi^2_{\alpha,n} = n\left[\sum_{i=0}^{6} n^{-\frac{i}{2}} \cdot d^i \cdot \left(a_i + \frac{b_i}{n} + \frac{c_i}{n^2}\right)\right]^3, \qquad (6.3)$$

де $d = 2.0637 \cdot \left(\ln\frac{1}{1-\alpha} - 0.16\right)^{0.4274} - 1.5774$ при $0.5 \le \alpha \le 0.999$;

$d = -2.0637 \cdot \left(\ln\frac{1}{\alpha} - 0.16\right)^{0.4274} + 1.5774$ при $0.001 \le \alpha \le 0.5$.

Коефіцієнти a, b, c наведені у таблиці:

a	b	c
1.0000886	-0.2237368	-0.01513904
0.4713941	0.02607083	-0.008986007
0.0001348028	0.01128186	0.02277679
-0.008553069	-0.01153761	-0.01323293
0.00312558	0.005169654	-0.006950356
-0.0008426812	0.00253001	0.001060438
0.00009780499	-0.001450117	0.001565326

Приклади розв’язування задач

Приклад 1. Для упорядкування графіка роботи співробітників досліджувана кількість клієнтів, що звернулися у фірму в першій половині робочого дня. Результати дослідження числа клієнтів залежно від часу роботи подані в табл. 6.1.

Таблиця 6.1

Номер інтервалу часу, i	1	2	3	4
Часи роботи (a_i; a_{i+1})	9 – 10	10 – 11	11 – 12	12 – 13
Кількість клієнтів ($\boldsymbol{n}_i$)	6	20	35	15

При рівні значущості $\alpha = 0{,}05$ потрібно перевірити гіпотезу H_0 про те, що кількість клієнтів, які звернулися у відзначені часи роботи, має нормальний розподіл.

Розв'язування. У даному випадку значеннями безперервної випадкової величини ξ є відліки моментів часу (x_i) *приходу клієнтів* в офіс фірми, які (відліки) можуть потрапити в різні часи роботи і сформувати в такий спосіб кількість значень випадкової величини ξ, що потрапили у від-повідний інтервал часу. Ці значення вже згруповані за $m = 4$ інтервалами, що дозволяє вважати середину кожного інтервалу – представницьким значенням усіх значень випадкової величини, що потрапили в цей інтервал. Процес розв'зання оформимо у вигляді таблиці (див. табл. 6.2).

Відзначимо, що застосування критерію Пірсона потребує, щоб частоти були не менше 5, тобто $n > 5$. В іншому випадку необхідно об'єднати інтервали до одержання необхідних частот. У випадку даного прикладу такої необхідності немає.

1. За варіантним рядом (див. табл. 6.1) знаходимо "*реальні*" оцінки невідомих параметрів *припущеного* теоретичного, у даному випадку – нормального закону розподілу $F(x)$. Для цього спочатку знайдемо загальну кількість клієнтів, що відвідали офіс фірми (див. табл. 6.2, рядок 3):

$$n = \sum_{i=1}^{4} n_i = 76.$$

Далі знайдемо середні (представницькі) значення ($\boldsymbol{x}_{icp}$) для кожного інтервалу (див. табл. 6.2, рядок 4). Потім домножимо кожне з цих середніх значень ($\boldsymbol{x}_{icp}$) на кількість випадкових величин ($\boldsymbol{n}_i$), що потрапили в цей інтервал (див. табл. 6.2, рядок 5), і знайдемо їх суму, яка дозволяє розрахувати оцінку математичного сподівання моментів часу появи клієнтів в офісі фірми:

$$\bar{x} = \tilde{m} = \frac{1}{n}\sum_{i=1}^{4} x_{icp} \cdot n_i = \frac{857}{76} = 11{,}276.$$

Для оцінки середнього *квадратичного* відхилення (S_x) моментів часу появи клієнтів спочатку (див. табл. 6.2, рядок 6) домножимо *квадрат* кожного із середніх значень інтервалів на кількість випадкових величин, які потрапили в цей інтервал ($\boldsymbol{x}^2_{icp} \cdot \boldsymbol{n}_i$), і знайдемо їх суму, що дозволяє розрахувати оцінку другого початкового моменту випадкового часу появи клієнтів в офісі фірми:

$$\tilde{M}[x^2] = \frac{1}{n}\sum_{i=1}^{4} x^2_{icp} \cdot n_i = \frac{9719}{76} = 127{,}882.$$

Потім знайдемо оцінки дисперсії зсунутої і виправленої:

$$\tilde{D}_в = \tilde{M}[x^2] - \tilde{m}_x^2 = 127{,}882 - 127{,}148 = 0{,}726;$$

$$\widetilde{D}_{в.\,випр} = \frac{n}{n-1} \cdot \widetilde{D}_{в.} = \frac{4}{3} \cdot 0{,}726 = 0{,}968\,,$$

що дозволяє одержати оцінку виправленого (незсуненого) середнього квадратичного відхилення:

$$S_x = \sigma_{x.\,випр} = \sqrt{\widetilde{D}_{в.\,випр}} = \sqrt{0{,}968} = 0{,}984\ .$$

2. Визначаємо *теоретичні* частоти на основі припущеного закону $F(x)$ розподілу в такий спосіб.

В даному випадку $\boldsymbol{\xi}$ – моменти часу приходу клієнтів в офіс фірми – безперервна випадкова величина, тому обчисляємо ***ймовірності влучення*** значень x_i випадкової величини $\boldsymbol{\xi}$ *у* кожний *і*-й ***інтервал***, тобто $p_i = P(a_i < x_i < a_{i+1}) = F(a_{i+1}) - F(a_i)$, для чого скористаємося функцією Лапласа і відомим виразом оцінки ймовірності влучення випадкової величини на інтервал (*a*; *b*):

$$P(a < x < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right),$$

одержимо формулу:

$$P(a_i < x < a_{i+1}) = \Phi\left(\frac{a_{i+1} - \widetilde{m}}{\sigma_{x.\,випр}}\right) - \Phi\left(\frac{a_i - \widetilde{m}}{\sigma_{x.\,випр}}\right). \qquad (4.4)$$

У (4.4) аргументами є нормовані границі інтервалів. Тому далі знайдемо значення нормованих границь цих інтервалів (див. табл. 6.2, рядки 7, 8) і, скориставшись таблицею $\Phi(x)$ – значень функції Лапласа (додаток 9), запишемо відповідні значення (див. табл. 6.2, рядки 9 і 10). Відзначимо, що для невід'ємного значення аргументу варто використовувати властивість непарності функції: $\Phi(-x) = -\Phi(x)$. У наступному рядку знайдемо різницю значень функції Лапласа для кожного інтервалу, що і є теоретичною оцінкою ймовірності (p_i) влучення значень випадкової величини в кожний інтервал (див. табл. 6.2, рядок 11).

Таблиця 6.2.

Схема оцінки правдоподібності гіпотези про нормальний розподіл безперервної випадкової величини

№ п/п	Параметри, що розраховуються	***i***				**Σ**	M[*]
		1	2	3	4		
1	a_i	9	10	11	12		
2	a_{i+1}	10	11	12	13		
3	$\boldsymbol{n}_i$	6	20	35	15	76	
4	$\boldsymbol{x}_{i\text{ср}}$	9,5	10,5	11,5	12,5		
5	$\boldsymbol{x}_{i\text{ср}} \cdot \boldsymbol{n}_i$	57	210	402,5	187,5	857	11,276

6	$\boldsymbol{x}^2_{\text{icp}} \cdot \boldsymbol{n}_i$	541,5	2205	4629	2344	9719	127,882
7	$x^н_{min.i} = (a_i - \bar{x})/S_x$	– 2,313	– 1,297	– 0,281	0,735	$\widetilde{\boldsymbol{D}}_{\boldsymbol{в}}$ =	**0,726**
8	$x^н_{max.i} = (a_{i+1} - \bar{x})/S_x$	– 1,297	– 0,281	0,735	1,752	$\widetilde{\boldsymbol{D}}_{\boldsymbol{в.випр}}$	**0,968**
9	$\Phi(x^н_{min.i})$	– 0,490	– 0,403	– 0,111	0,269	$\boldsymbol{S_x}$ =	**0,984**
10	$\Phi(x^н_{max.i})$	– 0,403	– 0,111	0,269	0,460		
11	$\boldsymbol{p}_i$	0,087	0,292	0,380	0,191	0,950	
12	$\boldsymbol{n}_i^{\text{T}}$	6,609	22,201	28,843	14,526		
13	$\boldsymbol{n}_i - \boldsymbol{n}_i^{\text{T}}$	– 0,609	– 2,201	6,157	0,474		
14	$(\boldsymbol{n}_i - \boldsymbol{n}_i^{\text{T}})^2 / \boldsymbol{n}_i^{\text{T}}$	0,056	0,218	1,314	0,016	**1,604 =**	$\chi^2_{спост}$

Потім за знайденими ймовірностями p_i і обсягом вибірки $\boldsymbol{n} = 71$ розрахуємо (див. табл. 6.2, рядок 12) значення ***теоретичних*** частот n_i^T за формулою:

$$n_i^T = n \cdot p_i .$$

3. За заданими в умові емпіричними частотами $\boldsymbol{n}_i$ (див. табл. 6.2, рядок 3) і отриманими теоретичними частотами n_i^T (див. табл. 6.2, рядок 12) обчислюємо величину $\chi^2_{спост}$. Для цього попередньо для кожного інтервалу знайдемо (див. табл. 6.2, рядок 13) різницю між емпіричною і теоретичною частотою влучення випадкової величини в цей інтервал: $(n_i - n_i^T)$. Потім знайдемо *окремі доданки* (див. табл. 6.2, рядок 14) для визначення значення показника "Хі-квадрат":

$$\chi_i^2 = \frac{(n_i - n_i^T)^2}{n_i^T}$$

і потім розрахуємо суму цих значень, що і є шуканою величиною для оцінки правдоподібності початкової гіпотези:

$$\chi^2_{спост} = \sum_{i=1}^{4} \frac{(n_i - n_i^T)^2}{n_i^T} = 1,604 .$$

У вибірці з $m = 4$ інтервалів оцінювалися *два* ($k = 2$) параметри – $\bar{x} = \tilde{m}$ і S_x, тому кількість ступенів свободи r дорівнюватиме:

$$r = m - k - 1 = 4 - 2 - 1 = 1.$$

За таблицею (додаток 10) знаходимо критичне значення:

$$\chi^2_{кр}(\alpha, r) = \chi^2_{кр}(0{,}05; 1) = 3{,}84; \qquad \chi^2_{спост} = 1{,}604 .$$

Значення $\chi^2_{спост} = 1{,}604$, тобто $\chi^2_{спост} < \chi^2_{кр}(\alpha, r)$, тому **гіпотеза** H_0 про те, що кількість клієнтів (випадкова величина ξ) має нормальний розподіл – **приймається**.

Приклад 2. При $\boldsymbol{n} = 757$ випробувань блоків радіоелектронної апаратури отримані і наведені в табл. 4.11 дані про кількість відмов на один блок.

Таблиця 6.3

Кількість відмов (x_i)	0	1	2	3	4	5	≥6
Кількість випадків ($\boldsymbol{n}_i$)	427	235	72	21	1	1	0

За рівнем значущості $\alpha = 0{,}05$ потрібно перевірити гіпотезу H_0 про те, що кількість відмов має розподіл Пуассона:

$$P_k = \frac{\lambda^k}{k!} \cdot e^{-\lambda} .$$

Розв'язування. **1.** Знайдемо оцінку параметра λ розподілу Пуассона, яка дорівнює середньому числу відмов. Процес розв'зання оформимо у вигляді розрахункових таблиць 6.4–6.6.

Із табл. 6.2 знаходимо середнє число відмов: $\bar{x} = \dfrac{\sum\limits_{i=1}^{7} x_i n_i}{\sum\limits_{i=1}^{7} n_i} = \dfrac{451}{757} = 0{,}6 .$

Отже, теоретичний закон розподілу числа відмов блоків апаратури має вигляд:

$$p_i = p(\xi = x_i) = \frac{(0{,}6)^{x_i}}{x_i!} e^{-0{,}6}, \qquad x_i = 0, 1, \ldots, 5 .$$

Таблиця 6.4 *Таблиця* 6.5 *Таблиця* 6.6

i	x_i	$\boldsymbol{n}_i$	$x_i \boldsymbol{n}_i$	p_i	$\boldsymbol{n}p_i$	$\boldsymbol{n}^{\mathrm{T}}{}_i$	$\boldsymbol{n}_i$	$\boldsymbol{n}^{\mathrm{T}}{}_i$	$(\boldsymbol{n}_i - \boldsymbol{n}^{\mathrm{T}}{}_i)^2$	$(\boldsymbol{n}_i - \boldsymbol{n}^{\mathrm{T}}{}_i)^2 / \boldsymbol{n}^{\mathrm{T}}{}_i$
1	0	427	0	0,5488	415,4504	416	427	416	121	0,291
2	1	235	235	0,3293	249,2702	249	235	249	196	0,787
3	2	72	144	0,0988	74,78107	75	72	75	9	0,120
4	3	21	63	0,0198	14,95621	15	23	17	36	2,118
5	4	1	4	0,0030	2,243432	2				
6	5	1	5	0,0004	0,269212	0				
7	6	0	0							
Σ		757	451			757			$\chi^2_{спост} = 3{,}316$	

2. На основі отриманого теоретичного закону розподілу знаходимо теоретичні частоти $\boldsymbol{n}^{\mathrm{T}}{}_i$, для чого складемо розрахункову табл. 6.5. Ймовірність p_i знаходимо або за таблицею, або обчисляємо за формулою Пуассона для всіх значень x_i.

3. За заданими в умові емпіричними частотами $\boldsymbol{n}_i$ і отриманими теоретичними частотами $\boldsymbol{n}^{\mathrm{T}}{}_i$ обчисляємо величину $\chi^2_{спост}$:

$$\chi^2_{спост.} = \sum_{i=1}^{m} \frac{(n_i - n_i^T)^2}{n_i^T}.$$

Для цього складемо розрахункову таблицю (див. табл. 6.5), об'єднавши останні три рядки з рядком для $x_i = 3$, тому що застосування критерію Пірсона потребує, щоб частоти не були малими.

У вибірці оцінювався *один* параметр λ, тому число ступенів свободи r дорівнюватиме:

$$r = \boldsymbol{m} - k - 1 = 4 - 1 - 1 = 2.$$

За таблицею (див. додаток 10) знаходимо критичне значення:

$$\chi^2_{кр}(\alpha, r) = \chi^2_{кр}(0{,}05;\ 2) = 5{,}99.$$

Отримане в спробі $\chi^2_{спост.} = 3{,}316$, і це значення виявляється меншим критичного, тобто $\chi^2_{спост.} < \chi^2_{кр}(\alpha, r)$, тому **гіпотеза** H_0 про розподіл кількості відмов блоків апаратури за законом Пуассона **приймається**.

Хід роботи

Організувати ввід з консолі $(N + 20)$ цілих чисел в діапазоні від 0 до 20 включно. Запрограмувати перевірку на допустимі значення. Запрограмувати перевірку апроксимації Голдштейна для різних значень α (контрольні значення – додаток А).

Ввести числа вручну випадковим чином. Організувати перевірку програмним чином на основі критерію χ^2 на відповідність введеної вибірки нормальному та рівномірному розподілу за двома різними рівнями значущості.

Вивести на екран вибірку та результати перевірок.

Контрольні запитання

1. Критерій узгодженості Пірсона.
2. Яким чином організована перевірка даних на допустимість.
3. Водоспадна модель ЖЦ.
4. Спіральна модель ЖЦ.

ЛАБОРАТОРНА РОБОТА № 7

Тема: проста лінійна регресія.

Мета: закріпити теоретичні знання та отримати практичні навички щодо використання ПЕОМ з метою оцінки параметрів та підбору емпіричних формул за допомогою простої вибіркової лінійної регресії методом найменших квадратів.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

Вибіркове рівняння прямої лінії регресії

Математична формалізація реальних процесів у вигляді функціональної залежності однієї змінної (функції y) від іншої змінної (аргументу x) $y = f(x)$ дозволяє кожному значенню аргументу однозначно поставити у відповідність одне заздалегідь відоме значення функції. Така залежність, як правило, є оборотною, тобто за відомим значенням функції завжди можна знайти відповідне значення її аргументу: $x = f^{-1}(y)$.

Проте на практиці реальні процеси демонструють іншу відповідність змінних. Так за один і той же робочий час тим самим підприємством може бути зроблена різна кількість продукції, та сама продукція може мати різну собівартість, та сама кількість продукції може бути продана за різну ціну, люди того самого зросту можуть мати різну вагу.

У підсумку, *тому* самому *значенню* аргументу x можуть відповідати *різні* значення функції y. Часто значення функції носять характер випадкових величин, які при зміні значень аргументу можуть мати різні центри групування, дисперсію і закон розподілу. У цьому випадку зворотне перетворення з метою відшукання значення аргументу x за відомим значенням функції y не є однозначним.

Для встановлення залежності між такими неоднозначними змінними використовують теорію кореляційно-регресійного аналізу, у якій досліджуються зміни середніх значень функції при зміні одного або багатьох аргументів. Проте формальні методи кореляційного аналізу не дають відповіді на питання “що є причиною, а що є наслідком або яка змінна є аргументом, а яка – функцією”? Відповідь на це питання має дати дослідник.

Метою побудови регресійних моделей може бути встановлення залежності між *середніми* значеннями двох змінних (параметрів), одну з яких дослідник *призначає* функцією, а другу – її аргументом.

У основі регресійного аналізу лежать дві гіпотези.

1. Передбачається, що досліджувана сукупність параметрів *має* внутрішній статистичний зв'язок, який може бути виявлений і формалізований у вигляді кореляційної (отже лінійної) залежності одного параметра від іншого або від інших. Тобто вважається, що існує внутрішній лінійний зв'язок *середніх* значень цих параметрів.

2. Передбачається, що випадковий розкид (дисперсія) значень (кожного) параметру має регулярну компоненту, яка залежить від деякого аргументу ("сигналу"), і випадкову компоненту ("шум"). Випадкова компонента ("шум") розподілена за нормальним законом.

Початковою інформацією для побудови лінійної однофакторної регресійної моделі є сукупність із n двовимірних точок (x_i, y_i), де кожна координата точки, як правило, має свій фізичний сенс, наприклад x_i – зріст людини в сантиметрах, y_i – її вага в кілограмах.

Під час формалізації постановки задачі розглянемо двовимірну випадкову величину (X, Y), над якою проведено $\boldsymbol{n}$ незалежних випробувань і в результаті випробувань отримана вибірка – $\boldsymbol{n}$ пар чисел (координат точки):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

де x_i – значення випадкової величини X у i-му випробуванні;

y_i – значення випадкової величини Y у i-му випробуванні.

Необхідно знайти *наближене* зображення значень однієї з випадкових величин як функції значень другої випадкової величини.

Вибірковим *рівнянням регресії* $\boldsymbol{Y}$ на $\boldsymbol{X}$ $(y \to x)$ називається рівняння, яке встановлює залежність змінної y від змінної x, тобто коли змінна y **вважається** *функцією*, а змінна x – *аргументом*: $y = f(x)$, при цьому початковою інформацією є вибірка з $\boldsymbol{n}$ пар чисел.

Вибірковим рівнянням регресії $\boldsymbol{X}$ ***на*** $\boldsymbol{Y}$ $(x \to y)$ називається рівняння $x = \varphi(y)$, у якому при тій же початковій інформації вже змінна $\boldsymbol{x}$ уважається *функцією*, а змінна $\boldsymbol{y}$ – її *аргументом*.

Лінійною *називається регресія* у випадку, коли залежності $f(x)$ і $\varphi(y)$ є *лінійними функціями*. Тоді рівняння регресії мають вигляд:

$$y = a \cdot x + b; \qquad x = c \cdot y + d.$$

Порядок розрахунку параметрів рівняння регресії розглянемо без виведення і для двох варіантів умов: при відсутності і при наявності збіжних точок у вибірці.

1. Нехай серед точок (x_i, y_i) вибірки збіжних точок немає. Для того щоб скласти вибіркове рівняння прямої лінії регресії, виконуються наступні розрахунки:

а) обчислюються середні значення величин: $\bar{x}$, $\bar{y}$, $\overline{x \cdot y}$, $\overline{x^2}$, $\overline{y^2}$ і знаходяться середні квадратичні відхилення $S_x = \sigma_x$, $S_y = \sigma_y$ із використанням формул, перерахованих з обліком доцільної послідовності їх застосування:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ; \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 ; \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 ;$$

$$\overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i ; \quad S_x^2 = \overline{x^2} - \left(\bar{x}\right)^2 ; \quad S_x = \sqrt{S_x^2} ; \quad S_y^2 = \overline{y^2} - \left(\bar{y}\right)^2 ; \quad S_y = \sqrt{S_y^2} ;$$

б) обчислюється значення вибіркового коефіцієнту кореляції r:

$$r = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y} .$$

Нагадаємо, що вибірковий коефіцієнт кореляції характеризує рівень лінійного кореляційного зв'язку двох випадкових величин. Чим ближче $|r|$ до одиниці, тим більш сильним є зв'язок двох величин, чим ближче $|r|$ до нуля, тим зв'язок слабше (див. п. 3.2.12);

в) для одержання рівняння регресії Y на X: $y = a \cdot x + b$ обчислюємо коефіцієнти a і b даного рівняння за формулами:

$$a = r\frac{S_y}{S_x} ; \quad b = \bar{y} - a \cdot \bar{x} .$$

Для одержання рівняння регресії X на Y: $x = c \cdot y + d$ обчислюємо коефіцієнти c і d даного рівняння за формулами:

$$c = r\frac{S_x}{S_y} ; \quad d = \bar{x} - c \cdot \bar{y} .$$

Обидві прямі лінії $y = a \cdot x + b$ і $x = c \cdot y + d$ проходять через точку $(\bar{x}, \bar{y})$. Для зображення обох прямих ліній на одному графіку друге рівняння варто подати у вигляді: $y = x/c - d/c$.

2. При великій кількості точок $\boldsymbol{n}$ у вибірці значення x_i може зустрітися $\boldsymbol{m}_i$ разів, значення y_j може зустрітися $\boldsymbol{n}_j$ разів. Та сама пара чисел (x_i, y_j) може зустрітися $\boldsymbol{n}_{ij}$ разів. У цьому випадку вибірку зручно подати у вигляді кореляційної таблиці (див. табл. 4.15) так, що кількість повторень $\boldsymbol{m}_i$ значень координат x_i, кількість повторень $\boldsymbol{n}_j$ значень координат y_j і обсяг вибірки $\boldsymbol{n}$ дорівнюватимуть:

$$m_i = \sum_{j=1}^{\ell} n_{ij} , \quad n_j = \sum_{i=1}^{k} n_{ij} ; \qquad n = \sum_{i=1}^{k}\sum_{j=1}^{\ell} n_{ij} = \sum_{j=1}^{\ell} n_j = \sum_{i=1}^{k} m_i = n .$$

Таблиця 7.1

Кореляційна таблиця

x_i	Y_j				
	y_1	y_2	⋯	y_ℓ	$\boldsymbol{m}_i$
x_1	$\boldsymbol{n}_{11}$	$\boldsymbol{n}_{12}$	…	$\boldsymbol{n}_{1\ell}$	$\boldsymbol{m}_1$
x_2	$\boldsymbol{n}_{21}$	$\boldsymbol{n}_{22}$	…	$\boldsymbol{n}_{2\ell}$	$\boldsymbol{m}_2$
…	…	…	…	…	…
x_k	$\boldsymbol{n}_{k1}$	$\boldsymbol{n}_{k2}$	…	$\boldsymbol{n}_{k\ell}$	$\boldsymbol{m}_k$
$\boldsymbol{n}_\text{j}$	$\boldsymbol{n}_1$	$\boldsymbol{n}_2$	…	$\boldsymbol{n}_\ell$	$\boldsymbol{n}$

Вибіркове рівняння прямої лінії регресії знаходимо аналогічно першому випадку, розрахунки величин $\bar{x}$, $\bar{y}$, $\overline{x \cdot y}$, $\overline{x^2}$, $\overline{y^2}$ виконуємо з урахуванням наявності повторюваних значень змінних за формулами:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot m_i ; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{\ell} y_i \cdot n_j ;$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot m_i ; \qquad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{\ell} y_i^2 \cdot n_j ; \qquad \overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{\ell} x_i \cdot y_i \cdot n_{ij} .$$

Якщо розглядається вибірка з генеральної сукупності безперервних випадкових величин X і Y, то кореляційна таблиця буде містити інтервали $[a_{i-1}, a_i)$ і $[b_{j-1}, b_j)$.

У цьому випадку для обчислення величин $\bar{x}$, $\bar{y}$, $\overline{x \cdot y}$, $\overline{x^2}$, $\overline{y^2}$ спочатку потрібно перейти до дискретних рядів, а потім виконати обчислення за розглянутими вище формулами.

Зауваження. Якщо значення координат точок (x_i, y_j) є дуже великими або дуже маленькими числами, то при обчисленні значень величин $\bar{x}$, $\bar{y}$, $\overline{x \cdot y}$, $\overline{x^2}$, $\overline{y^2}$ можна спочатку перейти до розглянутих раніше *умовних варіантів* u_i і v_j:

$$u_i = \frac{x_i - \alpha}{p}, \quad v_j = \frac{y_j - \beta}{q}$$

і знайти їх середні значення $\bar{u}$, $\bar{v}$ і S_u, S_v, а потім знайти *середні* значення початкових координат $\bar{x}$, $\bar{y}$ та їх середніх квадратичних відхилень S_x, S_y:

$$\bar{x} = p \cdot \bar{u} + \alpha , \qquad \bar{y} = q \cdot \bar{v} + \beta ; \qquad S_x = p \cdot S_u , \qquad S_y = q \cdot S_v .$$

Перехід до умовних варіантів не змінює величину вибіркового коефіцієнту кореляції, тому $r_{xy} = r_{uv} = r$.

Одне з головних завдань регресійного аналізу полягає у підборі відповідного виразу Y=f(X), графік якого проходить через емпіричні точки або близько від них. І таким чином зв'язані змінні X та Y. Даний вираз має назву рівняння регресії, функція f(X) називається функцією регресії. Графік даної

функції називається лінією даної регресії. Тобто регресійний аналіз виявляє кількісну залежність ознаки фактору від іншої ознаки.

Приклад. Знайти вибіркові рівняння прямих ліній регресії $\boldsymbol{Y}$ на $\boldsymbol{X}$ і $\boldsymbol{X}$ на $\boldsymbol{Y}$ за даними таблиці спостережень (див. табл. 7.2).

Таблиця 7.2

Таблиця спостережень

x	1,00	1,50	3,00	4,50	5,00
y	1,25	1,40	1,50	1,75	2,25

Розв'язування. Складемо розрахункову таблицю (див. табл. 7.3).

Таблиця 7.3

Розрахункова таблиця

i	x_i	y_j	x_i^2	y_j^2	$x_i\ y_j$
1	**1,00**	**1,25**	1,00	1,56	1,25
2	**1,50**	**1,40**	2,25	1,96	2,10
3	**3,00**	**1,50**	9,00	2,25	4,50
4	**4,50**	**1,75**	20,25	3,06	7,88
5	**5,00**	**2,25**	25,00	5,06	11,25
$\sum$	15,00	8,15	57,50	13,90	26,98
$\sum/\boldsymbol{n}$	**3**	**1,63**	**11,5**	**2,779**	**5,395**

За допомогою таблиці знаходимо оцінки середніх значень:

$$\bar{x}=\frac{1}{5}\sum_{i=1}^{5}x_i=\frac{15}{5}=3;\quad \bar{y}=\frac{1}{5}\sum_{j=1}^{5}y_j=\frac{8{,}15}{5}=1{,}63;\quad \overline{xy}=\frac{1}{5}\sum_{i=1}^{5}x_i\cdot y_i=\frac{26{,}975}{5}=5{,}395;$$

$$\overline{x^2}=\frac{1}{5}\sum_{i=1}^{5}x_i^2=\frac{57{,}5}{5}=11{,}5;\quad \overline{y^2}=\frac{1}{5}\sum_{i=1}^{5}y_i^2=\frac{13{,}897}{5}=2{,}779.$$

Потім визначаємо значення середніх квадратичних відхилень:

$$S_x=\sqrt{\overline{x^2}-\left(\bar{x}\right)^2}=\sqrt{11{,}5-9}=\sqrt{2{,}5}=1{,}58;$$

$$S_y=\sqrt{\overline{y^2}-\left(\bar{y}\right)^2}=\sqrt{2{,}779-2{,}657}=\sqrt{0{,}122}=0{,}35.$$

Знайдемо вибірковий коефіцієнт кореляції:

$$r = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y} = \frac{5{,}395 - 3 \cdot 1{,}63}{1{,}58 \cdot 0{,}35} = \frac{0{,}505}{0{,}553} = 0{,}913.$$

Для вибіркового рівняння прямої лінії регресії ***Y*** на ***X***, що має вигляд $y = a \cdot x + b$, обчислимо значення коефіцієнтів a і b за відомими формулами, одержимо:

$$a = r \frac{S_y}{S_x} = 0{,}913 \cdot \frac{0{,}35}{1{,}58} = 0{,}202;$$

$$b = \bar{y} - a \cdot \bar{x} = 1{,}63 - 0{,}202 \cdot 3 = 1{,}63 - 0{,}606 = 1{,}024.$$

Остаточно вибіркове рівняння прямої лінії регресії ***Y*** на ***X*** набуде вигляду:

$$y = 0{,}202 \cdot x + 1{,}024.$$

Для одержання рівняння регресії ***X*** на ***Y***: $x = c \cdot y + d$ обчисляємо коефіцієнти c і d за формулами:

$$c = r \frac{S_x}{S_y} = 0{,}913 \cdot \frac{1{,}58}{0{,}35} = 4{,}122; \quad d = \bar{x} - c \cdot \bar{y} = 3 - 4{,}122 \cdot 1{,}63 = -3{,}719.$$

Вибіркове рівняння прямої лінії регресії ***X*** на ***Y*** набуде вигляду:

$$x = 4{,}122 \cdot y - 3{,}719.$$

Для побудови даного рівняння в тій же системі координат, що і рівняння регресії ***Y*** на ***X***, скористаємося варіантом перерахунку $y = x/c - d/c$ і виразимо змінну y через змінну x, одержимо:

$$y = 0{,}242 \cdot x + 0{,}902.$$

Метод найменших квадратів

У ряді практичних задач обробки результатів вимірів, у тому числі значень випадкових величин, виникає необхідність згладженого подання значень однієї величини (наприклад, величини *y),* як функції іншої величини (наприклад, величини *x)*. Одним із найбільш поширених способів такого наближеного зображення є метод найменших квадратів.

За допомогою методу найменших квадратів розв'язується задача добору такої аналітичної залежності $y(x) = \Psi_m(x, a_0, a_1, \ldots, a_m)$, графік якої *не обов'язково* проходив би через *усі* задані точки, але максимально "згладжував" би випадкові похибки вимірюваних ординат функції $y_i = f(x_i)$ $(i = 0, 1, 2, \ldots, n)$, тобто щоб сума квадратів відхилень значень аналітичної залежності $\Psi_m(x, a_0, a_1, \ldots, a_m)$ від значень вимірюваних ординат $y_i = f(x_i)$ у цих точках була мінімальною:

$$S = \sum_{i=1}^{n} [y_i - \Psi_m(x_i)]^2 \Rightarrow \min . \Psi_m(x_i) \qquad (7.1)$$

Апроксимація за методом найменших квадратів виконується у два етапи:

– *на першому* етапі вибирають *вигляд* $\Psi_m(x, a_0, \ldots, a_m)$ шуканої формули;

– *на другому* етапі для формули обраного вигляду "підбирають" *значення параметрів* $a_0, a_1, \ldots, a_m$, виходячи з вимоги (4.20).

Процес добору полягає в одержанні системи з $(m+1)$-го рівняння для визначення значень усіх $(m+1)$ параметрів $a_0, a_1, \ldots, a_m$.

Виходячи з вимоги мінімізації відхилення значень $\Psi_m(x_i)$ від значень $y_i = f(x_i)$, система рівнянь формується шляхом відшукання похідних за кожним параметром (a_k) від рівняння (4.20) і прирівнювання їх до нуля:

$$\frac{\partial S}{\partial a_k} = -2\sum_{i=1}^{n} [y_i - \Psi_m(x_i, a_0, \ldots, a_k, \ldots, a_m)] \cdot \left(\frac{d\,\Psi_m}{da_k} \right)_{x_i} = 0; \; k = 0, 1, \ldots, m .$$

Розділивши ліву і праву частини на множник (–2), який "заважає", одержимо:

$$\sum_{i=1}^{n} [y_i - \Psi_m(x_i, a_0, \ldots, a_k, \ldots, a_m)] \cdot \left(\frac{d\,\Psi_m}{da_k} \right)_{x_i} = 0; \quad k = 0, 1, \ldots, m, \quad (7.2)$$

де $\left(\frac{d\,\Psi_m}{da_k} \right)_{x_i} = \Psi'_{a_k}(x_i, a_0, \ldots, a_k, \ldots, a_m)$ – значення часткової похідної від функції Ψ_m за параметром a_k у точці x_i.

Для розв'зання системи рівнянь (7.2) потрібно задатися конкретним видом апроксимуючої функції Ψ_m.

Лінійна апроксимація $(\Psi_1 = a_0 + a_1 \cdot x)$

У цьому випадку $\Psi'_{a_0} = 1; \; \Psi'_{a_1} = x$. Тоді рівняння (4.21) набуде вигляду:

$$\sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)] = 0 ; \qquad \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)] x_i = 0 .$$

У цьому рівнянні розкриємо дужки і зробимо підсумовування, що дозволяє одержати:

$$\sum_{i=1}^{n} y_i - n \cdot a_0 - a_1 \sum_{i=1}^{n} x_i = 0 ; \qquad \sum_{i=1}^{n} x_i y_i - a_0 \sum_{i=1}^{n} x_i - a_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Розділивши ліву і праву частини кожного рівняння на кількість точок n, одержимо:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - a_0 - a_1 \frac{1}{n}\sum_{i=1}^{n} x_i = 0; \qquad \frac{1}{n}\sum_{i=1}^{n} x_i y_i - a_0 \frac{1}{n}\sum_{i=1}^{n} x_i - a_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0.$$

Доданки в цих виразах дорівнюють:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = m_y = \overline{y}; \ \frac{1}{n}\sum_{i=1}^{n} x_i = m_x = \overline{x}; \ \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \alpha_{11} = \overline{xy}; \ \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \alpha_{2x} = \overline{x^2}$$

і мають сенс оцінок математичних сподівань (m_x, m_y), змішаного початкового другого моменту (α_{11}) і початкового другого моменту (α_{2x}), тому система рівнянь для пошуку значень коефіцієнтів a_0 і a_1 може бути записана більш компактно:

$$\left.\begin{aligned} m_y - a_0 - a_1 \cdot m_x = 0, \\ \alpha_{11} - a_0 \cdot m_x - a_1 \cdot \alpha_{2x} = 0, \end{aligned}\right\} \quad \rightarrow \quad \left.\begin{aligned} \overline{y} - a_0 - a_1 \cdot \overline{x} = 0, \\ \overline{xy} - a_0 \cdot \overline{x} - a_1 \cdot \overline{x^2} = 0. \end{aligned}\right\}.$$

Звідки випливає, що в апроксимуючому поліномі $\Psi_1 = a_0 + a_1 x$ коефіцієнти a_0 і a_1 можна визначити за формулами:

$$a_1 = \frac{\alpha_{11} - m_x \cdot m_y}{\alpha_{2x} - m_x^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2}; \qquad a_0 = m_y - a_1 \cdot m_x = \overline{y} - a_1 \cdot \overline{x}.$$

Аналогічно знаходяться параметри і для апроксимуючих функцій будь-якого іншого вигляду. Проте загального правила для вибору відповідного вигляду емпіричної формули не існує. Водночас, для найбільш поширених семи видів функцій існує процедура формальної оцінки їх придатності.

Параметри або коефіцієнти емпіричних моделей можуть встановлюватись методом найменших квадратів і методом максимуму правдоподібності.

Типове і основне завдання формулюють таким чином:

Необхідно встановити параметри a_0 і a_1 для залежності

$$\hat{y} = a_0 + a_1 X$$

на вибірці об`єму n.

У методі найменших квадратів дані параметри визначаються з умови мінімуму наступного критерію

$$R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 \times x_i)^2,$$

на основі чого було одержано наступні формули:

$$a_0 = \frac{\sum_{i=1}^{n} y_i \times \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i \cdot y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)},$$

$$a_1 = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)}.$$

Приклад. Для умов, розглянутих у попередньому прикладі (див. табл. 7.2), знайти лінійну апроксимацію залежності змінної ***y*** від змінної ***x*** за даними тієї ж таблиці спостережень, що повторюємо (див. табл. 7.3).

Таблиця 7.4

x	1,00	1,50	3,00	4,50	5,00
y	1,25	1,40	1,50	1,75	2,25

Розв'язування. Отримана раніше розрахункова таблиця має вигляд (див. табл. 7.5):

Таблиця 7.5

i	x_i	y_j	x_i^2	y_j^2	$x_i y_j$
1	**1,00**	**1,25**	1,00	1,56	1,25
2	**1,50**	**1,40**	2,25	1,96	2,10
3	**3,00**	**1,50**	9,00	2,25	4,50
4	**4,50**	**1,75**	20,25	3,06	7,88
5	**5,00**	**2,25**	25,00	5,06	11,25
Σ	15,00	8,15	57,50	13,90	26,98
Σ/*n*	**3**	**1,63**	**11,5**	**2,779**	**5,395**

Необхідні оцінки з таблиці були знайдені і мають такі значення:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{15}{5} = 3;\ \bar{y} = \frac{1}{5}\sum_{j=1}^{5} y_j = \frac{8{,}15}{5} = 1{,}63;\ \overline{xy} = \frac{1}{5}\sum_{i=1}^{5} x_i y_i = \frac{26{,}975}{5} = 5{,}395;$$

$$\overline{x^2} = \frac{1}{5}\sum_{i=1}^{5} x_i^2 = \frac{57{,}5}{5} = 11{,}5;\ \ \overline{y^2} = \frac{1}{5}\sum_{i=1}^{5} y_i^2 = \frac{13{,}897}{5} = 2{,}779.$$

Отже, для лінійної апроксимації $y = a_0 + a_1 x$ знаходимо:

$$a_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \frac{5{,}395 - 3\cdot 1{,}63}{11,5 - (3)^2} = \frac{5{,}395 - 4{,}89}{11,5 - 9} = \frac{5{,}395 - 4{,}89}{11,5 - 9} = \frac{0{,}505}{2,5} = 0,202;$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x} = 1{,}63 - 0{,}202 \cdot 3 = 1{,}63 - 0{,}606 = 1{,}024\ .$$

Результуючий вираз має вигляд: $y = 1{,}024 + 0{,}202 \cdot x$ і цілком збігається з отриманим раніше рівнянням регресії ***Y*** на ***X***:

$$y = 0{,}202 \cdot x + 1{,}024.$$

Графічне зображення цих ліній наведене на рис. 4.8.

Відзначимо, що лінія регресії $y \to x$ у даному випадку має менший кут нахилу, ніж лінія регресії $x \to y$. Читач може самостійно спробувати пояснити причину такого явища.

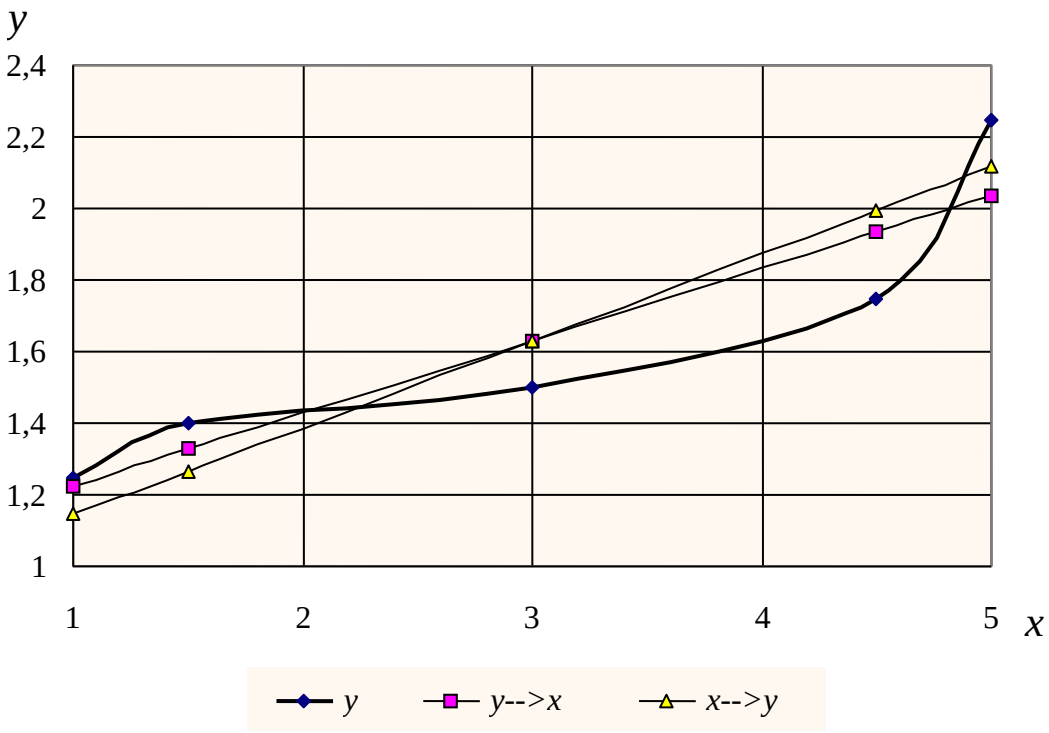


Рис. 7.1. Лінії регресії

Хід роботи

Задатись довільною лінією у формі $y = kx + b$. Одержати дискретне представлення лінії в деякому невід'ємному діапазоні (N+10) точок, де N – номер у журналі. Додати до цих даних "шум" з нормальним розподілом – розмах варіації (N/5). Знайти рівняння регресії на основі "зашумлених" даних, вивести графіки та коефіцієнти вихідної лінії та регресійної моделі.

Важливо знати

Багатьом із тих, хто стикається з науковими та інженерними розрахунками часто доводиться оперувати наборами значень, отриманих експериментальним шляхом чи методом випадкової вибірки. Як правило, на підставі цих наборів потрібно побудувати функцію, зі значеннями якої могли б з високою точністю збігатися інші отримувані значення. Така задача називається апроксимацією кривої. Інтерполяцією називають такий різновид апроксимації, при якій крива побудованої функції проходить точно через наявні точки даних.

Існує також близька до інтерполяції задача, що полягає в апроксимації якої-небудь складної функції іншою, простішою функцією. Якщо деяка функція занадто складна для продуктивних обчислень, можна спробувати обчислити її

значення в декількох точках, а за ними побудувати, тобто інтерполювати, простішу функцію. Зрозуміло, використання спрощеної функції не дозволяє одержати такі ж точні результати, які давала б початкова функція. Але, для деяких класів задач, досягнутий виграш у простоті і швидкості обчислень може переважити отриманий огріх у результатах.

Апроксима́ція (лат. approximate — наближати) — наближене вираження одних математичних об'єктів іншими, близькими за значенням, але простішими, наприклад, кривих ліній — ламаними, ірраціональних чисел — раціональними, неперервних функцій — многочленами.

Екстраполяція – наближення (приближення), знаходження за рядом даних значень функції інших її значень, що містяться поза цим рядом.

Екстраполювати, (рос.экстраполировать, англ. extrapolate, нім. extrapolieren) – поширювати висновки, одержані щодо однієї частини якоїсь системи, на іншу частину тієї самої системи.

Контрольні питання

1. Види програмних вимог.
2. Цілі і задачі регресійного аналізу.
3. Метод найменших квадратів.
4. Що таке інтерполяція.
5. Що таке апроксимація.
6. Що таке екстраполяція.

ЛАБОРАТОРНА РОБОТА № 8

Тема: Елементи кореляційного аналізу та їх програмна реалізація.

Мета роботи. Знайомство з комп´ютерними засобами кореляційно-регресійного аналізу на основі програмних засобів загального призначення, набуття навичок їх практичного застосування на прикладі процесора електронних таблиць (ПЕТ) MS Excel, застосування їх у програмному середовищі.

Технічне забезпечення: ПЕОМ, середовище програмування.

Короткі теоретичні відомості

Кореляційні зв’язки можна аналізувати на якісному рівні з діаграм розсіяння емпіричних значень змінних і відповідним чином їх інтерпретувати. Так, наприклад, якщо підвищення рівня однієї змінної супроводжується підвищення рівня іншої, то йдеться про позитивну кореляцію або прямий зв’язок. Якщо ж зростання однієї змінної супроводжується зниженням значень іншої, то маємо справу з негативною кореляцією.

Нульовою називається кореляція за відсутності зв’язку змінних. Проте нульова загальна кореляція може свідчити лише про відсутність лінійної залежності, а не про відсутність залежності між величинами взагалі. Кількісна міра кореляційного зв’язку оцінюється найчастіше за значеннями коефіцієнта кореляції від +1 до -1. Від’ємні значення коефіцієнта кореляції свідчать про зворотний зв’язок, додатні – про прямий.

Нульове значення може свідчити про відсутність зв’язку. Інтенсивність зв’язку (слабкий – помірний – суттєвий – сильний) оцінюється за абсолютним значенням коефіцієнта кореляції.

Методи розрахунку міри кореляційних зв’язків тісно пов’язані із вживаними вимірюваними шкалами.

Лінійний кореляційний зв’язок для емпіричних даних, виміряних за шкалою інтервалів або відношень, оцінюється за допомогою коефіцієнта кореляції Пірсона r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}, \qquad (8.1)$$

де x_i і y_i – значення змінних, $\bar{x}$ та $\bar{y}$ – серединні значення вибірок, n – обсяг вибірок.

Показник кореляції рангів. У тих випадках, коли одниниці досліджуваної часткової сукупності можуть бути у відношенні деякої ознаки розміщені в певному порядку по зростаючим (або спадаючим) номерам, або рангам, в якості статистики зв’язку служить показник кореляції рангів.

Ранг вказує те місце, яке займає дана одиниця сукупності серед інших одиниць. Якщо б кожна з цих одиниць відрізнялась у відношенні розглядуваної ознаки від всіх інших одиниць сукупності, то ранги являли собою б порядкові номери від 1 до числа n , рівного об'єму сукупності. Якщо ж деякі з одиниць сукупності є однаковими, то ранг всіх цих одиниць приймається середнім з їх відповідніх номерів.

Показник кореляції рангів рівний:

$$\rho = 1 - \frac{6\sum_{h=1}^{n} d_h^2}{n\left(n^2 - 1\right)}, \qquad (8.2)$$

де величини d_h являють собою різницю між рангами h_1 та h_2 одиниць, вибраних разом з двох послідовностей:

$$d_h = h_1 - h_2$$

Показник кореляції рангів змінюється від -1 до +1:

$$-1 \le \rho \le +1$$

Чим тісніший буде зв'язок між величинами, тим ближче до одиниці за своєю абсолютною величиною буде показник кореляції рангів; знак вказує, пряма залежність чи зворотна.

По суті задача виявлення залежності між величинами являє собою задачу перевірки гіпотези про існування залежності, але в силу специфічних особливостей задачі, її розповсюдженості і важливості, її часто розглядають як окрему задачу математичної статистики.

У статистиці розрізняють три види зв'язків між величинами:

- функціональний;
- стохастичний;
- статистичний.

Функціональний зв'язок – це зв'язок, який описується детермінованим співвідношенням виду $Y = f(X_1, X_2, ...X_n)$, так що кожному значенню однієї величини відповідає одне точно визначене значення іншої. Функціональний зв'язок не є предметом статистики.

Стохастичний зв'язок – це зв'язок, при якому зміна значення однієї величини, веде до зміни закону розподілу результативної величини.

Статистичний зв'язок – це зв'язок, при якому значення результативної ознаки змінюються в середньому залежно від того, яких значень набуває факторна величина.

Важливим частинним видом статистичного зв'язку є кореляційний зв'язок, який являє собою залежність умовного математичного сподівання результативної ознаки від значень факторної ознаки:

$$M\left(Y \middle| X = x\right) = \overline{y}(x),$$

тобто кожному значенню факторної ознаки ***x*** відповідає розподіл значень величини Y, причому математичне сподівання цього розподілу є функцією від значення величини X. Кореляційний зв'язок дуже часто зустрічається у всіх сферах діяльності.

Найбільш розробленими і широко вживаними є методи параметричного кореляційного аналізу, які призначені для виявлення і кількісної оцінки кореляційного зв'язку між ознаками досліджуваних об'єктів. Передумовами параметричного кореляційного аналізу є:

- усі спостереження є взаємно незалежними;
- спостереження мають нормальний закон розподілу.

Гіпотеза про існування залежності між величинами X та Y висувається на основі аналізу вигляду експериментальних даних у графічному поданні. Для цього використовується кореляційне поле, яке являє собою множину точок (x_i, y_i). Якщо точки кореляційного поля виявляють групування навколо деякої лінії (зокрема прямої), то є підстави припустити, що існує зв'язок між величинами X та Y, який описується залежністю y = f(x), графіком якої є лінія групування точок кореляційного поля.

Мірою кореляційного зв'язку між ознаками X та Y є коефіцієнт кореляції Пірсона (коефіцієнт лінійної парної кореляції), який розраховується на основі парної вибірки значень величин X та Y за формулою:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}},$$

де N – об'єм вибірки;

y_i, x_i – значення і-того елемента вибірок ***x*** та ***y*** відповідно;

$\bar{x}$, $\bar{y}$ – вибіркові середні ***x*** та ***y*** відповідно;

$\overline{xy}$ – середнє значення добутку x_i і y_i;

$\overline{y^2}$, $\overline{x^2}$ – середнє значення квадратів ознак ***y*** та ***x***.

Коефіцієнт кореляції Пірсона є мірою **тільки лінійного зв'язку**, що є його суттєвим обмеженням. Застосування цієї міри є коректним тільки тоді, коли вигляд кореляційне поля вказує на наявність лінійного зв'язку. Коли зв'язок нелінійний, навіть явно видимий з кореляційного поля, коефіцієнт кореляції Пірсона не дає правильних результатів – наприклад, показує відсутність зв'язку.

У випадку нелінійного зв'язку застосування застосування коефіцієнта кореляції Пірсона можливе після лінеаризації експериментальної залежності (наприклад, логарифмування даних, які підпорядковуються експоненціальній залежності). У випадку складної нелінійної залежності можлива її кусково-лінійна апроксимація, тобто заміна нелінійної залежності лінійною на окремих

ділянках (на графіку залежності це відповідає заміні окремих ділянок графіка нелінійної функції прямолінійними відрізками так, щоб вони мали відхилялись від графіка нелінійної залежності). Після цього коефіцієнт кореляції Пірсона коректно застосовується для оцінки сили лінійного зв'язку на окремих ділянках. При цьому має місце значне збільшення обсягів обчислень і ускладнення математичної моделі залежності.

Коефіцієнт кореляції Пірсона дозволяє оцінити силу (глибину, тісноту) лінійного зв'язку між величини, а також визначити напрям зв'язку. Коефіцієнт кореляції являє собою число в діапазоні від -1 до +1. Знак числа визначає напрям зв'язку, а абсолютне значення – силу зв'язку. Якщо коефіцієнт кореляції позитивний, то має місце прямий зв'язок, тобто при збільшенні значень однієї величини, друга також збільшується. Якщо ж коефіцієнт кореляції негативний, то зв'язок зворотний, тобто при збільшенні значень однієї величини, друга величина зменшується.

Силу зв'язку оцінюють на основі абсолютного значення коефіцієнта кореляції за такою шкалою:

Значення r	Оцінка зв'язку
$r < 0$	Зворотний зв'язок
$0 \le \|r\| < 0{,}1$	Зв'язок відсутній
$0{,}1 \le \|r\| < 0{,}3$	Слабкий
$0{,}3 \le \|r\| < 0{,}5$	Помірний
$0{,}5 \le \|r\| < 0{,}7$	Помітний
$0{,}7 \le \|r\| < 0{,}9$	Сильний
$0{,}9 \le \|r\| < 0{,}99$	Дуже сильний
$0{,}99 < \|r\| \le 1$	Повний (функціональний)

Коефіцієнт кореляції дозволяє тільки виявити наявність та оцінити силу зв'язку, причинно-наслідкові зв'язки ним ніяк не відображаються, що часто призводить до помилок. Змістовний аналіз кореляційного зв'язку має виконуватись спеціалістом на основі глибокого розуміння предмету дослідження.

Після встановлення факту існування зв'язку між величинами X та Y постає природне питання про його аналітичне визначення, тобто знаходження функції $y = f(x)$, яка описує виявлений зв'язок. Визначення функції $y = f(x)$ є задачею регресійного аналізу. Ця функція називається функцією регресії величини Y по X, або просто регресією Y по X. У випадку лінійної залежності функція регресії є лінійною:

$$y = kx+b,$$

так що задача регресійного аналізу полягає у визначення двох параметрів: коефіцієнтів k та b, які називаються коефіцієнтами регресії.

Для випадку лінійної залежності коефіцієнти регресії визначаються шляхом Розв'язування системи лінійних рівнянь, які включають елементи парної вибірки значень величин X та Y. Дана задача розв'язана для загального випадку, так що для коефіцієнтів регресії є готові загальні формули.

У випадку багатофакторної залежності результативної ознаки Y від факторів X_1, …, X_n методика кореляційного аналізу лишається тією самою. Спочатку візуально аналізуються кореляційні поля для усіх пар ознак (Y, X_i), на основі чого формулюються припущення про наявність і вид частинних залежностей між ознакою Y і кожним фактором X_i. Після цього розраховуються попарні коефіцієнти кореляції Пірсона для усіх пар факторів, і оцінюється сила зв'язків та їх вид. Можлива також оцінка кореляційної залежності між результативною ознакою Y та функціональними комбінаціями факторів X (наприклад, між H та попарними добутками факторів $X_i \cdot X_j$). Багатофакторний кореляційний аналіз є значно складнішим і більш трудомістким, ніж 1-фактроний. Побудова багатофакторної регресійної моделі являє собою надзвичайно складну задачу, для якої, на відміну від 1-факторної, немає загального розв'язку.

Зазвичай при виявленні зв'язку між двома величинами за допомогою кореляційного аналізу проводиться і його аналітичне визначення, тобто регресійний аналіз, внаслідок чого говорять про кореляційно-регресійний аналіз, предметом якого є як виявлення зв'язку, так і побудова його математичної моделі.

Кореляційно-регресійний аналіз є задачею, яка ефективно розв'язується за допомогою електронних таблиць. Типовим прикладом застосування електронних таблиць до кореляційно-регресійного аналізу є процесор електронних таблиць MS Excel (і його вільний аналог Open Office Calc). Розрахунок коефіцієнта кореляції Пірсона реалізується спеціальною вбудованою функцією КОРРЕЛ. Обчислення коефіцієнтів регресії ефективно реалізується за допомогою формул робочого аркуша, побудова яких полегшується наявністю вбудованих функцій, які реалізують проміжні розрахунки (такі як СУММПРОИЗВ – обчислює суму добутків відповідних елементів масивів, СУММКВ – обчислює суму квадратів тощо).

До складу пакету MS Excel входить надбудова Пакет аналізу, яка являє собою набір спеціалізованих інструментів, які реалізують типові складні функції статистичного аналізу даних. До складу пакету входять зокрема інструменти кореляційно-регресійного аналізу. Це – інструменти **Корреляция** и **Регрессия**. Обидва інструменти забезпечують виконання багатофакторного кореляційного і регресійного аналізу, включаючи 1-факторну модель як частинний випадок.

Вибір інструменту здійснюється у вікні надбудови Пакет аналізу (рис. 8.1.) зі списку інструментів.

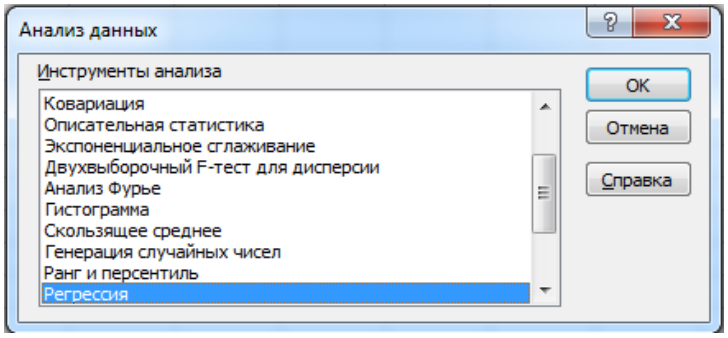


Рис. 8.1. Вікно меню надбудови **Пакет аналізу**.

Налаштування обраного інструменту виконується у його діалоговому вікні, де задаються діапазони даних та деякі допоміжні параметри. Діалогове вікно інструменту **Корреляция** показане на рис. 8.2, інструменту **Регрессия** – на рис. 8.3.

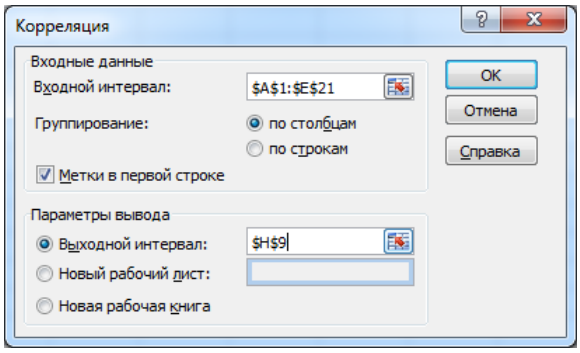


Рис. 8.2. Діалогове вікно інструменту **Корреляция**.

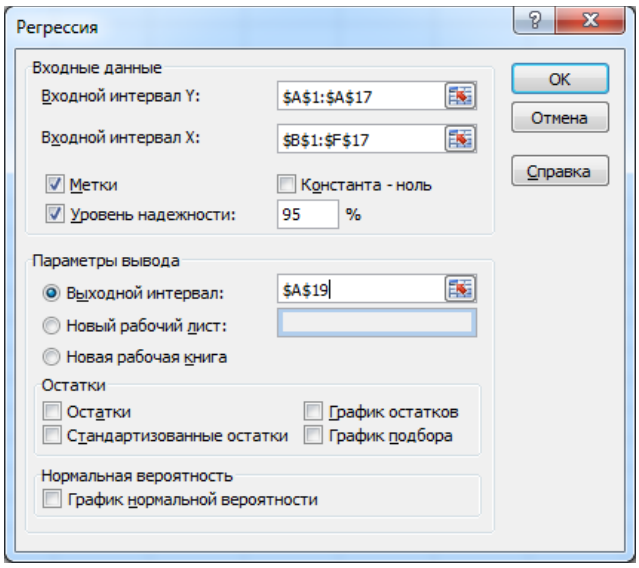


Рис. 8.3. Діалогове вікно інструменту **Регрессия**.

Результатом обробки даних інструментом **Корреляция** є кореляційна матриця, яка являє собою квадратну нижню трикутну матрицю, яка містить попарні коефіцієнти кореляції Пірсона усіх величин Y, X_1, …X_n .

В результаті обробки даних інструментом **Регрессия** створюється звіт, який відображається у заданому місці і містить коефіцієнти регресії і ряд статистичних оцінок, які характеризують значущість отриманої регресійної моделі і потрібні для подальшого поліпшення моделі.

Порядок виконання роботи

Наступні практичні завдання рекомендується виконувати на різних робочих аркушах однієї книги, іменуючи відповідним чином робочі аркуші, що полегшить і пришвидшить доступ до них.

I. Однофакторний кореляційно-регресійний аналіз.

Задача. Для виявлення зв'язку між вмістом фосфору в листку рослини *y* та вмістом фосфору у ґрунті *x* проведено серію аналізів і отримано такі результати:

x	5	4	3	7	13	11	23	28	18	9	12	13
y	64	71	54	71	93	76	87	109	83	75	79	87

Провести кореляційно-регресійний аналіз результатів.

Теоретична довідка

Найпростішим випадком є 1-факторний кореляційно-регресійний аналіз, коли аналізується зв'язок між результативною ознакою *y* та факторною ознакою *x* і відшукується рівняння залежності:

$$y = f(x).$$

Гіпотеза про існування залежності між величинами X та Y висувається на основі аналізу вигляду експериментальних даних у графічному поданні. Для цього використовується кореляційне поле, яке являє собою множину точок (x_i, y_i). Якщо точки кореляційного поля виявляють групування навколо деякої лінії (зокрема прямої), то є підстави припустити, що існує зв'язок між величинами X та Y, який описується залежністю y = f(x), графіком якої є лінія групування точок кореляційного поля.

Найпростішою залежністю є лінійна залежність, що описується рівнянням:

$$y = kx+b.$$

Тіснота лінійного зв'язку між ознаками ***x*** та ***y*** визначається за допомогою коефіцієнта лінійної парної кореляції (коефіцієнт кореляції Пірсона), який розраховується за вибірками ознак ***x*** та ***y*** за формулою:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}, \quad (8.3)$$

де N – об'єм вибірки;

y_i, x_i – значення і-того елемента вибірок ***x*** та ***y*** відповідно;

$\bar{x}$, $\bar{y}$ – вибіркові середні ***x*** та ***y*** відповідно;

$\overline{xy}$ – середнє значення добутку $\boldsymbol{x_i}$ і $\boldsymbol{y}_i$;

$\overline{y^2}$, $\overline{x^2}$ – середнє значення квадратів ознак ***y*** та ***x***.

Значення коефіцієнта кореляції змінюється від -1, що відповідає зворотному зв'язку, до +1, що відповідає прямо пропорційному зв'язку (значення 0 означає відсутність залежності).

Передумови застосування коефіцієнта кореляції Пірсона:

1) усі спостереження взаємно незалежні;
2) спостереження мають нормальний закон розподілу.

Зауваження. *Слід пам'ятати, що коефіцієнт кореляції Пірсона показує тісноту тільки лінійного зв'язку. У випадку більш складних залежностей (нелінійних) коефіцієнт кореляції буде показувати відсутність зв'язку.*

Значимість коефіцієнта кореляції r_{xy} перевіряється за допомогою критерію Стьюдента (фактично перевіряється гіпотеза про рівність коефіцієнта кореляції нулю). Для цього розраховується критеріальне значення:

$$t_p = \frac{\sqrt{r^2(N-2)}}{\sqrt{1-r^2}}, \quad (8.4)$$

де r – значення коефіцієнта кореляції;

N – об'єм вибірки.

Критичне значення критерію $t_{кр}$ визначають за розподілом Стьюдента при заданій значущості α і числі ступенів свободи ν = N – 2:

$$t_{кр} = t(p = 1 - \frac{\alpha}{2};\quad \nu = N - 2). \quad (8.5)$$

Якщо розрахункове значення t_p більше критичного ($t_p > t_{кр}$), то коефіцієнт кореляції є суттєвим (значущим) на рівні значущості α (рівень надійності $\gamma = 1 - \alpha$), тобто досліджуваний зв'язок має невипадковий характер.

Якщо виявлено наявність достатньо сильного лінійного зв'язку між ознаками ***x*** та ***y***, то постає задача знаходження рівняння:

$$y = b_1 x + b_0, \quad (8.6)$$

яке описує цей зв'язок. Розв'язування цієї задачі являє суть регресійного аналізу.

Параметри лінійної регресії b_1 і b_0 розраховуються на основі методу найменших квадратів (суть методу НК полягає в тому, щоб підібрати лінію з

такими параметрами b_1 і b_0, що сума квадратів залишкових відхилень $|y - y_i|$ буде мінімальною). Параметри лінійної регресії розраховуються за формулами:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad (8.7)$$

$$b_0 = \bar{y} - b_1 \bar{x}, \qquad (8.8)$$

де $\bar{x}$, $\bar{y}$ – вибіркові середні добутку x та y відповідно;

$\overline{xy} = \frac{1}{n}\sum_{i=1}^{N} x_i y_i$ – середнє значення добутку x_i, y_i;

$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{N} x_i^2$ – середній квадрат x.

Завдання 1. Побудова робочого аркуша для 1-факторного кореляційного аналізу.

Відкрити нову робочу книгу і присвоїти робочому аркушу ім’я «1-факторна кореляція».

Підготувати робочий аркуш за зразком рис. 8.1, розмістивши вибірку у перших двох стовпчиках. У комірки A2-A13 ввести значення фактора X, у комірки B2-B13 ввести відповідні значення фактора Y.

Побудувати кореляційне поле за даними задачі.

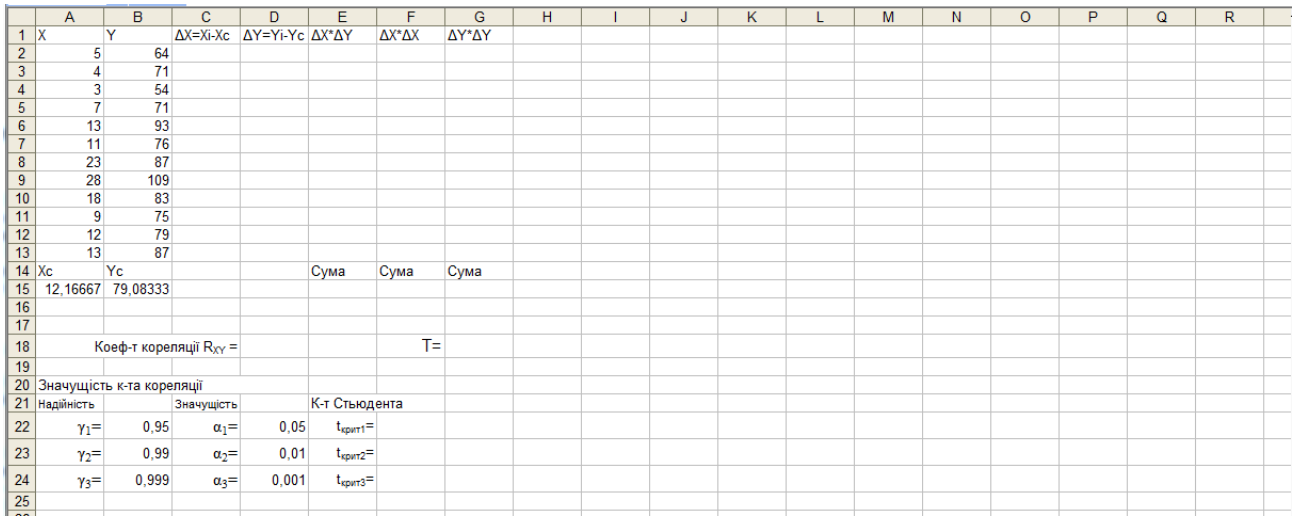
Для побудови кореляційного поля необхідно застосовувати діаграму типу «Точкова», який призначено саме для відображення пар значень. Діаграми інших типів для цієї мети не підходять, оскільки не забезпечують відображення зв’язаних пар значень.

Для побудови кореляційного поля слід:

- виділити діапазон даних A1:B13 і викликати майстра діаграм;
- обрати тип діаграми – «Точкова», вид – «Маркери»;
- задати параметри діаграми (назва, позначки осей, лінії сітки тощо) і розміщення – *на окремому аркушеві*.

Розглянути отримане кореляційне поле, щоб пересвідчитись у тому, що дані виявляють групування навколо прямої лінії, що є підставою для припущення про існування лінійної залежності між величинами X та Y.

Ввести у таблицю розрахункові формули. У комірки B22, B23, B24 ввести значення довірчої ймовірності γ_1=0,95, γ_2 = 0,99, γ_3=0,999 відповідно. У комірки A15 і B15 ввести формули для розрахунку вибіркових середніх $\bar{x}$ та $\bar{y}$ відповідно (використати функцію СРЗНАЧ).



	A	B	C	D	E	F	G
1	X	Y	ΔX=Xi-Xc	ΔY=Yi-Yc	ΔX*ΔY	ΔX*ΔX	ΔY*ΔY
2	5	64					
3	4	71					
4	3	54					
5	7	71					
6	13	93					
7	11	76					
8	23	87					
9	28	109					
10	18	83					
11	9	75					
12	12	79					
13	13	87					
14	Xc	Yc			Сума	Сума	Сума
15	12,16667	79,08333					
16							
17							
18			Коеф-т кореляції R_{XY} =			T=	
19							
20	Значущість к-та кореляції						
21	Надійність		Значущість		К-т Стьюдента		
22	γ_1=	0,95	α_1=	0,05	$t_{крит1}$=		
23	γ_2=	0,99	α_2=	0,01	$t_{крит2}$=		
24	γ_3=	0,999	α_3=	0,001	$t_{крит3}$=		
25							

Рис. 8.4. Зразок робочого аркуша до 1-факторного кореляційного аналізу.

У комірку C2 ввести формулу для розрахунку відхилення x_i від $\bar{x}$, $\Delta x_i = x_i - \bar{x}$ („=A2- $A$15”), а у комірку D2 – для розрахунку відхилення y_i від $\bar{y}$ („=B2-$B$15”). Скопіювати (розтягнути) введені формули на відповідні діапазони C2:C13 і D2:D13.

У комірках E2, F2, G2 створити формули для розрахунку відповідно: добутку відхилень вибіркових значень $\Delta x_i \cdot \Delta y_i$ „=C2*D2”; квадрату відхилення Δx „=C2^2”, квадрату відхилення Δy „=D2^2”. Створені формули розтягнути на відповідні діапазони E3:E13; F2:F13; G2:G13.

У комірки E15, F15, G15 ввести формули для обчислення сум $\sum_{i=1}^{N} x_i y_i$ („=СУММ(E2:E13)”), $\sum_{i=1}^{N} x_i^2$ („=СУММ(F2:F13)”), $\sum_{i=1}^{N} y_i^2$ („=СУММ(G2:G13)”. Функція СУММ належить до категорії ***Математические,*** а також подана кнопкою **Σ** на панелі інструментів.

У комірку D18 ввести формулу для розрахунку коефіцієнта кореляції „=E15/(КОРЕНЬ(F15*G15))” за допомогою майстра функцій (функція КОРЕНЬ належить до категорії ***Математические***).

Ввести формулу для розрахунку критерію значущості коефіцієнта кореляції:

$$t_p = \frac{\sqrt{r^2(N-2)}}{\sqrt{1-r^2}} \qquad (8.9)$$

Обрати комірку G18 і ввести до неї формулу =КОРЕНЬ(D18^2*(СЧЁТ(A1:A13)-2)/(1-D18^2)).

До комірок F22, F23, F24 ввести формули для знаходження критичного значення критерію $t_{кр}$ для заданого рівня значущості. Для цього:

- обрати комірку F22 і ввести формулу „=СТЬЮДРАСПОБР(1-B22;СЧЁТ($A$2: $A$13)-2)”;
- скопіювати формулу з комірки F22 на комірки F23 і F24.

***Зауваження**. Зверніть увагу, що аргументом функції СТЬЮДРАСПОБР є рівень значущості α, який розраховується через заданий рівень надійності γ (комірки D20-D22): α = 1 – γ.*

На основі отриманих значень коефіцієнта кореляції ***r***, розрахункового значення t-критерія *T* і критичних рівнів для трьох значень надійності γ_1=0,95, γ_2=0,99, γ_3=0,999 зробити висновок про вид і глибину кореляційного зв’язку та його значущість і занести висновок до протоколу.

Продумати і запропонувати, як побудувати робочий аркуш (модифікувати створений у завданні), щоб можна було працювати з вибірками довільного розміру. Як треба змінити формули робочого аркуша.

Записати висновок у протокол і зберегти робочу книгу у своїй робочій теці.

Завдання 2. Розрахунок коефіцієнта кореляції за допомогою функції КОРРЕЛ.

2.1. Скопіювати на новий робочий аркуш вхідні дані задачі, для чого:

- на робочому аркушеві завд.1 виділити діапазон A1:B13, що містить дані задачі і скопіювати його до буферу обміну;
- перейти на новий робочий аркуш, обрати на ньому комірку A1 і вставити вміст буферу обміну.

2.2. До комірки D5 ввести текст „r = ”, а до комірки E5 – формулу „=КОРРЕЛ(A2:A13; B2:B13)”, для чого:

- обрати комірку E5 і викликати майстра функцій;
- у категорії ***Статистические*** обрати функцію КОРРЕЛ;
- для функції КОРРЕЛ в якості аргумента ***Массив 1*** задати діапазон A2:A13, а в якості аргумента ***Массив 2*** – діапазон B2:B13;
- дати ЛК на кнопці ***ОК*** вікна майстра функцій і отримати результат у комірці E5.

2.3. Записати до протоколу висновок і зберегти робочу книгу.

2.4. За допомогою довідкової системи ознайомитись з інформацією по функції КОРРЕЛ та занотувати основні відомості.

Завдання 3. Розрахунок рівняння лінійної регресії і побудова лінії регресії.

3.1. Перейти на наступний робочий аркуш (при потребі вставити новий аркуш командою ***Вставка/Лист***), присвоїти йому ім’я «Регресія» і підготувати його згідно зразка рис. 2, для чого скопіювати на нього діапазон вхідних даних задачі завдання 1 і формули для обчислення вибіркових середніх (діапазон A1:B15) та виконати необхідні доповнення.

	A	B	C	D	E	F	G	H	I	J	K
1	X	Y	X*Y	X^2	Yteop						
2	5	64				Коеф-ти регресії					
3	4	71									
4	3	54				b0=					
5	7	71				b1=					
6	13	93									
7	11	76									
8	23	87									
9	28	109									
10	18	83									
11	9	75									
12	12	79									
13	13	87									
14	Xc	Yc									
15	12,16667	79,08333									
16											
17											

Рис. 8.5. Зразок робочого аркуша для розрахунку рівняння регресії.

3.2. До комірки C2 ввести формулу обчислення добутку відповідних елементів вибірок „=A2*B2”, після чого розтягнути формулу на діапазон C2:C13.

3.3. До комірки D2 ввести формулу обчислення квадрату елементів вибірки *x* „=A2*A2”, і розтягнути її на діапазон D2:D13.

3.4. До комірки C15 ввести формулу для обчислення середнього значення добутку $x_i \cdot y_i$ за допомогою функції СРЗНАЧ.

3.5. Ввести до комірки D15 формулу для розрахунку середнього значення квадрату елементів вибірки.

3.6. До комірки G5 занести формулу для обчислення коефіцієнта регресії b_1:

=(C15-A15*B15)/(D15-A15*A15),

а до комірки G4 – формулу для обчислення коефіцієнта b_0 (вільного члена рівняння регресії):

=B15-G5*A15.

3.7. Занести до протоколу розрахункові формули і отримане рівняння регресії..

3.8. Побудувати кореляційне поле з лінією регресії. Для цього:

– до комірки E2 ввести формулу рівняння регресії (1.5), яка буде мати вигляд

=$G$5*A2+$G$4;

– скопіювати формулу комірки E2 на діапазон E2:E13;

– виділити на робочому аркушеві «Регресія» діапазони A2:B13 і E2:E13 (несуміжні діапазони виділяються протягуванням вказівника при натиснених лівій кнопці маніпулятора і клавіші CTRL);

– викликати майстра діаграм і обрати діаграму типу «Точкова» з маркерами і прямими відрізками;

– виконати налаштування діаграми (назва, осі тощо).

3.9. Записати робочу книгу до своєї робочої теки. Записати до протоколу висновок.

Завдання 4. Виконання кореляційно-регресійного аналізу за допомогою інструментів пакету аналізу.

Зауваження. *Надбудова Пакет аналізу може бути неактивною і, відповідно, не відображатися у меню програми, у цьому випадку її слід активізувати, для чого потрібно виконати команду* ***Надстройки*** *і у її вікні встановити прапорець вибору для надбудови* ***Пакет анализа****.*

4.1. Визначення коефіцієнта лінійної кореляції величин X та Y.

Перейти на новий робочий аркуш, назвати його «Пакет аналізу» і скопіювати на нього дані задачі (діапазон A1:B13).

Відкрити надбудову **Пакет анализа** і в її вікні обрати інструмент **Корреляция.**

У вікні інструменту **Корреляция** виконати наступне:

– у полі **Входной интервал** вказати діапазон A1:B13;
– встановити кнопку групування по стовпчикам;
– встановити прапорець **Метки в первой строке;**
– у полі **Выходной интервал** задати довільну комірку, наприклад D4.

Після виконання налаштувань клацнути кнопку ОК у вікні інструменту **Корреляция.** В результаті на робочому аркуші відобразиться кореляційна матриця рис.3, яка у комірці на перетині стовпчика X і рядка Y містить коефіцієнт кореляції величин X та Y.

	X	*Y*
X	1	
Y	0,884562	1

Рис. 8.6. Кореляційна матриця

4.2. Регресійний аналіз засобами пакету аналізу.

На робочому аркушеві «Кореляція» відкрити надбудову **Пакет анализа** і в її вікні обрати інструмент **Регрессия.**

У вікні інструменту **Регрессия** виконати наступне:

– **Выходной интервал** задати довільну у полі **Входной інтервал Y** вказати діапазон B1:B13;
– у полі **Входной інтервал X** вказати діапазон A1:A13;
– встановити прапорець **Метки;**
– (необов'язково) встановити прапорець **Уровень надёжности** і ввести у відповідне поле значення надійності у відсотках (наприклад, 95 або 97);
– у полі комірку, наприклад D9.

Після виконання налаштувань клацнути кнопку ОК у вікні інструменту**.**

В результаті на робочому аркуші відобразиться прив'язаний лівим верхнім кутом до заданої комірки D9 звіт інструменту **Регрессия** (рис. 8.7), який містить

значення коефіцієнтів лінійної регресії та ряд статистичних оцінок результатів. Значення коефіцієнтів регресії містяться у комірках на перетині стовпчика **Коэффициенты** з рядками Y-пересечение та X (обведені рамкою). При цьому число X – це коефіцієнт b_1 рівняння регресії (1.4), а число Y-пересечение – коефіцієнт b_0.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,884561612
R-квадрат	0,782449245
Нормированный R-квадрат	0,76069417
Стандартная ошибка	6,985329007
Наблюдения	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	1754,968453	1754,968453	35,96628505	0,000132605
Остаток	10	487,9482134	48,79482134		
Итого	11	2242,916667			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	58,99352667	3,909969922	15,08797455	3,30519E-08	50,28157081	67,70548253	50,28157081	67,70548253
X	1,651216986	0,275331789	5,997189763	0,000132605	1,037739533	2,264694439	1,037739533	2,264694439

Рис. 8.7. Звіт інструменту **Регрессия**.

4.3. Записати результати та висновок до протоколу і зберегти робочу книгу у своїй робочій теці.

Завдання 5. Задача для самостійного розв'язування.

За допомогою створених розрахункових таблиць виконати кореляційно-регресійний аналіз (побудувати кореляційне поле, знайти коефіцієнт кореляції, розрахувати рівняння регресії) залежності ознаки ***y*** від ознаки ***x*** за наступною вибіркою:

x	2,5	2,9	3,4	3,8	4,1	4,4	5,0	3,8	2,9	4,1	5,0	3,8
y	1,5	2,0	2,5	3,0	3,5	4,0	4,5	3,2	2,4	3,4	4,4	2,9

Застосувати побудований у завд.1 робочий аркуш і інструменти пакету аналізу. Занести до протоколу результати.

ІІ. Багатофакторний кореляційно-регресійний аналіз.

Завдання 6. Провести багатофакторний кореляційно-регресійний аналіз даних за допомогою пакету аналізу.

Задача. Потрібно знайти залежність показника Y від факторів X_1, X_2 X_3, X_4 на основі наявних даних про їх значення.

Y	X_1	X_2	X_3	X_4
9,7	1,59	0,26	2,05	0,32
10,2	1,8	0,55	3,22	0,4
11,5	2,5	0,71	4,6	0,72
9,9	1,6	0,49	2,57	0,39
8,6	0,7	0,67	2,95	0,2
9	1,1	0,88	3,6	0,25
10,7	2	0,92	4,31	0,64
10,1	1,6	0,63	4,24	0,55
9,5	1,2	0,74	3,86	0,28
11,3	1,9	0,82	3,9	0,68
11	1,8	0,95	4,24	0,62

6.1. Відкрити новий робочий аркуш, назвати його «Багатофакторний» і занести на нього дані задачі за зразком рис. 8.8.

	A	B	C	D	E	F	G	H
1	Y	X_1	X_2	X_3	X_4			
2	9,7	1,59	0,26	2,05	0,32			
3	10,2	1,8	0,55	3,22	0,4			
4	11,5	2,5	0,71	4,6	0,72			
5	9,9	1,6	0,49	2,57	0,39			
6	8,6	0,7	0,67	2,95	0,2			
7	9	1,1	0,88	3,6	0,25			
8	10,7	2	0,92	4,31	0,64			
9	10,1	1,6	0,63	4,24	0,55			
10	9,5	1,2	0,74	3,86	0,28			
11	11,3	1,9	0,82	3,9	0,68			
12	11	1,8	0,95	4,24	0,62			
13								

Рис. 8.8. Зразок робочого аркуша для багатофакторного кореляційно-регресійного аналізу.

6.2. Аналогічно п.1.3 завдання 1 побудувати і проаналізувати кореляційні поля для усіх пар факторів X_1, X_2 X_3, X_4 з величиною Y, щоб побачити, чи є підстави припускати наявність лінійного зв'язку між величиною Y і факторами X_1, X_2 X_3, X_4 .

6.3. Відкрити надбудову **Пакет анализа** і в її вікні обрати інструмент **Корреляция.**

У вікні інструменту **Корреляция** виконати наступне:

– у полі **Входной интервал** вказати діапазон A1:E12;
– встановити кнопку групування по стовпчикам;
– встановити прапорець **Метки в первой строке ;**

– у полі **Выходной интервал** задати довільну комірку, наприклад G2.

Після виконання налаштувань клацнути кнопку ОК у вікні інструменту **Корреляция.** В результаті на робочому аркуші відобразиться кореляційна матриця рис.9. У комірці на перетині стовпчика однієї змінної з рядком другої міститься коефіцієнт парної кореляції цих змінних. Наприклад, коефіцієнт кореляції змінних X_1 та X_3 дорівнює 0,465513:

$$r(X_1, X_3) = 0{,}465513.$$

З кореляційної матриці видно сили взаємної залежності між факторами. Так в даному випадку, видно (з вигляду кореляційних полів і значення коефіцієнту кореляції), що є дуже сильний лінійний зв'язок величини Y з факторами X_1 та X_4 , помітний – з фактором X_3 і слабкий з фактором X_2. Це дозволяє твердити, що фактори X_1 , X_3 та X_4 мають бути враховані в регресійній моделі. При цьому між факторами X_1 та X_4 також є сильний прямий зв'язок (тобто один з них може бути виражений через інший), так що один з цих факторів може бути виключений з рівняння регресії.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Y	X_1	X_2	X_3	X_4								
2	9,7	1,59	0,26	2,05	0,32			*Y*	*X1*	*X2*	*X3*	*X4*	
3	10,2	1,8	0,55	3,22	0,4		Y	1					
4	11,5	2,5	0,71	4,6	0,72		X1	0,933047	1				
5	9,9	1,6	0,49	2,57	0,39		X2	0,284802	0,079733	1			
6	8,6	0,7	0,67	2,95	0,2		X3	0,584652	0,465513	0,782969	1		
7	9	1,1	0,88	3,6	0,25		X4	0,959464	0,880486	0,387486	0,689918	1	
8	10,7	2	0,92	4,31	0,64								
9	10,1	1,6	0,63	4,24	0,55								
10	9,5	1,2	0,74	3,86	0,28								
11	11,3	1,9	0,82	3,9	0,68								
12	11	1,8	0,95	4,24	0,62								
13													

Рис. 8.9. Кореляційна матриця.

6.4. Відкрити надбудову **Пакет анализа** і в її вікні обрати інструмент **Регрессия.**

У вікні інструменту **Регрессия** виконати наступне:

– у полі **Входной интервал Y** вказати діапазон A1:A12;

– у полі **Входной интервал X** вказати діапазон B1:E12;

– встановити прапорець **Метки;**

– (необов'язково) встановити прапорець **Уровень надёжности** і ввести у відповідне поле значення надійності у відсотках (наприклад, 95 або 97);

– у полі **Выходной интервал** задати довільну комірку, наприклад A14.

Після виконання налаштувань клацнути кнопку ОК у вікні інструменту**.**

В результаті на робочому аркуші відобразиться прив'язаний лівим верхнім кутом до заданої комірки A14 звіт інструменту **Регрессия** (рис. 7), який містить значення коефіцієнтів лінійної регресії та ряд статистичних оцінок результатів.

Значення коефіцієнтів регресії містяться у комірках стовпчика **Коэффициенты** напроти імен відповідних факторів. Y-пересечение – це вільний член рівняння регресії.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,980030608
R-квадрат	0,960459992
Нормированный R-квадрат	0,934099986
Стандартная ошибка	0,237861635
Наблюдения	11

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	4	8,245985602	2,061496401	36,43625921	0,000239937
Остаток	6	0,339468943	0,056578157		
Итого	10	8,585454545			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	7,588060473	0,504959672	15,02706233	5,4721E-06	6,352468668	8,823652277	6,352468668	8,823652277
X1	0,846641713	0,409726017	2,066360636	0,084310983	-0,155921732	1,849205157	-0,155921732	1,849205157
X2	0,565761126	0,711779022	0,794855016	0,456978528	-1,175899395	2,307421646	-1,175899395	2,307421646
X3	-0,173612024	0,203793176	-0,851903033	0,426956607	-0,67227596	0,325051912	-0,67227596	0,325051912
X4	3,073880871	1,175431749	2,615107915	0,039846611	0,197702999	5,950058743	0,197702999	5,950058743

Рис. 8.10. Звіт інструменту **Регресія**.

6.5. Записати рівняння регресії, округливши коефіцієнти до сотих. Коефіцієнт при змінній фактора X_i – це число напроти імені фактора у стовпчику **Коефіцієнти**. Рівняння регресії: $Y = 0{,}847 \cdot X_1 + 0{,}566 \cdot X_2 - 0{,}174 \cdot X_3 + 3{,}07 \cdot X_4$.

Зауваження. *Отримане рівняння регресії являє собою тільки перше наближення. Далі зазвичай виконується процедура його поліпшення на основі врахування взаємозв'язків факторів, їх значущості тощо.*

6.6. Записати до протоколу рівняння регресії і висновок. Зберегти робочу книгу до своєї робочої теки.

Хід роботи

1 Задатись вибіркою А із 100елементів , що відповідає 1-му періоду синусоїди з параметрами Т-секунд, Т – номер студента.

2. Задатись вибіркою Б, що відповідає синусоїді з пункту 1 у протифазі з половинною амплітудою.

3. Задатись вибіркою В, що утворюється з вибірки А шляхом додавання рівномірного шуму амплітудою 25 % від початкової.

4. Здійснити кореляційний аналіз А і Б, А і В.

Контрольні питання

1. Поняття про кореляційний аналіз.
2. Лінійний кореляційний зв язок.
3. Показник кореляції рангів.
4. Множинна кореляція
5. Аналіз вимого до програмного забезпечення.
6. Сертифікація вимого до програмного забезпечення.

ПЕРЕЛІК ЗАВДАНЬ ДЛЯ КОНТРОЛЬНОЇ РОБОТИ

Варіант 1. Нижче наведено дані про ширину (в дюймах з точністю до 0,0001 дюйма) 115 зразків з цинку марки ВВ.

0,4998	0,5002	0,5	0,4995	0,5001	0,5005	0,5	0,5008	0,5	0,5005
0,5009	0,5005	0,5001	0,5004	0,5	0,4998	0,4997	0,4998	0,4994	0.4999
0,5001	0,5003	0,4996	0,4994	0,5001	0,4999	0,4992	0,4999	0,5001	0,5004
0,4998	0,5005	0,5005	0,5002	0,5003	0,5003	0,5008	0,5	0,4999	0,4999
0,5	0,5005	0,5008	0,5007	0,5008	0,5002	0,5	0,4994	0,4998	0,4991
0,5008	0,5009	0,501	0,5005	0,5006	0,4997	0,4997	0,4999	0,5001	0,5003
0,5	0,5001	0,5002	0,4995	0,4996	0,5001	0,5003	0,499	0,4993	0,5
0,4993	0,4994	0,4999	0,4996	0,4997	0,4999	0,4998	0,5001	0,5006	0,4998
0,4995	0,4995	0,4992	0,4995	0,4992	0,5	0,4999	0,5	0,5002	0,4997
0,4994	0,4998	0,499	0,5	0,5	0,5002	0,4996	0,5	0,5007	0,5005
0,5	0,4993	0,5002	0,5001	0,5003	0,4998	0,5004	0,4998	0,4999	0,5
0,4994	0,5	0,4996	0,4997	0,5001					

Побудувати гістограму, полігон частот та полігон накопичених частот. Обчислити середнє значення ширини,вибіркову дисперсію та вибіркове середнє квадратичне відхилення.

Варіант 2. Вивчався розподіл часу уповільнення нейтронів до різних енергій у водневому уповільнювачі. Наведені в таблиці дані являють собою груповану вибірку, що містить 1000 значень часу уповільнення нейтронів (в мікросекундах) до енергії 0,025 еВ.

Час мкс	Частота	Час мкс	Частота
0	40	11	20
1	104	12	14
2	124	13	10
3	150	14	6
4	121	15	1
5	106	16	2
6	104	17	0
7	76	18	1
8	49	19	0
9	42	20	0
10	29	21	1

Побудувати гістограму, полігон частот та полігон накопичених частот,емпіричну функцію розподілу. Обчислити середнє значення часу уповільнення,вибіркову дисперсію і вибіркове середнє квадратичне відхилення.

Варіант 3. В виробництві радіоламп важливо, щоб їх виходи були прямими,інакше за умов автоматичного виробництва вони не потраплять в потрібні отвори,що викличе виникнення неякісної продукції. Прямизна виходу перевіряється за допомогою оптичного компаратора з закріпленим одним кінцем провідника. Прогин визначається як різниця між максимальним та мінімальним положеннями не затиснутого кінця провідника мінус його діаметр. Таким чином,прогин абсолютно прямого провідника дорівнює нулеві. Нижче в таблиці наведені дані про прогин 487 провідників.

Прогин, 10^{-5} м	Частота	Прогин, 10^{-5} м	Частота
1	12	23	18
3	26	25	14
5	36	27	12
7	50	29	4
9	45	31	7
11	49	33	2
13	51	35	5
15	44	37	2
17	44	39	2
19	40	41	1
21	32	43	1

Обчислити вибіркове середнє значення прогину,вибіркову дисперсію та вибіркове середнє квадратичне відхилення. Побудувати гістограму та емпіричну функцію розподілу.

Варіант 4. Нижче наведено дані про наслідки 150 аналізів відносно вмісту триокису сірки в суміші (у відсотках), проведених на протязі місяця.

15,8	16,0	15,7	16,0	15,7	15,9	16,0	15,7	15,8	15,7
15,4	15,7	15,8	15,7	15,9	16,0	15,7	15,7	15,7	15,8
15,8	15,6	15,9	15,8	15,5	16,0	15,7	15,7	15,7	15,8
15,9	15,7	15,8	16,0	15,8	15,9	16,2	15,7	15,5	15,9
15,7	15,7	15,3	15,6	16,1	15,7	16,1	15,9	15,8	16,0
16,1	15,7	15,5	15,6	15,8	15,6	15,8	15,8	15,6	15,7
15,6	15,9	15.8	15,8	15,8	15,9	15,6	15,8	15,8	15,9
15,5	15,8	15,4	15,5	15,5	15,7	15,6	15,9	15,8	15,5
15,9	15,8	15,5	15,9	15,6	15,8	15,6	15,7	15,7	15,7
15,7	15,7	16,0	16,1	15,6	15,5	15,6	15,5	16,0	15,5
15,8	15,8	15,9	16,1	15,5	15,7	16,0	15,9	15,7	15,5
16,1	15,7	15,7	15,5	16,2	15,7	15,6	16,0	15,6	15,7
15,3	15,5	15,4	16,0	15,7	15,5	15,8	15,4	15,7	16,3
15,9	15,6	15,7	15,4	15,9	15,6	16,0	15,7	15,8	15,9
16,0	16,0	15,8	15,9	15,7	15,6	15,6	15,9	15,6	15,5

Побудувати гістограму та емпіричну функцію розподілу. Обчислити вибіркове середнє значення відсоткового вмісту сірки в суміші, вибіркове середнє квадратичне відхилення,визначити моду, медіану та розмах вибірки.

Варіант 5. Наведені нижче дані являють собою кількості виробів, виготуваних на протязі однієї години деяким дрібним виробником.

Побудувати груповану вибірку, визначити медіану, моду та розмах вибірки. Обчислити середню кількість виробів за годину, вибіркове середнє квадратичне відхилення.

136	122	132	128	123	133	130	131
134	149	138	127	119	137	133	130
143	134	128	131	118	133	131	132
118	128	122	130	139	145	122	130
128	136	132	126	124	117	139	132
141	144	138	133	127	150	144	133

134	125	140	135	129	138	138	147
150	126	135	136	150	135	138	140
122	142	127	127	132	145	140	133
127	142	144	125	132	145	137	132

Варіант 6. Нижче наведено дані про наслідки 150 аналізів відносно вмісту триокису сірки в суміші (у відсотках), проведених на протязі місяця.

15,8	16,0	15,7	16,0	15,7	15,9	16,0	15,7	15,8	15,7
15,4	15,7	15,8	15,7	15,9	16,0	15,7	15,7	15,7	15,8
15,8	15,6	15,9	15,8	15,5	16,0	15,7	15,7	15,7	15,8
15,9	15,7	15,8	16,0	15,8	15,9	16,2	15,7	15,5	15,9
15,7	15,7	15,3	15,6	16,1	15,7	16,1	15,9	15,8	16,0
16,1	15,7	15,5	15,6	15,8	15,6	15,8	15,8	15,6	15,7
15,6	15,9	15.8	15,8	15,8	15,9	15,6	15,8	15,8	15,9
15,5	15,8	15,4	15,5	15,5	15,7	15,6	15,9	15,8	15,5
15,9	15,8	15,5	15,9	15,6	15,8	15,6	15,7	15,7	15,7
15,7	15,7	16,0	16,1	15,6	15,5	15,6	15,5	16,0	15,5
15,8	15,8	15,9	16,1	15,5	15,7	16,0	15,9	15,7	15,5
16,1	15,7	15,7	15,5	16,2	15,7	15,6	16,0	15,6	15,7
15,3	15,5	15,4	16,0	15,7	15,5	15,8	15,4	15,7	16,3
15,9	15,6	15,7	15,4	15,9	15,6	16,0	15,7	15,8	15,9
16,0	16,0	15,8	15,9	15,7	15,6	15,6	15,9	15,6	15,5

Побудувати гістограму та емпіричну функцію розподілу. Обчислити вибіркове середнє значення відсоткового вмісту сірки в суміші, вибіркове середнє квадратичне відхилення,визначити моду, медіану.

Варіант 7. Наведені нижче дані являють собою кількості виробів, виготуваних на протязі однієї години деяким дрібним виробником.

Побудувати груповану вибірку, визначити медіану, моду вибірки. Обчислити середню кількість виробів за годину, вибіркове середнє квадратичне відхилення.

136	122	132	128	123	133	130	131
134	149	138	127	119	137	133	130
143	134	128	131	118	133	131	132
118	128	122	130	139	145	122	130
128	136	132	126	124	117	139	132

141	144	138	133	127	150	144	133
134	125	140	135	129	138	138	147
150	126	135	136	150	135	138	140
122	142	127	127	132	145	140	133
127	142	144	125	132	145	137	132

Варіант 8

Дискретна випадкова величина X задана рядом розподілу

X	0	1	2	3	4
P	0,1	0,2	0,35	0,20	0,15

Знайти $M[X]$ і σ_x.

Варіант 9

Час, що був витрачений опитуваними на ранкову гімнастику, склав 11, 15, 12, 0, 16, 19, 6, 11, 12, 13, 16, 8, 9, 14, 5, 11, 3 хв. Для цієї вибірки знайти об'єм, розмах, побудувати варіаційний ряд та ряд розподілу частот.

Варіант 10

За 16 місяців витрати міського бюджету на озеленення склали 17, 18, 16, 16, 17, 18, 19, 17, 15, 17, 19, 18, 16, 16, 18, 18 тис. грн. Знайти об'єм цієї вибірки, її розмах, побудувати варіаційний ряд та ряд розподілу частот.

Варіант 11

У групі на заняттях із статистики провести експеримент щодо реєстрації місяця народження студентів (опитування виконати, наприклад, за списком групи). Побудувати варіаційний ряд та ряд розподілу частот отриманої вибірки. Зобразити результат у вигляді секторної діаграми.

Варіант 12

Нехай розподіл зарплати баскетболістів команди NBA у млн дол. США має вигляд 0.5, 0.6, 0.9, 1, 1, 1.3, 1.9, 2.33, 2.4, 2.4, 2.5, 3, 11.4, 21, 25 (дані чемпіонату США 2010–2011 рр.). Обчислити моду, медіану і середнє та інтерпретувати їх.

Варіант 13. Нехай маємо 4 вибірки з елементами 1, 2, 3, 4 та 5, записані в ряд розподілу:

Вибірка	1	2	3	4	5
a	0,46	0,04	0,00	0,04	0,46
b	0,04	0,08	0,77	0,08	0,04
c	0,19	0,19	0,23	0,19	0,19

d 0,04 0,08 0,12 0,19 0,58

Порівняти дисперсії цих вибірок.

Варіант 14.

Нехай задано ряд розподілу частот для кількості *Xi* неправильних з'єднань за хвилину на телефонній станції:

Xi	0	1	2	3	4
mi	75	25	12	5	2

Знайти вибіркову дисперсію та вибіркове середнє цієї величини.

Варіант 15. Нехай задано груповану вибірку для часу обробки тесту (у хв) студентами першого курсу:

Xi	[0; 4)	[4; 8)	[8; 12)	[12; 16)	[16; 20]
mi	2	7	28	10	3

Знайти вибіркову дисперсію та вибіркове середнє цієї величини.

Варіант 16. Під час дослідження часу обслуговування покупця біля касового апарата, зафіксовано тривалості обслуговування (у хв).

Дані наведено у таблиці:

Xi	[0; 0,2)	[0,2; 0,4)	[0,4; 0,6)	[0,6; 0,8)	[0,8; 1,0)	[1,0; 1,2]
mi	3	6	46	120	20	5

Знайти вибіркову дисперсію та вибіркове середнє для тривалості обслуговування.

Варіант 17.

Побудувати групову вибірку, визначити медіану, моду вибірки.

Обчислити вибіркове середнє значення,вибіркову дисперсію та вибіркове середнє квадратичне відхилення.

№ 1

Елементами вибірки є зріст 20 опитаних людей (у см):

172, 175, 173, 172, 177, 182, 179, 175, 176, 173, 169, 174, 173, 178,
180, 179, 176, 172, 173, 176.

Варіант 18.

№ 2

Елементами вибірки є вік респондентів:

20, 25, 37, 28, 34, 22, 23, 20, 17, 18, 19, 25, 71, 35, 42, 39, 28, 31,
24, 55.

Варіант 19.

№ 3

Елементами вибірки є вага респондентів (у кг):

75, 59, 80, 79, 63, 55, 85, 83, 79, 78, 53, 61, 90, 59, 69, 23, 47, 31, 24, 55.

Варіант 20.

№ 4

Елементами вибірки є кількість дітей у сім'ї респондента:

1, 0, 2, 2, 1, 3, 0, 1, 2, 0, 1, 2, 1, 0, 2, 4, 2, 3, 2, 1.

Варіант 21.

№ 5

Елементами вибірки є кількість років, які респондент витратив на освіту:

12, 16, 18, 20, 17, 18, 16, 14, 13, 12, 19, 13, 17, 19, 18, 14, 16, 15, 15, 12.

ПЕРЕЛІК ТЕОРЕТИЧНИХ ПИТАНЬ НА ЕКЗАМЕН З ДИСЦИПЛІНИ: «ЕМПІРИЧНІ МЕТОДИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ».

1. Емпіричні методи в наукових дослідженнях.

2 Об'єкти дослідження емпіричними засобами програмної інженерії.

2.1. Життєвий цикл програмного забезпечення.

2.2. Процеси тестування показників програмних засобів.

2.3. Вимоги до продуктивності та якості пз. Зовнішні і внутрішні характеристики програм.

2.4. Метрики програмного забезпечення.

2.5. Приклад опису значень характеристик і оцінок. Експертне оцінювання і метрики в програмуванні.

3. Статистичний аналіз метрик та експертних оцінок.

3.1. Первинний статистичний аналіз.

3.2. Основні статистичні оцінки для кожної метрики.

3.3. Варіаційний ряд, гістограма.

3.4. Емпірична функція розподілу.

3.5. Числові характеристики вибірки.

3.6. Довірчий інтервал, довірча ймовірність.

3.7. Методи перевірки статистичних гіпотез:

a. Основні поняття теорії гіпотез.
b. Побудова критичної області.
c. Критерій згоди.
d. Перевірка статистичних гіпотез.
e. Критерій Пірсона χ^2.
f. Критерій Колмогорова (χ^2).
g. Критерій знаків.
h. Критерій Вілкоксона.
i. Критерій серій.

3.8. Вибіркове рівняння прямої лінії регресії.

3.9. Метод найменших квадратів.

3.10. Кореляційний аналіз.

3.11. Регресійний аналіз.

3.12. Статичний аналіз коду.

3.13. Інструменти статичного аналізу.

4 Емпіричні методи оцінки надійності програмного забезпечення.

5 Обробка та узагальнення результатів експериментів.

5.1. Методи експертних оцінок.

5.1.1. Методи ранжирування.

5.1.2. Обробка результатів ранжування.

5.1.3. Нормування отриманих оцінок.

5.2. Методи багатомірного шкалування.

6. Методи статистичної обробки результатів експериментального дослідження.

6.1. Особливості характеристик параметрів вимірювання. Види похибок.

6.2. Основні положення теорії похибок.

7. Емпіричні методи дослідження декомпозиції програмних систем, зв'язаності і зчепленості їх компонентів.

7.1. Декомпозиція підсистем на модулі.

7.2. Особливості структурних програм.

7.2.1. Мета структурного програмування.

7.2.2. Програмування з використанням покрокової деталізації.

7.2.3. Низхідне і висхідне програмування.

7.2.4. Модульність.

7.2.5. Інформаційна закритість.

7.3. Зв'язність модуля.

7.4. Слабка зв'язність. Закон Деметри.

7.5. Зчеплення модулів.

8. Емпіричні методи тестування програмного забезпечення.

8.1. Поняття та принципи тестування.

8.2. Тестування „чорного ящика".

8.3. Тестування „білого ящика".

8.3.1. Тестування базового шляху.

8.3.2. Способи тестування умов.

8.3.3. Тестування циклів.

8.4. Налаштування програмного забезпечення.

8.5. Засоби і методи виявлення помилок в ПЗ та його наладки.

8.6. Категорії помилок в програмному забезпеченні.

9. Застосування емпіричних методів на етапах експлуатації та супроводу програмного виробу.

9.1. Етап передачі програмного виробу в експлуатацію.

9.2. Етап планування випробувань.

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДО САМОСТІЙНОЇ РОБОТИ

ВСТУП

Самостійна робота виконується з метою закріплення, поглиблення та узагальнення теоретичних знань, набутих студентами під час вивчення дисципліни, розвитку навичок їх практичного застосування, самостійного та комплексного Розв’язування конкретних фахових завдань. Робота має також за мету навчити користуватися довідковою літературою, таблицями та іншими матеріалами, які фахівець використовує під час своєї професійної діяльності.

Також домашнє завдання надає студентам можливість поглиблення теоретичних та практичних навичок самостійної кваліфікованої праці на рівні фахівця певної галузі діяльності з використанням сучасних комп'ютерних інформаційних технологій при обробці інформації та проведенні обчислень.

Самостійна робота складається з двох частин. ***Перша частина*** складається з двох завдань. Перше завдання присвячено знаходженню характеристик стаціонарного випадкового процесу. В другому завданні вирішується задача знаходження одномірної щільності розподілу стаціонарного випадкового процесу на виході безінерційного пристрою при відомому розподілі процесу на вході та заданій характеристиці пристрою. В ***другій частині*** вирішується задача аналізу ефективності параметричного алгоритму виявлення сигналів. Під час виконання другого завдання вивчаються особливості виявлення інформаційного сигналу на фоні білого гауссового шуму методом накопичення відліків огинаючої випадкового процесу.

Порядок оформлення роботи

Робота оформляється у вигляді звіту на аркушах формату А4 та на електронному носії. Звіт по виконання домашнього завдання кожної окремої частини повинен виконуватися засобами MS Word та містити наступні листи:

1. титульний;
2. завдання;
3. хід виконання роботи.
4. висновки
5. список використаної літератури.

ЧАСТИНА 1

Тема: «Дослідження випадкових процесів»

Перша частина домашнього завдання складається з двох завдань.

Перше завдання присвячено знаходженню характеристик стаціонарного випадкового процесу.

В другому завданні вирішується задача знаходження одномірної щільності розподілу стаціонарного випадкового процесу на виході безінерційного пристрою при відомому розподілі процесу на вході та заданій характеристиці пристрою.

Завдання 1

Характеристики випадкових процесів

Метою даного завдання є дослідження заданого і теоретичного закону розподілу випадкового процесу. Згідно варіанту задана одномірна щільність розподілу ймовірності $f_{\xi}(x)$ стаціонарного у вузькому сенсі випадкового процесу $\xi(x)$.

За заданої щільності розподілу визначити, який це закон розподілу. У результаті побудувати щільність розподілу, функцію розподілу, а також гістограму розподілу випадкового процесу $\xi(x)$. Одномірні щільності розподілу задано в табл. 1. Після проведених досліджень визначити кількісні характеристики випадкового процесу: математичне сподівання, дисперсію, «асиметрія» і ексцес.

Всі розрахунки випадкових процесів призвести в середовищі MathCAD 11/12/13/14.

Порядок вибору варіанта:

Номер варіанту завдання відповідає останній цифрі номеру студентського квитка. Значення заданого параметра розподілу відповідає передостанній цифрі номера студентського квитка (якщо передостання цифра дорівнює нулю, то значення заданого параметра вибирають рівним 10).

Методичні вказівки

Основні співвідношення до виконання завдання 1 приведені в списку рекомендованої літератури: [1-4], [6], [8].

Випадковий (стохастичний) процес характеризує зміну будь-якої випадкової фізичної величини при її спостереженні.

У радіотехніці зазвичай розглядають випадкові процеси, залежні від одного аргументу – від часу, а фізичними величинами є електричні величини – напруга, струм, напруженість поля, фаза та інші.

Для математичного опису випадкового процесу вводиться ***поняття випадкової функції.***

Випадковий процес $X(t)$ може бути представлений сукупністю його реалізацій $x_1(t),...,x_m(t)$, де m- число незалежних дослідів, в результаті яких одержані ці реалізації (рис.1.1)

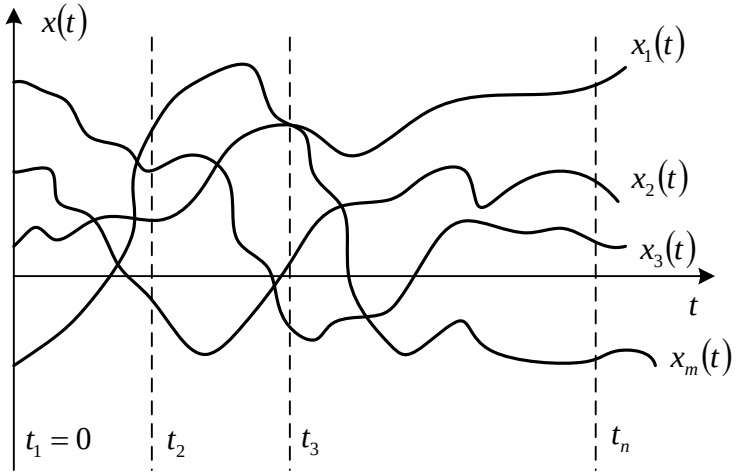


Рис. 1.1 Реалізація випадкового процесу.

Результат спостереження випадкового процесу $x(t)$ у фіксовані моменти часу $t_1,...,t_n$ - це система n випадкових величин $X(t_1),...,X(t_n)$.

Стаціонарний випадковий процес (СВП) - це процес, який протікає в часі приблизно однорідно тобто істотно не змінюються його імовірнісні характеристики. Він має місце в установленому режимі роботи системи при незмінних зовнішніх умовах.

Випадковий процес $X(t)$ називається стаціонарним у вузькому сенсі, якщо всі його багатовимірні щільності розподілу ймовірності $f(x_1,...,x_n,t_1,...,t_n)$ не змінюються при будь-якому зсуві всієї групи точок $t_1,...,t_n$ вздовж осі часу, тобто за будь-яких значеннях n і τ справедливо рівність:

$$f(x_1,...,x_n,t_1,...,t_n)= f(x_1,...,x_n,t_1-\tau,...,t_n-\tau) \qquad (1)$$

Аналогічні рівності повинні виконуватися і для інших імовірнісних характеристик (функцій розподілу, моментних і кореляційних функцій). З виразу (1) випливає, що при I, виконується умова

$$f(x_1,t_1)= f(x_1,t_1-t_1)= f(x_1) \qquad (2)$$

тобто одномірна щільність розподілу ймовірності СВП не залежить від вибраного моменту часу. Для СВП математичне сподівання і дисперсія є константами, тобто

$$m_x(t)= m_x = const \qquad D_x(t)= D_x = const \qquad (3)$$

Умова для математичного очікування не є суттєвою, так як від вихідного процесу завжди можна перейти до центрованого процесу, для якого математичне сподівання тотожно дорівнює нулю і є постійною величиною, тобто процес, нестаціонарний тільки за рахунок змінного математичного очікування, завжди може бути приведений до СВП.

Так як **математичне сподівання** стаціонарного випадкового процесу не залежить від часу його визначають за формулою

$$m = M_{\xi}(t) = \int_{-\infty}^{\infty} x f(x) dx \qquad (4)$$

де $f(x)$ - щільність розподілу ймовірностей випадкового процесу; межі інтегрування $-\infty$ і ∞ визначаються межами області значень випадкового процесу (див. табл. 1).

Формула для обчислення **дисперсії** прийме вигляд:

$$D_{\xi}(t) = \int_{-\infty}^{\infty} [x - m]^2 f(x) dx \qquad (5)$$

Формула для обчислення **асиметрії** СВП прийме вигляд:

$$Sk_{\xi}(t) = \frac{\int_{-\infty}^{\infty} [x - m]^3 f(x) dx}{\sigma_3} = \frac{\mu_3}{\sigma^3}, \qquad (6)$$

де μ_3 - третій центральний момент випадкового процесу.

Формула для обчислення **ексцесу** СВП прийме вигляд:

$$Ex_{\xi}(t) = \frac{\int_{-\infty}^{\infty} [x - m]^4 f(x) dx}{\sigma^3} = \frac{\mu_4}{\sigma^4} - 3, \qquad (7)$$

де μ_4 - **четвертий центральний** момент випадкового процесу, $\sigma = \sqrt{D_{\xi}(t)}$ - **середньоквадратичне відхилення.**

Зв'язок між функцією розподілу $F_{\xi}(x)$ і щільністю розподілу ймовірностей стаціонарного випадкового процесу $f_{\xi}(x)$ визначають співвідношенням

$$F_{\xi}(x) = \int_{-\infty}^{\infty} f_{\xi}(x) dx \qquad (8)$$

Таблиця 1

Щільності розподілу випадкових процесів

Номер варіанта	Щільності розподілу ймовірності $f_\xi(x)$	Область значень випадкових величин	Вказаний параметр
1	**2**	**3**	**4**
0	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$-\infty < x < \infty$	σ
1	$\frac{1}{b}\exp\left(-\frac{x}{b}\right)$	$x \geq 0$	σ
2	$\frac{1}{b}$	$0 \leq x \leq b$	b
1	**2**	**3**	**4**
3	$\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$x \geq 0$	σ
4	$\frac{x}{b}\exp\left(-\frac{x}{b}\right)$	$x \geq 0$	b
5	$\frac{1}{b}$	$1 \leq x \leq b$	b
6	$\frac{x}{b^2}\exp\left(-\frac{x}{b^2}\right)$	$x \geq 0$	b
7	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2}\right)$	$-\infty < x < \infty$	m
8	$\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$	$x \geq 0$	m
9	$\frac{1}{b-a}$	$a \leq x \leq b$	a, b

Завдання 2

Перетворення випадкових процесів

Метою даного завдання є дослідження перетворення випадкових процесів, а саме щільності розподілу на виході безінерційний пристрою.

На безінерційний радіотехнічний пристрій впливає стаціонарний випадковий сигнал $\xi(x)$ і має щільність розподілу $f_\xi(x)$. Знайти в загальному вигляді щільність розподілу $f_\gamma(y)$ сигналу на виході цього пристрою по заданій щільності розподілу ймовірностей розрахованій в завданні 1 (див. табл.1) та заданій характеристиці пристрою (детермінованій функції) $y = q(x)$ (табл. 2).

За результатами проведених теоретичних та практичних досліджень побудувати графік залежності щільності розподілу $f_\gamma(y)$ на виході безінерційний пристрою. В результаті проведених досліджень порівняти вхідну $f_\xi(x)$ та отриману щільність $f_\gamma(y)$ розподілу випадкового процесу на виході цього пристрою.

Порядок вибору варіанта:

Номер варіанта завдання дорівнює числу (m - n), де m - передостання; n - остання цифри номера студентського квитка (табл. 2).

Методичні вказівки

Основні співвідношення до завдання 2 наведені в списку рекомендованої: [1-2], [6]; [8].

Серед нелінійних перетворень безінерційні є найпростішими. Під час неінерційного нелінійного перетворення значення випадкового процесу на виході у будь-який момент часу визначається лише значенням вхідного впливу. Випадок безінерційного перетворення найпростіший при дослідженні нелінійних ланцюгів (рис.1.2). Тоді сигнал на виході ланцюга $\gamma(t)$ визначається значенням вхідного сигналу $\xi(t)$ в той же момент часу t : $\gamma(t)=q(\xi(t))$, де $\gamma(t)$ і $\xi(t)$ - випадкові процеси на вході і виході безінерційного нелінійного кола відповідно; $y=q(x)$ - деяка детермінована функція

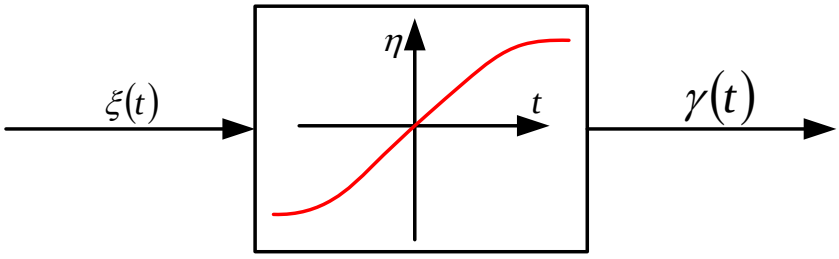


Рис.1.2 Безінерційне нелінійне коло

Для апроксимації функції $q(\xi(t))$ застосовують різні методи. До найбільш вживаних відносяться: поліноміальний, метод апроксимації кусково-ламаній характеристикою, трансцендентними функціями (експонентою, синусоїдою і т.д.).

Припустимо, нам відома щільність розподілу $f_\xi(x)$ випадкової величини $\xi(t)$ і треба знайти щільність розподілу випадкової величини $\gamma=q(\xi)$ в якийсь момент часу t. Припустимо, що існує однозначна зворотна функція $\xi=q^{-1}(\gamma)$. Це справедливо, якщо $q(\xi)$ - монотонно зростаюча або спадна функція. Будемо

при цьому виходити з того, що, якщо величина ξ знаходиться в інтервалі $[x_0, x_0 + \Delta x]$, то величина обов'язково буде в інтервалі, $[y_0, y_0 + \Delta y]$, де $y_0 = q(x_0)$, $y_0 + \Delta y = q(x_0 + \Delta x)$ рис. 1.3.

Тоді рівні і імовірності цих двох подію:

$$f_\gamma(y)\Delta y = f_\xi(x_0)\Delta x \qquad (9)$$

У цьому випадку припускаємо, що інтервали Δx і Δy малі і далі переходимо до межі $\Delta x, \Delta y \to 0$, отримуємо:

$$f_\gamma(y)dy = f_\xi(x)dx \qquad (10)$$

або

$$f_\gamma(y) = f_\xi(x)\frac{dx}{dy} = f_\xi\left(q^{-1}(y)\right)\frac{dq^{-1}(y)}{dy} \qquad (11)$$

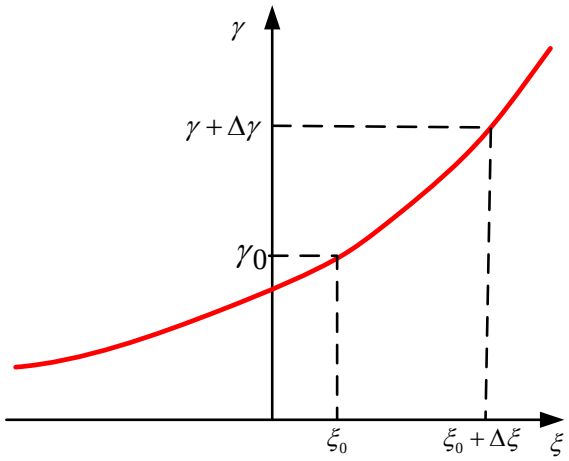


Рис. 1.3. Нелінійне перетворення випадкової величини

Оскільки щільність ймовірності - величина позитивна, а в разі спадної функції $q^{-1}(y)$ похідна буде негативна, то в формулу треба поставити модуль похідної. Таким чином,

$$f_\gamma(y) = f_\xi\left(q^{-1}(y)\right)\left|\frac{dq^{-1}(y)}{dy}\right| \qquad (12)$$

Більш складним є випадок, коли залежність $q(y)$ не є монотонною функцією. У цьому випадку не існує однозначної оберненої функції $q^{-1}(y)$: кожному значенню y відповідає кілька значень x. Нехай будуть дві гілки функції $q^{-1}(y)$: $q_1^{-1}(y)$ і $q_2^{-1}(y)$. В цьому випадку вірогідність попадання на інтервал $[y_0, y_0 + \Delta y]$ дорівнює сумі ймовірностей попадання на інтервали $[x_1, x_1 + dx]$ і $[x_2, x_2 + dx]$:

$$f_\gamma(y)dy = f_\xi dy = f_\xi(x_1)dx_1 + f_\xi(x_2)dx_2 \qquad (13)$$

Висловивши x через y, отримаємо остаточне вираз:

$$f_\gamma(y) = f_\xi\left(q_1^{-1}(y)\right)\left|\frac{dq_1^{-1}(y)}{dy}\right| + f_\xi\left(q_2^{-1}(y)\right)\left|\frac{dq_2^{-1}(y)}{dy}\right| \qquad (14)$$

Якщо гілок оберненої функції багато, то вираз набуде вигляду:

$$f_\gamma(y) = \sum_{i=1}^{n} f_\xi\left(q_i^{-1}(y)\right)\left|\frac{dq_i^{-1}(y)}{dy}\right| \qquad (15)$$

Таблиця 2

Характеристичні функції пристрою

Номер варіанта	Характеристика пристрою, $y = q(x)$
0	$y = \cos(x)$
1	$y = x^2$
2	$y = \ln\|x\|$
3	$y = e^x$
4	$y = 10x + 5$
5	$y = \sin(x)$
6	$y = 1/x$
7	$y = x^4$
8	$y = 2e^x$
9	$y = 2/x$

Контрольні запитання

1. Який випадковий процес називають стаціонарним ?
2. Основні статисти стичні характеристики СВП?
3. Які лінійні перетворення називаються неінерційними?
4. Як визначити одновимірну функцію розподілу ймовірностей на виході без інерційного нелінійного кола?
5. Чи зміниться вид щільності розподілу після безінерційного перетворення?

ЧАСТИНА 2

Тема: «Дослідження параметричних алгоритмів виявлення сигналів»

Мета:

1. Ознайомлення з основними алгоритмами виявлення сигналів.
2. Вивчення особливостей виявлення нормального сигналу на фоні нормального шуму методом накопичення відліків згинаючої випадкового процесу.

3. Вивчення особливостей виявлення нормального сигналу на фоні нормального шуму використовуючи різні критерії прийняття рішення згідно заданому варіанту.

4. Оцінка ефективності алгоритмів виявлення сигналів методом математичного проектування в середовищі MathCAD 11/12/13/14.

Порядок вибору варіанта:

Номер варіанту завдання відповідає цифрі номера студента в журналі (табл.4).

Методичні вказівки

Основні співвідношення до виконання другої частини наведені в літературі: [6-11].

Одним з найважливіших практичних розділів статистичної радіотехніки є можливість розробки алгоритмів виявлення корисних сигналів на фоні завад, та оцінити ефективність роботи.

Задача виявлення сигналів полягає у прийнятті однозначного рішення: або сигнал є (рішення γ_1), або сигналу немає (рішення γ_0).

Ефективність роботи алгоритмів виявлення оцінюється рядом характеристик, до числа яких відносять залежності ймовірностей правильного виявлення, помилкової тривоги і пропуску сигналу від вихідних даних задачі. Перша залежність розраховується як функція відношення сигнал / шум :

$$D = D\left(\frac{P_S}{P_n}\right) \qquad (16)$$

де P_s і P_n - потужності (дисперсії) сигналу і завади.

Найважливішою характеристикою алгоритму виявлення є його ефективність, яка оцінюється пороговим сигналом.

Пороговим сигналом називається те мінімальне відношення сигнал / шум за потужністю $b = P_s/P_n$, яке при фіксованому обсязі вибірки n і заданої ймовірності помилкової тривоги F забезпечує необхідне значення ймовірності правильного виявлення D.

Значення F, D і n визначаються характером завдання, зокрема, в задачах радіолокаційного виявлення зазвичай прагнуть забезпечити

$F = 10^{-7} - 10^{-9}$, $D = 0{,}9 - 0{,}99$, $n = 8,\ 16,\ 32$

Розглянемо алгоритм виявлення з накопиченням відліків огинаючої випадкового процесу на прикладі задачі виявлення ідеального сигналу на фоні нормального некорельованого шуму. Структурна схема виявлення показана на рис.1.4

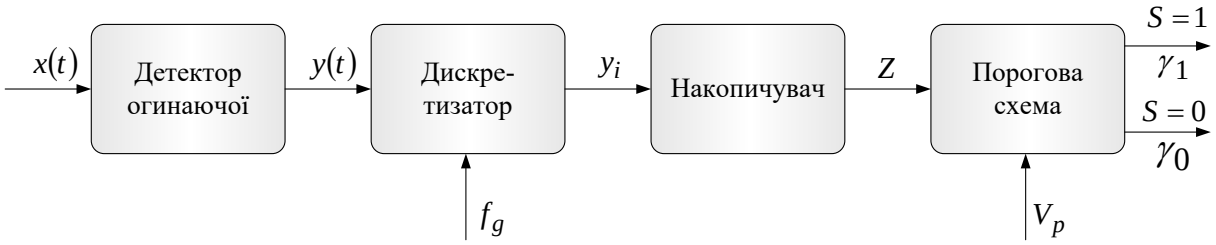


Рис.1.4. Структурна схема параметричного детектора

На вхід детектора огинаючої за відсутності корисного сигналу (S_0) надходить вузькополосний випадковий процес, який представляє собою стандартний (гаусівський) шум з математичний очікуванням і має щільність розподіл ймовірності виду:

$$f(x/S_0)=\frac{1}{\sigma_n\sqrt{2\pi}}exp\left(-\frac{x^2}{2{\sigma_n}^2}\right) \quad , \tag{17}$$

де ${\sigma_n}^2$- дисперсія (потужність) шуму. При наявності на вході, детектора корисного сигналу (S_1) з математичним очікуванням щільність розподілу адитивної суміші сигналу і шуму також має нормальний розподіл:

$$\begin{aligned} f(x/S_1)&=\frac{1}{\sqrt{2\pi\left(\sigma_s^2+{\sigma_n}^2\right)}}exp\left(-\frac{x^2}{2\left(\sigma_s^2+{\sigma_n}^2\right)}\right)= \\ &=\frac{1}{\sqrt{2\pi\left(\sigma_{sh}^2\right)}}exp\left(-\frac{(x-m)^2}{2\sigma_{sh}^2}\right), \end{aligned} \tag{18}$$

де ${\sigma_{sh}}^2$ - дисперсія (потужність) адитивної суміші сигналу і шуму, ${\sigma_s}^2$ - потужність сигналу (потужністю інформативного сигналу виступає його амплітуда).

При виведенні формули (18) використана теорема складання дисперсій: дисперсія суми некорельованих випадкових велич дорівнює сумі дисперсій доданків. Крім того, відомо, що сума нормальних процесів також розподілена за нормальним законом. Розподіл (18) зручно записати у вигляді

$$f(x/S=1)=\frac{1}{\sqrt{2\pi{\sigma_n}^2(1+b)}}exp\left(-\frac{x^2}{2{\sigma_n}^2(1+b)}\right), \tag{19}$$

де $b=\sigma_s^2/{\sigma_n}^2$ - відношення огинаючої потужності сигналу до потужності завади.

На виході детектора виділяється огинаюча вхідного випадкового процесу. Відомо, що щільність розподілу обвідної нормального випадкового процесу при лінійному детектуванні описується законом Релея:

при відсутності сигналу

$$f(y/S_0)=\frac{y}{\sigma_n{}^2}exp\left(-\frac{y^2}{2\sigma_n{}^2}\right) \qquad (20)$$

і при наявності сигналу

$$f(y/S_1)=\frac{y}{\sigma_n{}^2(1+b)}exp\left(-\frac{y^2}{2\sigma_n{}^2(1+b)}\right) \qquad (21)$$

Огинаюча $y(t)$ випадкового процесу надходить на дискретизатор за часом, на виході якого формуються дискретні y_1, відліки амплітуди яких дорівнюють миттєвим значенням огинаючої. Відліки огинаючої визначаються частотою дискретизації f_g. Отримані відліки надходять на накопичувач, який здійснює підсумовування поточних відліків. Вихідна напруга Z, накопичувача в цьому випадку, дорівнює:

$$Z=\sum_{i=1}^{n}y_i \qquad (22)$$

Накопичена сума порівнюється з розрахованим порогом прийняття рішення V_p.

Якщо в результаті порівняння значення суми виявиться більше V_p, то приймається рішення про наявність сигналу (γ_1), в іншому випадку - альтернативне рішення (γ_0), тому що вид рішення залежить від виконання умови

$$Z=\sum_{i=1}^{n}y_i > V_p \qquad (23)$$

Розглянемо задачу оцінки ефективності детектора по схемі рис. 1. Накопичена за вибіркою обсягом сума n (22) називається перевірочною статистикою.

Згідно з центральною граничною теоремою, якщо $y_1,...,y_n$ - це незалежні випадкові величини, то при необмеженому збільшенні n - закон розподілу суми цих величин наближається до нормального. На практиці при $n>10$ закон розподілу суми вважається нормальним. При малих n розподіл суми підпорядковуються закону Ерланга.

На підставі центральної граничної теореми можна записати вираз для щільності розподілу перевірочної статистики Z

$$f(Z)=\frac{1}{\sqrt{2\pi D_Z}}exp\left(-\frac{(Z-m_Z)}{2D_Z^2}\right) \qquad (24)$$

У формулі (24) m_Z і D_Z - математичне очікування і дисперсія статистики:

$$m_Z=n\cdot m_y,\ D_Z=n\cdot D_y \qquad (25)$$

m_y і D_y - математичне очікування і дисперсія дискретних релеївських відліків рівні:

при відсутності сигналу

$$m_y = \sigma\sqrt{\pi/2}\ ,\ D_y = \frac{4-\pi}{2}\sigma^2 \qquad (26)$$

і при наявності сигналу

$$m_{ys} = \sqrt{\frac{\sigma^2(1+b)\pi}{2}}\ ,\ D_{ys} = \frac{4-\pi}{2}\sigma^2(1+b) \qquad (27)$$

На рис. 1.5 показані криві розподілу відліків огинаючої процесу за відсутності та за наявності сигналу.

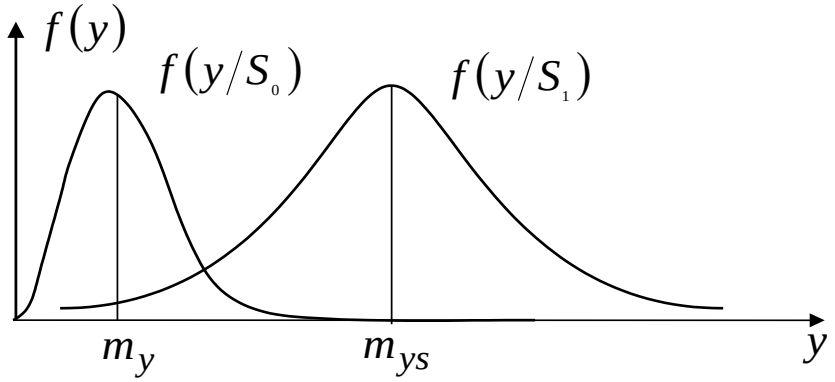


Рис.1.5. Криві розподілу відліків огинаючої процесу за відсутності та за наявності сигналу

На рис. 1.6 – відповідні криви розподілу перевірочної статистики Z. Для прийняття рішення S_1 про те, що на вході детектора є корисний сигнал, необхідно, щоб випадкова величина Z перевищила поріг V_p.

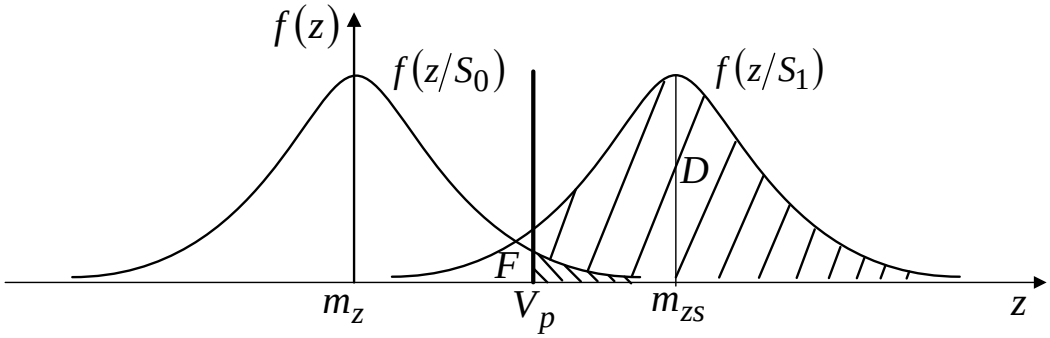


Рис.1.6. Криві розподілу перевірочної статистики Z

Значення порогу при виявлення сигналів вибирають при розрахунках відповідно до критерію Неймана-Пірсона так, щоб ймовірність перевищення його статистикою Z за відсутності сигналу була б не більш наперед заданої. Ця ймовірність F називається ймовірністю помилкової тривоги (помилка 1 роду) (див. мал. 1.6)

$$F = \int_{V_p}^{\infty} f(z/S_0)dz \qquad (28)$$

Підставивши формулу (22) у формулу (28), після спрощення отримаємо

$$F = 1 - \Phi\left[\frac{V_p - m_z}{\sqrt{D_z}}\right]$$

де $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{t^2}{2}\right)dt$ - табульований інтеграл ймовірності.

При заданому значенні ймовірності помилкової тривоги F значення порогу вирішенні V_p може бути знайдено за допомогою таблиць з рівняння (28).

Вірогідність D правильного виявлення сигналу (див.мал.1.6) визначається виразом:

$$D = \int_{V_p}^{\infty} f(z/S_1)dz$$

рівним з урахуванням формули (22)

$$D = 1 - \Phi\left[\frac{V_p - m_{zs}}{\sqrt{D_{zs}}}\right]$$

Для оцінки якості функціонування параметричного алгоритму виявлення сигналу побудуємо значення ймовірності D правильного виявлення сигналу від відношення сигнал/шум b. Характеристики правильного виявлення $D(b)$ для різних обсягів вибірки n показані на рис.1.7.

З графіків видно, що задана ймовірність правильного виявлення $D_{зад}$ при збільшенні обсягу накопичення n може бути досягнута при меншому значенні відношенні сигнал/шум b.

Іншими словами, при заданому b збільшення n забезпечує збільшення ймовірності правильного виявлення сигналів на фоні шумів.

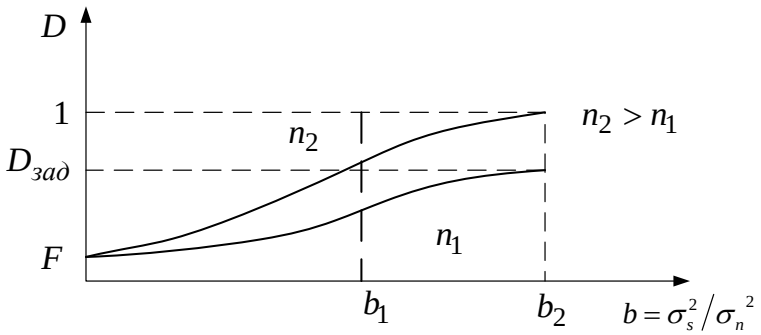


Рис.1.7.Характеристиктики виявлення D(b) в залежності від співвідношення сигнал/шум і обсягів вибірки

Практична частина

У ході виконання роботи треба:

1. сформувати випадковий процес згідно заданому варіанту (табл.4);

2. сформувати інформаційний сигнал згідно заданому варіанту(табл.4);

3. сформувати адитивну суміш;

4. визначити кількісні характеристики (математичне сподівання, дисперсію, середньоквадратичне відхилення) для випадкового процесу, інформаційного сигналу, адитивної суміші;

5. побудувати щільність розподілу і розрахувати вузькополосний випадковий процес на вході детектора огинаючої за відсутності корисного сигналу (S_0);

6. побудувати щільність розподілу і розрахувати корисний сигнал у вигляді адитивної суміші сигналу і шуму на вході детектора огинаючої (S_1);

7. побудувати і розрахувати густину обвідної нормального випадкового процесу при лінійному детектуванні, яке описується законом Релея при відсутності сигналу і при наявності сигналу;

8. побудувати і розрахувати криві розподілу відліків обвідної процесу за відсутності та за наявності сигналу;

9. побудувати і розрахувати криві розподілу перевірочної статистики Z;

10. розрахувати поріг прийняття рішення на основі кривих розподілу перевірочної статистики Z дослідним шляхом для різної кількості відліків перевірочної статистики;

11. на підставі розрахованого порогу прийняття рішення знайти ймовірність правильного прийняття рішення D та ймовірності помилкової тривоги F для різної кількості відліків перевірочної статистики;

12. провести оцінку точності процедури прийняття рішення на підставі побудови графічної залежності ймовірності правильного прийняття рішення $D(b)$ від співвідношення сигнал/шум b і кількості відліків n перевірочної статистики;

13. результати сформувати у вигляді зведеної таблиці (табл.3) і побудувати графічні залежності ймовірності правильного прийняття рішення $D(b)$ від співвідношення сигнал / шум σ_s/σ_1 і кількості відліків перевірочної статистики n_1.

Таблиця 3

Зведена таблиця результатів проведених досліджень

Дані отримані дослідним шляхом	Перевірочна статистика, n_i											
	n_1			n_2			n_3			n_4		
	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3

V_p - поріг прийняття рішення												
F - імовірність помилкової тривоги												
D - імовірність правильного прийняття рішення												

Таблиця 4

Варіанти домашнього завдання (частина 2)

№ п/п	Випадковий процес Нормальний закон				Інформаційний сигнал			*n*-перевірочна статистика (кількість відліків)			
	μ	σ									
		σ_1	σ_2	σ_3	Тип	Амплітуда *мВ*	Структура сигналу				
1.	0	0.5	1	2	Відеоімпульс	2	10101010	2	4	10	100
2.	0	0.5	1	2	Радіоімпульс	4	11101010	2	4	10	100
3.	0	0.5	1	2	Відеоімпульс	6	11111010	2	4	10	100
4.	0	0.5	1	2	Радіоімпульс	10	11111110	2	4	10	100
5.	0	0.5	1	2	Відеоімпульс	12	11111100	2	4	10	100
6.	0	0.5	1	2	Радіоімпульс	15	11101100	2	4	10	100
7.	0	0.5	1	2	Відеоімпульс	20	11001100	2	4	10	100
8.	0	0.5	1	2	Радіоімпульс	2	00001100	2	4	10	100
9.	0	0.5	1	2	Відеоімпульс	4	10001100	2	4	10	100
10.	0	0.5	1	2	Радіоімпульс	6	11001111	2	4	10	100
11.	0	0.5	1	2	Відеоімпульс	10	11000011	2	4	10	100
12.	0	0.5	1	2	Радіоімпульс	12	11011011	2	4	10	100
13.	0	0.5	1	2	Відеоімпульс	15	01011010	2	4	10	100
14.	0	0.5	1	2	Радіоімпульс	20	11101011	2	4	10	100
15.	0	0.5	1	2	Відеоімпульс	4	00101011	2	4	10	100
16.	0	0.5	1	2	Радіоімпульс	5	00111011	2	4	10	100
17.	0	0.5	1	2	Радіоімпульс	6	00111011	2	4	10	100

18.	0	0.5	1	2	Відеоімпульс	8	11111110	2	4	10	100
19.	0	0.5	1	2	Радіоімпульс	10	10111110	2	4	10	100
20.	0	0.5	1	2	Радіоімпульс	12	11110010	2	4	10	100
21.	0	0.5	1	2	Відеоімпульс	14	11010010	2	4	10	100
22.	0	0.5	1	2	Радіоімпульс	16	10000010	2	4	10	100
23.	0	0.5	1	2	Відеоімпульс	20	10010010	2	4	10	100
24.	0	0.5	1	2	Радіоімпульс	2	10011010	2	4	10	100
25.	0	0.5	1	2	Відеоімпульс	4	10111010	2	4	10	100
26.	0	0.5	1	2	Радіоімпульс	6	10111011	2	4	10	100
27.	0	0.5	1	2	Відеоімпульс	8	00111011	2	4	10	100
28.	0	0.5	1	2	Радіоімпульс	10	01111011	2	4	10	100
29.	0	0.5	1	2	Відеоімпульс	12	11011011	2	4	10	100
30.	0	0.5	1	2	Радіоімпульс	14	11011000	2	4	10	100

Контрольні запитання

1. Назвіть основні характеристики алгоритмів виявлення сигналів?
2. Запишіть вираз для щільності розподілу ймовірності шуму та поясніть смисл параметрів розподілу?
3. Запишіть вираз для щільності розподілу ймовірності нормального (ідеального) сигналу та шуму та поясніть смисл параметрів розподілу?
4. Чи залежить ймовірність правильного виявлення від розміру статистичної вибірки при заданому співвідношенні сигнал/шум?
5. Який критерій використовується при оцінці ефективності функціонування алгоритму виявлення?

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ ДО САМОСТІЙНОЇ РОБОТИ

1. *Баскаков С И.* Радиотехнические цепи и сигналы. - М: Высш. шк., 1988. – 448 с.

2. *Баскаков С. И.* Радиотехнические цепи и сигналы. Руководство к решению задач. - М: Высш. шк., 1987. – 207 с.

3. *Бойко 1.Ф., Давлет'янц О. І.* та ін. Статистична радіотехніка: Навч. поібник, – К .: КМУЦА, 1998. – 124 с.

4. *Гоноровский И. С.* Радиотехнические цепи и сигналы. - М, Сов. радио, 1986. – 608с.

5. *Горяинов В. К., Журавлев А. Г., Тихонов В. И.* Статистическая радиотехника: Примеры и задачи. - М.; Сов. радио, 1980. – 544 с.

6. *Заездный А.М.* Основы расчетов по статистической радиотехнике. – М.: Связь, 1969. – 448 с.

7. *Левин Б. Р.* Теоретические основы статистической радиотехники. Кн. 1. - М.: Сов. радио, 1974. – 552 с.

8. *Левин Б. Р.* Теоретические основы статистической радиотехники. Кн. 2. - М: Сов. радио, 1975. – 392 с.

9. *Левин Б. Р.* Теоретические основы статистической радиотехники. Кн. 3. - М: Сов. радио, 1976. – 448 с.

10. *Тихонов В. И.* Статистическая радиотехника. – М.: Сов. радио, 1966. – 678 с.

11. *Тихонов В.И., Харисов В.Н.* Статистический анализ и синтез радиотехнических устройств и систем. – М.: Наук. думка, 1975. – 144с.

ЛІТЕРАТУРА

До розділу: Теорія ймовірності.

1. Гмурман В. Е. Теория вероятностей и математическая статистика: учеб. пособие для бакалавров / В. Е. Гмурман. — 12-е изд. — М.: Издательство Юрайт, 2013. — 479 с. : ил. — Серия : Бакалавр. Базовый курс. [стр. 17–26].
2. Чжун К.Л., АитСахлиа Ф. Элементарный курс теории вероятностей. Стохастические процессы и финансовая математика. — Пер. с англ. — М.: БИНОМ. Лаборатория знаний., 2007. — 455 с. [стр. 31–53].
3. Прохоров Ю.В., Розанов Ю.А. Теория вероятностей (Основные понятия. Предельные теоремы. Случайные процессы). — М.: Наука, 1973. — 494 с. [стр. 39–45].

Статистичний аналіз коду

1. Скринкаст: Статический анализ Си++ кода // Блог компании PVS-Studio, Habrahab.r
2. О безошибочных программах // «Открытые системы», № 07, 2004.
3. Первые шаги к решению проблемы верификации программ // «Открытые системы», № 08, 2006.
4. Сертификация и тестирование программного обеспечения // НПО Эшелон
5. Что такое «Parallel Lint»? // Viva64.
6. Статический анализ безопасности кода // Программная инженерия и информационная безопасность. 2013 № 1, стр 50 119.
7. Лагутин М.Б. Наглядная математическая статистика: учебное пособие — М.: БИНОМ. Лаборатория знаний., 2009. — 472 с. [стр. 71–76].
8. Кривенцов А.С., Ульянов М.В. Интервальная оценка параметров бета-распределения при определении доверительной трудоемкости алгоритмов // Известия ЮФУ. 2012. №7(132). С. 210–219.

Зв’язність.

1. ISO/IEC/IEEE 24765-2010 Systems and software engineering — Vocabulary.
2. Бадд, 1997, 17.1.2. Разновидности связности.
3. Вендров А. М. CASE-технологии. Современные методы и средства проектирования информационных систем. 2.2.3. Типы связей между функциями
4. Пирогов В. Ю. Информационные системы и базы данных: организация и проектирование — СПб, БХВ-Петербург, 2009. С.203-204.
5. ISO/IEC TR 19759:2005, Software Engineering — Guide to the Software Engineering Body of Knowledge (SWEBOK).
6. W. Stevens, G. Myers, L. Constantine, «Structured Design», IBM Systems Journal, 13 (2), 115—139, 1974.

7. Тимоти Бадд Объектно-ориентированное программирование в действии = An Introduction to Object-Oriented Programming. — СПб.: «Питер», 1997. — 464 с. — (В действии). — ISBN 5-88782-270-8.
8. Стив Макконнелл Совершенный код = Code Complete. — 2-е издание. — М.: Русская редакция, 2010. — С. 163-166. — 896 с. — (Мастер-класс). — ISBN 978-5-7502-0064-.1

Математична статистика и програмна інженерія.

1. Петрушин В.Н, Ульянов М.В. Информационная чувствительность компьютерных алгоритмов. — М.: ФИЗМАТЛИТ, 2010. — 224 с. [стр. 138–163].
2. Ульянов М.В., Наумова О.А., Яковлев И.А. Прогнозирование временных оценок для табличного алгоритма решения задачи оптимальной упаковки на основе функции трудоёмкости // Бизнес-Информатика 2008. №3(5). С. 37-46.
3. Головешкин В.А., Петрушин В.Н., Ульянов М.В. Количественные оценки информационной чувствительности алгоритмов // Информационные технологии и вычислительные системы. 2011. №4. С. 45-57.
4. Ульянов М.В., Петрушин В.Н., Кривенцов А.С. Доверительная трудоемкость – новая оценка качества алгоритмов // Информационные технологии и вычислительные системы. 2009. №2. С. 23 – 37.

Додаток А

Таблиця 1 – Квантилі статистики хіта-квадрат

	0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99
1	0,0002	0,0010	0,0039	0,0158	0,0642	0,1485	0,2750	0,4549	0,7083	1,0742	1,6424	2,7055	3,8415	5,0239	6,6349
2	0,0201	0,0506	0,1026	0,2107	0,4463	0,7133	1,0217	1,3863	1,8326	2,4079	3,2189	4,6052	5,9915	7,3778	9,2103
3	0,1148	0,2158	0,3518	0,5844	1,0052	1,4237	1,8692	2,3660	2,9462	3,6649	4,6416	6,2514	7,8147	9,3484	11,3449
4	0,2971	0,4844	0,7107	1,0636	1,6488	2,1947	2,7528	3,3567	4,0446	4,8784	5,9886	7,7794	9,4877	11,1433	13,2767
5	0,5543	0,8312	1,1455	1,6103	2,3425	2,9999	3,6555	4,3515	5,1319	6,0644	7,2893	9,2364	11,0705	12,8325	15,0863
6	0,8721	1,2373	1,6354	2,2041	3,0701	3,8276	4,5702	5,3481	6,2108	7,2311	8,5581	10,6446	12,5916	14,4494	16,8119
7	1,2390	1,6899	2,1673	2,8331	3,8223	4,6713	5,4932	6,3458	7,2832	8,3834	9,8032	12,0170	14,0671	16,0128	18,4753
8	1,6465	2,1797	2,7326	3,4895	4,5936	5,5274	6,4226	7,3441	8,3505	9,5245	11,0301	13,3616	15,5073	17,5345	20,0902
9	2,0879	2,7004	3,3251	4,1682	5,3801	6,3933	7,3570	8,3428	9,4136	10,6564	12,2421	14,6837	16,9190	19,0228	21,6660
10	2,5582	3,2470	3,9403	4,8652	6,1791	7,2672	8,2955	9,3418	10,4732	11,7807	13,4420	15,9872	18,3070	20,4832	23,2093
11	3,0535	3,8157	4,5748	5,5778	6,9887	8,1479	9,2373	10,3410	11,5298	12,8987	14,6314	17,2750	19,6751	21,9200	24,7250
12	3,5706	4,4038	5,2260	6,3038	7,8073	9,0343	10,1820	11,3403	12,5838	14,0111	15,8120	18,5493	21,0261	23,3367	26,2170
13	4,1069	5,0088	5,8919	7,0415	8,6339	9,9257	11,1291	12,3398	13,6356	15,1187	16,9848	19,8119	22,3620	24,7356	27,6882
14	4,6604	5,6287	6,5706	7,7895	9,4673	10,8215	12,0785	13,3393	14,6853	16,2221	18,1508	21,0641	23,6848	26,1189	29,1412
15	5,2293	6,2621	7,2609	8,5468	10,3070	11,7212	13,0297	14,3389	15,7332	17,3217	19,3107	22,3071	24,9958	27,4884	30,5779
16	5,8122	6,9077	7,9616	9,3122	11,1521	12,6243	13,9827	15,3385	16,7795	18,4179	20,4651	23,5418	26,2962	28,8454	31,9999
17	6,4078	7,5642	8,6718	10,0852	12,0023	13,5307	14,9373	16,3382	17,8244	19,5110	21,6146	24,7690	27,5871	30,1910	33,4087
18	7,0149	8,2307	9,3905	10,8649	12,8570	14,4399	15,8932	17,3379	18,8679	20,6014	22,7595	25,9894	28,8693	31,5264	34,8053
19	7,6327	8,9065	10,1170	11,6509	13,7158	15,3517	16,8504	18,3377	19,9102	21,6891	23,9004	27,2036	30,1435	32,8523	36,1909
20	8,2604	9,5908	10,8508	12,4426	14,5784	16,2659	17,8088	19,3374	20,9514	22,7745	25,0375	28,4120	31,4104	34,1696	37,5662
21	8,8972	10,2829	11,5913	13,2396	15,4446	17,1823	18,7683	20,3372	21,9915	23,8578	26,1711	29,6151	32,6706	35,4789	38,9322
22	9,5425	10,9823	12,3380	14,0415	16,3140	18,1007	19,7288	21,3370	23,0307	24,9390	27,3015	30,8133	33,9244	36,7807	40,2894
23	10,1957	11,6886	13,0905	14,8480	17,1865	19,0211	20,6902	22,3369	24,0689	26,0184	28,4288	32,0069	35,1725	38,0756	41,6384
24	10,8564	12,4012	13,8484	15,6587	18,0618	19,9432	21,6525	23,3367	25,1063	27,0960	29,5533	33,1962	36,4150	39,3641	42,9798
25	11,5240	13,1197	14,6114	16,4734	18,9398	20,8670	22,6156	24,3366	26,1430	28,1719	30,6752	34,3816	37,6525	40,6465	44,3141
26	12,1981	13,8439	15,3792	17,2919	19,8202	21,7924	23,5794	25,3365	27,1789	29,2463	31,7946	35,5632	38,8851	41,9232	45,6417
27	12,8785	14,5734	16,1514	18,1139	20,7030	22,7192	24,5440	26,3363	28,2141	30,3193	32,9117	36,7412	40,1133	43,1945	46,9629
28	13,5647	15,3079	16,9279	18,9392	21,5880	23,6475	25,5093	27,3362	29,2486	31,3909	34,0266	37,9159	41,3371	44,4608	48,2782
29	14,2565	16,0471	17,7084	19,7677	22,4751	24,5770	26,4751	28,3361	30,2825	32,4612	35,1394	39,0875	42,5570	45,7223	49,5879
30	14,9535	16,7908	18,4927	20,5992	23,3641	25,5078	27,4416	29,3360	31,3159	33,5302	36,2502	40,2560	43,7730	46,9792	50,8922

Додаток Б

Значення критерію Стьюдента

f	***p***							
	0.80	**0.90**	**0.95**	**0.98**	**0.99**	**0.995**	**0.998**	**0.999**
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619
2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.363	1.795	2.201	2.718	3.105	3.496	4.024	4.437
12	1.3562	1.7823	2.1788	2.6810	3.0845	3.4284	3.929	4.178
13	1.3502	1.7709	2.1604	2.6503	3.1123	3.3725	3.852	4.220
14	1.3450	1.7613	2.1448	2.6245	2.976	3.3257	3.787	4.140
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860	3.732	4.072
16	1.3360	1.7450	2.1190	2.5830	2.9200	3.2520	3.6860	4.0150
17	1.3334	1.7396	2.1098	2.5668	2.8982	3.2224	3.6458	3.965
18	1.3304	1.7341	2.1009	2.5514	2.8784	3.1966	3.6105	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.3253	1.7247	2.08600	2.5280	2.8453	3.1534	3.5518	3.8495
21	1.3230	1.7200	2.2.0790	2.5170	2.8310	3.1350	3.5270	3.8190
22	1.3212	1.7117	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.315	1.705	2.059	2.478	2.778	3.0660	3.4360	3.7060
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0360	3.3962	3.8494
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460

32	1.3080	1.6930	2.0360	2.4480	2.7380	3.0140	3.3650	3.6210
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.9520	3.3479	3.6007
36	1.3050	1.6883	2.0281	2.4345	2.7195	9.490	3.3326	3.5821
38	1.3042	1.6860	2.0244	2.4286	2.7116	3.9808	3.3190	3.5657
40	1.303	1.6839	2.0211	2.4233	2.7045	3.9712	3.3069	3.5510
42	1.320	1.682	2.018	2.418	2.6980	2.6930	3.2960	3.5370

f	**p**							
	0.80	**0.90**	**0.95**	**0.98**	**0.99**	**0.995**	**0.998**	**0.999**
44	1.301	1.6802	2.0154	2.4141	2.6923	3.9555	3.2861	3.5258
46	1.300	1.6767	2.0129	2.4102	2.6870	3.9488	3.2771	3.5150
48	1.299	1.6772	2.0106	2.4056	2.6822	3.9426	3.2689	3.5051
50	1.298	1.6759	2.0086	2.4033	2.6778	3.9370	3.2614	3.4060
55	1.2997	1.673	2.0040	2.3960	2.6680	2.9240	3.2560	3.4760
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.9146	3.2317	3.4602
65	1.2947	1.6686	1.997	2.3851	2.6536	3.9060	3.2204	3.4466
70	1.2938	1.6689	1.9944	2.3808	2.6479	3.8987	3.2108	3.4350
80	1.2820	1.6640	1.9900	2.3730	2.6380	2.8870	3.1950	3.4160
90	1.2910	1.6620	1.9867	2.3885	2.6316	2.8779	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737	3.3905
120	1.2888	1.6577	1.9719	2.3578	2.6174	2.8598	3.1595	3.3735
150	1.2872	1.6551	1.9759	2.3515	2.6090	2.8482	3.1455	3.3566
200	1.2858	1.6525	1.9719	2.3451	2.6006	2.8385	3.1315	3.3398
250	1.2849	1.6510	1.9695	2.3414	2.5966	2.8222	3.1232	3.3299
300	1.2844	1.6499	1.9679	2.3388	2.5923	2.8279	3.1176	3.3233
400	1.2837	1.6487	1.9659	2.3357	2.5882	2.8227	3.1107	3.3150
500	1.2830	1.6470	1.9640	2.3330	2.7850	2.8190	3.1060	3.3100