

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ДВНЗ «ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ»
КАФЕДРА ПРИКЛАДНОЇ МАТЕМАТИКИ ТА ІНФОРМАТИКИ

МЕТОДИЧНІ ВКАЗІВКИ І ЗАВДАННЯ
до лабораторних занять з дисципліни
«СТАТИСТИЧНІ МЕТОДИ ОЦІНЮВАННЯ»
для студентів ОС «Магістр», що навчаються за спеціальністю
121 Інженерія програмного забезпечення
всіх форм навчання

Дрогобич – 2025

УДК 004.4(072)

М 54

Методичні вказівки і завдання до лабораторних занять з дисципліни «Статистичні методи оцінювання» для студентів ОС «Магістр», що навчаються за спеціальністю 121 Інженерія програмного забезпечення всіх форм навчання / уклад. І.А. Назарова. - Дрогобич: ДВНЗ «ДонНТУ», 2025. - 49с.

Укладач:

Ірина НАЗАРОВА, к.т.н.,
доцент кафедри ПМІ

Рецензент:

Віталій ШАМАЄВ, к.т.н.,
доцент кафедри ЕТКІ

Відповідальний за випуск:

Ярослав ДОРОГИЙ, д.т.н.,
в.о. завідувача кафедри ПМІ

Затверджено навчально-методичним відділом ДВНЗ ДонНТУ,
протокол № 6 від 29.04.2025р.

Розглянуто на засіданні кафедри ПМІ, протокол №4 від 3 квітня 2025р.

І.А. Назарова, 2025

ЗМІСТ

Вступ	4
1 Лабораторне заняття. Метод найменших квадратів, нормальні рівняння нелінійної регресії	5
1.1 Завдання до лабораторного заняття	5
1.2 Варіанти індивідуальних завдань	6
1.3 Вказівки до виконання завдань	15
1.4 Зміст звіту	18
1.5 Контрольні питання	19
2 Лабораторне заняття. Розробка множинної регресійної моделі	21
2.1 Завдання до лабораторного заняття	21
2.2 Варіанти індивідуальних завдань	22
2.3 Вказівки до виконання завдань	27
2.4 Зміст звіту	33
2.5 Контрольні питання	34
3 Лабораторне заняття. Розробка алгоритмічної та програмної реалізації ГПВЧ	35
3.1 Завдання до лабораторного заняття	35
3.2 Варіанти індивідуальних завдань	35
3.3 Вказівки до виконання завдань	37
3.4 Зміст звіту	37
3.5 Контрольні питання	43
Список рекомендованої літератури	45

ВСТУП

Методичні вказівки і завдання до лабораторних занять з дисципліни «Статистичні методи оцінювання» призначені для студентів ОС «Магістр», що навчаються за спеціальністю 121 Інженерія програмного забезпечення всіх форм навчання.

Основною метою проведення лабораторних занять є отримання теоретичних знань та практичних навичок в обробці великих обсягів даних, що отримані на основі експерименту/опитування/емуляції тощо.

За структурою методичні вказівки до проведення лабораторних занять містять тему, мету, завдання, індивідуальні варіанти, склад звіту, а також рекомендації до виконання і контрольні питання. За вмістом вказівки присвячені двом розділам статистичної науки, а конкретно – кореляційно-регресійному аналізу та методам Монте-Карло. Головним завданням перших двох занять є розробка економетричних моделей на основі методу найменших квадратів, оцінювання їх невідомих параметрів, перевірка їх на значущість та всебічний аналіз отриманих результатів. Останнє лабораторне заняття присвячене методам генерації псевдо-випадкових чисел, оцінюванню їх спроможності та ефективності.

Таким чином, методичні вказівки містять необхідну інформацію для ефективної організації роботи магістрів в процесі проведення лабораторних занять з дисципліни «СМО», не дублюють відповідний лекційний курс, направлені на опанування засадами нелінійного та багатофакторного кореляційно-регресійного аналізу та на отримання практичних навичок в проектуванні програмних емуляторів псевдо-випадкових чисел, що мають різні закони розподілу.

1 Лабораторне заняття

Метод найменших квадратів, нормальні рівняння нелінійної регресії

Мета заняття: формування теоретичних знань, практичних вмінь та навичок у застосуванні методики побудови і аналізу якості моделей парної регресії на основі методу найменших квадратів.

1.1 Завдання до лабораторного заняття

Відповідно до варіантів індивідуальних завдань (табл. 1.1-1.2), знайти рівняння парних регресій у від x , розрахувати середню помилку апроксимації для кожної з моделей, провести порівняння якості обох моделей.

Зауваження: згідно із завданням необхідно побудувати рівняння парної регресії двох типів: для всіх варіантів обов'язково лінійну та задану варіантом одну з нелінійних регресій.

Для моделі лінійної парної регресії виконати:

- 1) розрахунок вибіркової коваріації та вибіркового коефіцієнта кореляції, оцінити щільність, напрямок та тип зв'язку аналізованих змінних;
- 2) розрахунок коефіцієнта середньої еластичності рівняння парної регресії та коефіцієнта детермінації;
- 3) побудову довірчих інтервалів для коефіцієнтів лінійної регресії з рівнем значущості α ;
- 4) оцінку статистичної значущості коефіцієнтів регресії за допомогою t -критерію Стюдента;
- 5) оцінку статистичної надійності результатів регресійного моделювання за допомогою F -критерію Фішера;

6) розрахунок прогнозного значення результату, якщо прогнозне значення фактора збільшиться на $c\%$ від його середнього.

1.2 Варіанти індивідуальних завдань

Вибір типу рівнянь парної регресії робиться відповідно за таблицею 1.1, рівня значущості α та константи c – за таблицею 1.2.

Таблиця 1.1 – Варіанти вибору типу парної регресії

№ варіанта	Лінійна регресія $y = ax + b$	Степенева регресія $y = ax^b$	Показова регресія $y = ab^x$	Гіперболічна регресія $y = a + b/x$
1	+			+
2	+	+		
3	+			+
4	+			+
5	+		+	
6	+	+		
7	+			+
8	+			+
9	+			+
10	+		+	
11	+	+		
12	+		+	
13	+			+
14	+			+
15	+			+
16	+		+	
17	+	+		
18	+		+	

№ варіанта	Лінійна регресія $y = ax + b$	Степенева регресія $y = ax^b$	Показова регресія $y = ab^x$	Гіперболічна регресія $y = a + b/x$
19	+		+	
20	+			+

Таблиця 1.2 – Варіанти для вибору α та c

№ варіанта	Коефіцієнт значущості, α	Константа c (%)
1	0,05	25
2	0,1	34
3	0,01	32
4	0,05	26
5	0,1	29
6	0,01	45
7	0,05	46
8	0,1	44
9	0,01	42
10	0,01	31
11	0,1	37
12	0,05	39
13	0,01	56
14	0,1	45
15	0,05	46
16	0,01	42
17	0,1	31
18	0,05	35
19	0,1	38
20	0,05	39

Експериментальні дані за індивідуальними варіантами наведено нижче.

1. Досліджується залежність обсягів продажу програмного продукту від витрат на рекламу.

Обсяг продажу y , (тис. грн.)	110	145	112	124	131	221	190	240	235	207
Витрати на рекламу x , (тис. грн.)	1,2	2,5	1,8	1,7	1,9	3,4	4,2	4,1	5,5	5,1

2. Досліджується залежність витрат від обсягів продажу програмних продуктів та послуг.

Витрати ІТ-компаній, y (тис. грн.)	90	20	23	18	36	40	47	100	67	12
Обсяг продажу, x (тис. грн.)	170	67	78	59	90	110	127	195	150	35

3. Досліджується залежність прибутку ІТ-компаній від чисельності персоналу.

Прибуток ІТ-компаній, y (тис. грн.)	330	680	540	1420	314	400	490	674	1200	279
Чисельність персоналу, x (чол.)	100	150	130	190	90	110	120	155	187	98

4. Досліджується залежність середньомісячної продуктивності праці розробників програмного коду на мові програмування Python від їх віку.

Продуктивність праці, y (тис. грн.)	12	8	13	14	18	19	16	17	12	15
Вік програміста, x (років)	22	25	23	29	30	28	32	31	26	35

5. Досліджується залежність між ціною на однорідні програмні продукти та прибутком ІТ-компанії від їх продажу.

Прибуток ІТ-компанії, y (тис. грн.)	12	20	19	22	27	29	16	17	34	35
Ціна продукту, x (тис. грн.)	5	10	7	12	14	18	7	11	45	42

6. Досліджується залежність коефіцієнту прискорення паралельного алгоритму від кількості процесорів високопродуктивної архітектури, що реально задіяні при вирішенні завдання за цим алгоритмом.

Кількість процесорів, y	51	24	55	43	32	58	16	34	19	41
Прискорення паралельного алгоритму, x	15	10	18	14	12	18	7	14	5	17

7. Досліджується залежність коефіцієнту ефективності алгоритму від кількості процесорів багатопроцесорної паралельної обчислювальної системи.

Кількість процесорів, y (тис.)	51	24	55	43	32	58	16	34	19	41
Коефіцієнт ефективності, x	0.65	0.1	0.8	0.4	0.5	0.9	0.07	0.4	0.5	0.7

8. Досліджується залежність обсягу програмного коду від загального часу на його розробку для десяти юніорів ІТ-компанії.

Час на розробку, y (час.)	9	6	5	4	3,7	3,6	3,5	6	7	3,5
Обсяг програмного коду, x (тис. рядків)	10	20	30	40	50	60	70	15	12	25

9. Досліджується залежність продуктивності праці тестувальника програмного коду від рівня автоматизації робіт.

Продуктивність праці y , (т/ч)	7	22	30	29	39	15	32	22	20	20
Рівень автоматизації, x (%)	10	25	56	38	75	23	66	34	17	21

10. Досліджується залежність коефіцієнту ефективності паралельного алгоритму Кеннона для матричного множення на суперкомп'ютері від порядку вхідних та результуючих матриць.

Порядок матриці, x	Коефіцієнт ефективності, y
1600	0,466
7100	0,933
6750	0,883
6130	0,801
6110	0,848
4190	0,730
3850	0,514
3680	0,566
3650	0,717
2680	0,672
2600	0,626

11. Досліджується залежність частки операцій міжпроцесорного обміну для паралельного алгоритму множення квадратних матриць Фокса від порядку матриць.

Порядок матриць, x (тис.)	35	13	21	23	18	26	16	30	13	32
Частка операцій міжпроцесорного обміну, y (%)	65	21	21	65	44	87	22	75	25	75

12. Досліджується залежність частки арифметичних операцій паралельного алгоритму Фокса з блоковим розподілом даних для множення квадратних матриць від порядку матриць.

Порядок матриць, x (тис.)	24	55	60	13	30	81	68	41	82	41
Частка арифметичних операцій, y (%)	36	24	17	7	15	49	45	19	52	40

13. Досліджується залежність коефіцієнту ефективності паралельного алгоритму Штрассена-Вінограда швидкого рекурсивного матричного множення на суперкомп'ютері від порядку вхідних та результуючих матриць.

Порядок матриць, x	Коефіцієнт ефективності, y
1200	0,50
1290	0,57
1400	0,60
1500	0,65
1800	0,76
1965	0,71
2100	0,85
1900	0,72
2200	0,62
2625	0,87
1200	0,34

14. Досліджується залежність коефіцієнту ефективності використання паралельного обладнання для алгоритму матричного множення зі стрічковим горизонтальним розподілом даних при множенні щільно заповнених матриць на суперкомп'ютері з розподіленою пам'яттю від порядку вхідних та результуючих матриць.

Порядок матриць, x	Коефіцієнт ефективності, y
2160	0,91
1700	0,57
1930	0,90
1498	0,51
1645	0,76
1632	0,41
1701	0,85
1605	0,62
1805	0,82
1711	0,71
2180	0,95

15. Досліджується залежність коефіцієнту прискорення паралельного алгоритму вирішення систем лінійних алгебраїчних рівнянь (СЛАР) методом Гауса від порядку системи для однакової кількості процесорів розподіленої архітектури.

Порядок СЛАР, x	Коефіцієнт прискорення, y
1600	14,9
7100	11,7
6750	11,7
6130	18,8
6110	10,7
4190	10,9
3850	34,8
3680	41,7
3650	22,8
2680	20,7
2600	17,7

16. Досліджується залежність коефіцієнту прискорення паралельного алгоритму вирішення систем нелінійних алгебраїчних рівнянь методом Ньютона від порядку системи для однакової кількості процесорів розподіленої архітектури.

Порядок СНАР, x	Коефіцієнт прискорення, y
56000	365,69
44000	307,29
30000	247,06
13000	103,89
14000	172,05
53000	360,41
35000	218,74
29000	503,91
13000	110,98
15000	120,34
66000	388,86

17. Досліджується залежність споживчих витрат розробників програмних продуктів (в умовних одиницях (у.о.) та в розрахунку на одну людину) від грошових доходів (також умовних одиницях та в розрахунку на одну людину) для десяти ІТ-фірм.

Споживчі витрати в розрахунку на одну людину, y (у.о.)	Грошові доходи на одну людину, x (у.о.)
385	954
376	981
380	1006
338	994
370	971
358	983
361	954
356	980
372	979
369	960

18. Досліджується залежність споживчих витрат для QA інженерів (в умовних одиницях (у.о.) та в розрахунку на одну людину) від їх

середньої заробітної плати (також умовних одиницях та в розрахунку на одну людину) для десяти ІТ-фірм.

Споживчі витрати в розрахунку на одну людину, y (у.о.).	Середня заробітна плата, x (у.о.)
495	1010
420	995
480	1130
475	1002
415	965
397	997
423	981
445	1005
490	1120
462	1003

19. Досліджується залежність коефіцієнту прискорення паралельного алгоритму вирішення систем звичайних диференціальних рівнянь методом Рунге-Кутти від порядку системи для однакової кількості процесорів розподіленої архітектури.

Порядок СЗДР, x	Коефіцієнт прискорення, y
90000	109
80000	106
70000	105
60000	103
50000	100
40000	85
30000	69
20000	63
10000	52
5000	45
4000	30

20. Досліджується залежність коефіцієнту ефективності паралельного алгоритму вирішення системи звичайних диференціальних

рівнянь методом локальної екстраполяції Річардсона від кількості процесорів паралельної системи.

Кількість процесорів, x	Коефіцієнт ефективності, y
32768	0,92
16384	0,86
8192	0,83
4096	0,76
2048	0,72
1024	0,65
512	0,69
256	0,61
128	0,52
32	0,41

1.3 Вказівки до виконання завдань

При виконанні розрахунків для прикладів завдань до лабораторних робіт використовується редактор електронних таблиць Microsoft Excel.

Рівняння парної регресії може бути подане у наступному вигляді:

$$y = f(\underline{\beta}, x) + u,$$

де y – залежна, пояснювана, ендогенна змінна;

x – незалежна, факторна, пояснююча, екзогенна змінна;

$\underline{\beta}$ – коефіцієнти регресії;

u – випадкова помилка.

Наприклад, вхідні вибіркові дані x та y наведено на рис. 1.1, обсяг вибірки дорівнює $n = 7$.

N	x	y
1	1	3,4
2	2	5
3	7	12
4	5	11

5	12	21
6	14	23
7	3	4,8

Рисунок 1.1 – Вхідні вибіркові дані

Основні статистичні характеристики за вхідними даними розраховуються відповідними функціями Microsoft Excel:

- 1) вибіркові математичні очікування: $m_x^* = \underline{x} = 6,29$, $m_y^* = \underline{y} = 11,46$;
- 2) вибіркові дисперсії: $D_x^* = S^2(x) = 21,63$, $D_y^* = S^2(y) = 53,68$;
- 3) вибіркова коваріація: $K_{xy}^* = cov(x, y) = 33,81224$.

За отриманими даними необхідно здійснити підрахунок оцінок коефіцієнтів, по-перше, лінійної регресії за формулами, які виведені з використанням методу найменших квадратів:

$$\hat{\beta}_1 = \frac{K_{xy}^*}{D_x^*}, \hat{\beta}_0 = m_y^* - \hat{\beta}_1 m_x^*.$$

Для даного прикладу: $\hat{\beta}_1 = 1,563019$, $\hat{\beta}_0 = 1,632453$.

Таким чином, після округлення до 2 знаків після коми рівняння лінійної парної регресії має вигляд:

$$\hat{y} = 1,63 + 1,56x.$$

Розраховані за регресійною моделлю значення $\hat{y}$ позначаються $\hat{y}$ (рис. 1.2).

N	x	y	$\hat{y}$
1	1	3,4	3,19547 2
2	2	5	4,75849 1
3	7	12	12,5735 8
4	5	11	9,44754 7
5	12	21	20,3886 8

6	14	23	23,5147 2
7	3	4,8	6,32150 9

Рисунок 1.2 – Експериментальні та розраховані дані

Середня похибка апроксимації в умовах прикладу складає:

$$\underline{A} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = 9,51\%.$$

Результати розрахунку вибіркового коефіцієнта кореляції:

$$r_{xy}^* = \frac{K_{xy}^*}{\sigma_x^* \sigma_y^*} = 0,992261.$$

Отримане значення вибіркового коефіцієнта кореляції (близьке до одиниці) та середня похибка апроксимації говорять про наявність високої позитивної кореляції між змінними x та y та наближений до лінійного позитивний зв'язок.

Для розрахунку коефіцієнта середньої еластичності рівняння парної регресії використовуємо формулу:

$$\underline{E} = f'(x) \cdot \frac{m_x^*}{m_y^*} = 1,56 \cdot \frac{6,29}{11,46} = 0,86,$$

тобто при зміні середнього значення факторної змінної x на 1% середнє значення результуючої змінної y зміниться на 0,86%.

Довірчий інтервал за визначенням – це такий інтервал, у який із заданою довірчою ймовірністю потрапить значення параметра генеральної сукупності.

Побудова довірчих інтервалів для коефіцієнтів регресії складається з наступних основних етапів:

- 1) встановлюється рівень значущості α ;
- 2) обчислюється оцінка дисперсії випадкового відхилення:

$$\sigma_u^2 = \frac{\sum_{i=1}^n u_i^2}{n - 2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}.$$

3) обчислюється величина квадрата стандартної помилки коефіцієнтів:

$$D(\hat{\beta}_0) = \frac{\sigma_u^2 \cdot \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - m_x^*)^2}, \quad D(\hat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - m_x^*)^2}.$$

4) довірчі інтервали для точних значень коефіцієнтів (β_0, β_1) регресії можна записати у вигляді:

$$I_\gamma[\beta_0] = [\hat{\beta}_0 - t_\gamma \cdot \sigma(\hat{\beta}_0); \hat{\beta}_0 + t_\gamma \cdot \sigma(\hat{\beta}_0)], \\ I_\gamma[\beta_1] = [\hat{\beta}_1 - t_\gamma \cdot \sigma(\hat{\beta}_1); \hat{\beta}_1 + t_\gamma \cdot \sigma(\hat{\beta}_1)],$$

де t_γ – квантиль розподілу Стюдента, для двосторонньої області обчислюється як: $t_\gamma = t_{1-\frac{\alpha}{2}}(n - k - 1) = t_{1-\frac{\alpha}{2}}(n - 2)$.

Результати розрахунку довірчих інтервалів для заданого прикладу при $\alpha=0,1$:

$$\sigma_u^2 = 1,158581, t_{0,95}(5) = 2,02, \\ \sigma(\hat{\beta}_0) = 0,683962, \sigma(\hat{\beta}_1) = 0,08747: \\ I_\gamma[\beta_0] = [0,250849; 3,014057], \\ I_\gamma[\beta_1] = [1,386329; 1,739708].$$

Коефіцієнт детермінації розраховується за формулою:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - m_y^*)^2} = 0,98,$$

що свідчить про високу якість побудованого рівняння регресії.

Статистична надійність регресійної моделі в цілому оцінюється за допомогою F -критерію Фішера. F -тест полягає у перевірці основної гіпотези H_0 про статистичну незначущість рівняння регресії та показника щільності зв'язку. Для цього виконується порівняння фактичного та табличного/критичного значень F -критеріїв:

$$F = \frac{R^2}{1 - R^2} (n - 2), \\ F_{kp} = F_{1-\alpha}(k, n - k - 1).$$

Для розглянутого прикладу за таблицею Фішера: $F_{kp} = 4,06$ при $k = 1, n = 7, \alpha = 0$, а фактичне значення дорівнює $F^* = 319,31$. Оскільки $F^* >$

$F_{кр}$, то H_0 відхиляється, тобто визнається статистична значимість та надійність рівняння регресії.

Побудоване рівняння регресії може бути використано для прогнозування. Припускається, що значення факторної змінної збільшилося на $c = 30\%$ від середнього значення. Тоді, відповідно до рівняння регресії $y = 1,63 + 1,56x$:

$$f(\underline{x} \cdot 1,3) = f(6,29 \cdot 1,3) = f(8,18) = 14,42.$$

1.4 Зміст звіту

Звіт з лабораторного заняття повинен бути оформлений відповідно до ДСТУ-3008:15 і містити:

- титульну сторінку;
- таблицю з вхідними даними;
- для двох видів парної регресії (згідно з варіантом):
 - 1) розрахунок основних статистичних показників;
 - 2) знаходження вибірових коефіцієнтів регресії;
 - 3) побудову ряду розрахункових значень даних у відповідно до отриманої залежності;
 - 4) розрахунок середньої помилки апроксимації;
 - 5) побудову точкової діаграми з вхідними даними, отриманими даними та лінією тренду (для всіх видів діаграм крім гіперболічної);
- для моделі лінійної парної регресії, що побудована, виконати:
 - 1) розрахунок вибіркової коваріації та вибіркового коефіцієнта кореляції, оцінити щільність зв'язку аналізованих змінних;
 - 2) розрахунок коефіцієнта середньої еластичності рівняння парної регресії;
 - 3) побудову довірчих інтервалів для коефіцієнтів лінійної регресії з рівнем значимості;

- 4) оцінку статистичної значущості коефіцієнтів регресії за допомогою t -критерію Стюдента;
- 5) оцінку статистичної надійності результатів регресійного моделювання за допомогою F -критерію Фішера;
- 6) розрахунок коефіцієнта детермінації;
- 7) розрахунок прогнозного значення результативної змінної y , якщо прогнозне значення факторної змінної x збільшиться на $s\%$ від її середнього рівня.

1.5 Контрольні питання

1. Описати регресійну модель з одним рівнянням.
2. Дати визначення ендогенних змінних регресійних моделей.
3. Дати визначення екзогенних змінних регресійних моделей.
4. Які основні сфери застосування регресійних моделей?
5. Описати модель парної регресії.
6. Що характеризує коефіцієнт еластичності?
7. Які характеристики використовують для оцінки щільності зв'язку між аналізованими змінними?
8. Дати визначення та записати формулу для обчислення коефіцієнта детермінації.
9. Як можна оцінити якість рівняння регресії?
10. Яку гіпотезу формулюють для оцінки статистичної значущості коефіцієнтів регресії та кореляції?
11. Дати визначення довірчого інтервалу.
12. Сформулювати теорему Гауса-Маркова.
13. Як за допомогою моделі парної регресії можна виконувати прогноз?
14. Види нелінійної регресії. Способи лінеаризації.
15. Нелінійна модель регресії, нелінійна регресія за параметрами.

Література [1-3]

2 Лабораторне заняття

Розробка множинної регресійної моделі

Мета заняття: засвоєння теоретичного матеріалу та набуття практичних навичок у застосуванні засад регресійного параметричного багатовимірної аналізу, розглянути способи визначення та усунення мультиколінеарності емпіричних даних багатовимірної регресійної моделі.

2.1 Завдання до лабораторного заняття

1. Для моделі, заданої відповідно до варіанта, побудувати рівняння множинної лінійної регресії:

- за допомогою рішення системи нормальних рівнянь (СНР) у координатній формі (МНК), рішення СНР навести у загальному вигляді через відомі початкові та центральні моменти;

- за допомогою операцій матричної алгебри;

- за допомогою вбудованої функції Excel.

2. Побудувати рівняння множинної регресії у «стандартизованому масштабі», оцінити ступінь впливу кожного чинника.

3. Скласти рівняння лінійної парної регресії, залишивши по черзі кожен із факторів та фактор-фактор.

4. Для даних та рівняння множинної регресії, провести перевірку мультиколінеарності (розрахунок коефіцієнтів кореляції між пояснювальними змінними обов'язковий).

5. Усунути кореляцію в залежності від варіанта одним із наступних методів:

- варіанти 1, 4, 7, 10, 13, 16, 19 – метод виключення пояснюючої змінної;

- варіанти 2, 5, 8, 11, 14, 17, 20 – метод збільшення обсягу вибірки;

– варіанти 3, 6, 9, 12, 15, 18 – спосіб зміни специфікації моделі (або інший вид рівняння регресії (нелінійний), або введення додаткових пояснюючих змінних).

Додаткові дані, які потрібні для збільшення обсягу вибірки і методу зміни специфікації моделі, записати самостійно.

2.2 Варіанти індивідуальних завдань

1. Досліджується залежність валового національного продукту Y (млрд. у.о.) за 10 років від споживання $X1$ (млрд. у.о.) та інвестиціями $X2$ (млрд. у.о.).

Y	8	9,5	11	12	13	14	15	16,5	17	18
$X1$	1,65	1,8	2,0	2,1	2,2	2,4	2,65	2,85	3,2	3,55
$X2$	14	16	18	20	23	23,5	25	26,5	28,5	30

2. Досліджується залежність обсягу пропозиції Y деякого продукту фірми в умовах конкуренції від ціни $X1$ цього продукту та заробітної плати $X2$ співробітників фірми, що виробляють цей продукт.

Y	10	15	20	25	40	37	43	35	38	55	50	35
$X1$	12	10	9	9	8	8	6	4	4	5	3	1
$X2$	20	35	30	45	60	69	75	90	105	120	130	130

3. Досліджується прибутковість підприємств Y (оцінюється за 5-ти бальною шкалою) в залежності від числа працівників $X1$ (тис. чол.) і витрат на рекламу $X2$ (у.о.).

Y	2	3	3	3	4	4	5	5	5	5
$X1$	4	4	4	3	5	6	7	6	5	6
$X2$	1	1	2	2	3	3	4	4	5	5

4. Досліджується залежність продуктивності праці Y від рівня механізації праці $X1$ (%) та середнього віку працівника $X2$ (років).

Y	32	30	36	40	41	47	56	54	60	55	61	69
$X1$	33	32	41	39	46	43	34	38	42	35	39	40
$X2$	20	24	28	30	31	33	34	37	38	40	41	45

5. Досліджується залежність прибутковості підприємств Y (тис. у.о.) в залежності від витрат на заробітну плату працівників $X1$ (тис. у.о.) і витрат на оренду виробничих приміщень $X2$ (тис. у.о.).

Y	2	3	3,5	3,67	3,89	4	4,1	4,12	4,15	4,4
$X1$	1,5	1,55	1,68	1,8	1,97	2,01	2,11	2,03	2	2,34
$X2$	25	27	27,5	29	31	34	35	37	38	41

6. Досліджується залежність дохідності підприємств Y (тис. у.о.) в залежності від витрат на рекламу $X1$ (тис. у.о.) і витрат на енергоносії $X2$ (тис. у.о.).

Y	0,1	0,12	0,22	0,24	0,2	0,21	0,23	0,25	0,27	0,24
$X1$	0,8	0,97	0,89	0,99	1,1	1,08	1,06	0,99	1,0	1,12
$X2$	10	11	13	13,7	16	17,4	18	18,5	19	21

7. Вивчається залежність коефіцієнту ефективності паралельних обчислень динамічного завдання Y від розміру входу $X1$ (тисяч одиниць) та кількості процесорів паралельної архітектури, на якій його реалізовано $X2$ (тисяч одиниць).

Y	40	70	80	100	130	200	300	400	500	600
$X1$	10	15	17	25	20	50	70	100	130	300
$X2$	0,61	0,65	0,67	0,68	0,7	0,8	0,83	0,84	0,85	0,88

8. Досліджується попит Y на нерухомість на ринку первинного житла в залежності від ціни за кв. м. $X1$ (у.о.) та району будівництва об'єкта $X2$ (умовно позначимо: 3 – центральний район міста, 2 – прилеглий до центрального району, 1 – спальний віддалений район).

Y	900	990	1050	1200	1300	1450	1700	1800	1900	2000
$X1$	1	1	2	1	3	2	3	3	2	3
$X2$	34	37	44	47	55	56	60	63	67	70

9. Досліджується попит на тури Y для туристичної фірми в залежності від ціни $X1$ (у.о.) та пори року $X2$ (1 – осінь, 2 – весна, 3 – зима, 4 – літо).

Y	400	500	560	575	540	600	650	700	720	750
$X1$	1	2	3	2	2	3	4	4	4	4
$X2$	28	29	33	31	32	34	42	41	40	45

10. Вивчається залежність коефіцієнту ефективності паралельних обчислень динамічного завдання Y від розміру входу $X1$ та кількості процесорів паралельної архітектури, на якій його реалізовано $X2$ (тисяч одиниць).

Y	3400	3700	3800	5400	5300	6000	6800	8000	8100	9600
$X1$	10	15	17	25	20	50	70	100	130	300
$X2$	0,67	0,76	0,676	0,82	0,93	0,84	0,73	0,61	0,62	0,4

11. За підприємствами концерну вивчається залежність вироблення продукції на одного працівника Y (тис. грн) від введення в дію нових основних фондів $X1$ (% від вартості фондів на кінець року) і від ваги робочих високої кваліфікації в загальній кількості робочих $X2$ (%).

Y	3,9	4	4,8	4,4	6,8	6,4	7,2	8,2	8,5	9,0
X1	14	16	19	20	20	22	25	29	31	36
X2	7,0	7,0	7,0	8,0	10,0	11,0	11,0	12,0	12,0	14,0

12. Досліджується залежність попиту на свинину Y залежно від ціни на неї $X1$ (грн.) та від ціни на яловичину $X2$ (грн.).

Y	27,0	29,0	30,0	29,8	31,0	32,0	32,5	33,0	34,5	37,0
X1	23,0	24,0	23,5	24,0	26,0	26,0	27,5	29,0	31,0	32,0
X2	58	61	59,5	63	66	67	71	74,5	81	83

13. За підприємствами регіону вивчається залежність вироблення продукції на одного працівника Y (у. о.) від введення в дію основних фондів $X1$ (% від вартості фондів) та від питомої ваги робітників високої кваліфікації у загальній чисельності робітників $X2$ (%).

Y	2,87	3,3	3,8	5,3	5,36	6,0	6,8	8,1	8,34	9,5
X1	12	14	17	18	21	22	24	29	31	34
X2	6,0	7,0	7,0	8,0	8,0	9,0	9,0	12,0	13,0	14,0

14. За областями країни вивчається залежність середньорічного доходу Y (тис. грн.) від частки зайнятих важкою фізичною працею в загальній чисельності зайнятих $X1$ (%) та від частки економічно активного населення в чисельності всього населення $X2$ (%).

Y	31	29	28	33	35	37	25	24	27	34
X1	47	49	52	41	44	41	36	39	41	48
X2	52	50	51	48	54	53,07	34,5	35	41	55

15. За областями країни вивчаються залежність середньорічного доходу Y (тис. грн.) від частки зайнятих в сфері

обслуговування в кількості зайнятих X_1 (%) і частки економічно активного населення від всієї чисельності населення X_2 (%).

Y	30	22	25	31	32	39	26	47	29	24
X_1	42	54	35	40	44	55	39	31	45	33
X_2	227,1	228,3	168,05	225,41	248	336	190	286	233	153

16. Для сукупності підприємств концерну вивчається залежність прибутку Y (тис. грн.) від вироблення продукції на одного працівника X_1 (од.) та індексу цін на продукцію X_2 (%).

Y	52	54	45	39	74	76	77	74	78	79
X_1	36	32	41	44	39	38	42	39	34	33
X_2	20	21	22,5	24	26	26,6	29	31	33,4	34

17. Для підприємств, що випускають продукцію А, вивчається залежність споживання електроенергії Y (тис. квт ч.) від виробництва продукції X_1 (тис. од.) та рівня механізації робіт – X_2 (%).

Y	52	57	51	53	44	38	37	33	27	29
X_1	32	39	35	31	40	43	36	38	28	30
X_2	37	35	34	31	30	28	27	25	21	17

18. За 10 господарствами досліджується споживання м'яса Y (кг на одиницю населення) від доходу X_1 (грн. на одного члена господарства) та від споживання риби X_2 (кг на одиницю населення) на місяць.

Y	1010	1200	1350	950	1205	1800	2010	2000	1670	2500
X_1	2	2,5	3	5	4	3	3,2	2,9	3,4	3,1
X_2	9,3	11	12,3	7,5	10,4	16	18	19	15	20

19. Вивчається залежність на підприємствах концерну споживання матеріалів Y (т) від енергоозброєності праці X_1 (Квт/годин на одного робітника) та обсягу виробленої продукції X_2 (тис. од.)

Y	63	57	51	56	50	52	43	37	36	32
X_1	40	35	31	38	34	30	39	42	35	37
X_2	39	37	36	34	33	30	29	26	27	20

20. Досліджується попит на нерухомість на ринку первинного житла Y в залежності від ціни за кв. м X_1 (у.о.) і середнього доходу населення X_2 (у.о. в міс.).

Y	800	850	910	970	1000	990	965	1100	1200	1330
X_1	550	590	630	750	1000	650	590	1050	690	1500
X_2	30	29	27	33	37	31	18	32	16	45

2.3 Вказівки до виконання завдань

Рівняння множинної регресії може бути подане у вигляді:

$$Y = f(\underline{\beta}, X) + U,$$

де $X = (X_1, X_2, \dots, X_m)$ – вектор незалежних (пояснюючих/екзогенних) змінних, m – кількість факторів,

$\underline{\beta}$ – вектор параметрів (ендогенних змінних, що підлягають визначенню),

U – випадкова помилка (відхилення),

Y – залежна, пояснювана змінна.

Представимо вхідні дані та відповідні коефіцієнти в матричній формі (n – кількість експериментів/ спостережень):

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_m \end{bmatrix}, u = \begin{bmatrix} u_0 \\ u_1 \\ \dots \\ u_n \end{bmatrix}.$$

Коефіцієнти (значення ендогенних змінних) розраховують згідно з МНК за формулою:

$$\underline{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y.$$

Метод найменших квадратів може бути застосований і до рівняння множинної регресії в «стандартизованому масштабі»:

$$t_y = \alpha_1 t_{x_1} + \alpha_2 t_{x_2} + \dots + \alpha_m t_{x_m} + \varepsilon,$$

де $t_y, t_{x_1}, t_{x_2}, \dots, t_{x_m}$ – стандартизовані змінні: $t_y = \frac{y - \bar{y}}{\sigma_y}$, $t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$, які

отримані зі змінних на основі операцій центрування та нормування, тому для них середнє значення дорівнює нулю: 0, а середнє квадратичне відхилення дорівнює одиниці: 1, α_i – стандартизовані коефіцієнти регресії.

З огляду на те, що всі змінні центровані та нормовані, стандартизовані коефіцієнти регресії можна порівнювати між собою. Порівнюючи їх один з одним, можна ранжувати фактори за силою їхнього впливу на результат. У цьому основна перевага стандартизованих коефіцієнтів регресії на відміну коефіцієнтів «чистої» регресії, які незрівнянні між собою.

Коефіцієнти «чистої» регресії β_i та стандартизовані коефіцієнти регресії α_i пов'язані між собою зі таким чином:

$$\beta_i = \alpha_i \frac{\sigma_y}{\sigma_{x_i}}.$$

Розраховують стандартні помилки коефіцієнтів регресії за формулами:

$$\sigma(\hat{\beta}_j) = \sqrt{\frac{\sum u_i^2}{n-m-1} \cdot z_{jj}^{-1}},$$

$$Z^{-1} = (X^T \cdot X)^{-1}.$$

Для роботи з матрицями Microsoft Excel використовують такі вбудовані функції.

«МУМНОЖ» – множення матриць.

Параметри: (масив1; масив2).

Приклад множення матриць $C = A * B$ наведено на рис. 2.1.

	A	B	C	D
1				
2		$\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}_A$		$\begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}_B$
3				
4				
5		$\begin{pmatrix} 3 & 4 \\ 12 & 13 \end{pmatrix}_C$		
6				
7				

Рисунок 2.1– Приклад множення матриць

Формулу потрібно вводити як формулу масиву. Якщо формула не буде введена як формула масиву, буде видано єдине значення. Щоб запровадити формулу як формулу масиву треба:

1. Ввести формулу в комірку.

A5 := МУМНОЖ(A2:B3; D2:E3).

2. Виділити весь діапазон, де буде масив-результат.

3. Натиснути F2 і потім клавіші Ctrl + Shift + Enter.

4. У результаті в кожній комірці формула буде у фігурних дужках {} і в комірках буде відповідний результат.

5. «МОБР» – визначення зворотної матриці. Параметр: (масив).

«ТРАНСП» – транспонування матриці. Параметр: (масив).

Зворотна матриця може не існувати, якщо визначник дорівнює нулю (знаходження визначника матриці - функція МОПРЕД (масив)).

Крім того, функція "ЛИНЕЙН" (категорія "Статистичні") повертає значення оцінок параметрів регресії.

«ЛИНЕЙН» (відомі_значення_y; відомі_значення_x; [конст]; [статистика]);

[конст] – логічне значення, яке вказує чи потрібно, щоб вільний член дорівнював нулю.

[статистика] — логічне значення, яке вказує – чи потрібно повернути додаткову статистику щодо регресії.

Додаткова регресійна статистика:

1) $S_{\beta_1}, S_{\beta_2}, \dots, S_{\beta_m}$ – стандартні значення помилок для коефіцієнтів;

2) S_{β_0} – стандартне значення помилки β_0 ;

3) R^2 – коефіцієнт детермінації;

4) S_Y – стандартна помилка для оцінювання Y ;

5) F – статистика;

6) df – ступінь свободи ($\nu_1 = n - k - 1, \nu_2 = k$);

7) SS_{rescel} – залишкова сума квадратів.

Приклад часткового виконання завдання до лабораторної роботи.

Нехай проведено $n = 8$ експериментів і вхідні дані занесено в таблицю (рис. 2.2).

Y	120	140	129	145	117	139	151	110
X1	200	210	201	225	190	199	234	190
X2	89	94	91	95	88	92	95	85

Рисунок 2.2 – Вхідні дані для побудови рівняння
лінійної множинної регресії

На рис. 2.3-2.4 наведено основні проміжні таблиці обчислень.

X^T	1	1	1	1	1	1	1	1
	200	210	201	225	190	199	234	190
	89	94	91	95	88	92	95	85

$X^T * X$	8	1649	729
	1649	341683	150614
	729	150614	66521

Рисунок 2.3 – Проміжні таблиці

$(X^T * X)^{-1}$	138,066	0,324	-2,247					
	0,324	0,002	-0,009					
	-2,247	-0,009	0,044					
$(X^T * X)^{-1} * X^T$	2,915	-5,079	-1,255	-2,465	1,921	-4,149	0,451	8,661
	0,005	-0,016	-0,010	0,009	-0,009	-0,024	0,029	0,017
	-0,041	0,094	0,039	0,008	0,001	0,100	-0,070	-0,131

Рисунок 2.4 – Проміжні таблиці

Результат обчислення вибірових коефіцієнтів регресії показано на рис. 2.5.

$\hat{\beta}_0$	-211,72
$\hat{\beta}_1$	0,12888 3
$\hat{\beta}_2$	3,47357 3

Рисунок 2.5 – Оцінки коефіцієнтів множинної лінійної регресії

Таким чином, отримане рівняння має вигляд:

$$\hat{Y} = -211,72 + 0,129X_1 + 3,474X_2.$$

Серйозною проблемою при побудові моделей множинної лінійної регресії є мультиколінеарність – лінійний взаємозв'язок двох або декількох факторних змінних.

Існує кілька ознак, за якими може бути встановлена наявність мультиколінеарності.

1) Вибірковий коефіцієнт детермінації досить високий та статистично значущий (критерій Фішера), але оцінки коефіцієнтів регресії статистично незначущі, тобто вони мають низькі t -статистики.

2) Парна кореляція між малозначущими пояснювальними змінними досить висока. Однак, ця ознака можна враховувати лише у разі двох пояснюючих змінних. При великій кількості більш доцільним є використання часткових коефіцієнтів кореляції.

3) Високі часткові коефіцієнти кореляції (0,7 та вище):

$$r_{X_1 X_2}^* = \frac{K_{X_1 X_2}^*}{\sigma_{X_1}^* \sigma_{X_2}^*}.$$

Часткові коефіцієнти кореляції визначають силу лінійної залежності між двома змінними без урахування впливу інших змінних. Однак, при вивченні багатовимірних зв'язків у ряді випадків парні коефіцієнти кореляції можуть давати зовсім неправильні уявлення про характер зв'язків між двома змінними.

Наприклад, між змінними X і Y може бути високий позитивний коефіцієнт кореляції не тому, що одна з них стимулює зміну іншої, а через те, що обидві ці змінні змінюються в одному напрямку під впливом інших змінних, як врахованих в моделі, так і не врахованих.

Коефіцієнт кореляції між двома змінними, очищений від інших змінних, називається частковим коефіцієнтом кореляції.

4) Сильна допоміжна регресія.

Мультиколінеарність може мати місце через те, що будь-яка з пояснюючих змінних є лінійною комбінацією інших пояснюючих змінних. Для аналізу будуються рівняння регресії кожної з пояснюючих

змінних X_j на пояснюючі змінні, що залишилися. Обчислюються відповідні коефіцієнти детермінації та розраховується їхня статистична значимість на підставі F -статистики:

$$F_j = \frac{R_j^2}{1 - R_j^2} \cdot \frac{n - m}{m - 1}.$$

Якщо коефіцієнт R_j^2 статистично незначущий, то x_j не є лінійною комбінацією інших змінних, і x_j можна залишити в рівнянні.

Методи усунення мультиколінеарності.

1. Виключення змінної з моделі

Найпростішим методом усунення мультиколінеарності є виключення з моделі однієї або низки корельованих змінних. Однак у цій ситуації можливі помилки специфікації. Наприклад, при дослідженні попиту на деякий продукт у якості пояснюючих змінних можна використовувати ціну даного продукту та ціну замінників, які часто корелюють один з одним. Виключивши з моделі ціни замінників, швидше за все припустимося помилки специфікації, тому можна зробити необґрунтовані висновки.

2. Зміна обсягу вибірових даних або отримання додаткових даних чи нової вибірки

Іноді для зменшення мультиколінеарності достатньо збільшити обсяг вибірки. Наприклад, від використання щорічних даних можна перейти до поквартальних даних. Збільшення кількості даних скорочує дисперсії коефіцієнтів у регресії і цим збільшує їх статистичну значимість. Однак такий підхід може збільшити автокореляцію, що обмежує можливості використання цього методу.

3. Зміна специфікації моделей

У ряді випадків проблема мультиколінеарності може бути вирішена шляхом зміни специфікації моделей: або змінюється форма моделей; або додаються пояснювальні змінні, які не враховані в початковій

моделі, але існуючі та впливають на залежні змінні. Якщо цей метод має підстави, його використання зменшує суму квадратичних відхилень, цим скорочує стандартну помилку регресій.

2.4 Зміст звіту

Звіт до лабораторного заняття повинен бути оформлений відповідно до ДСТУ-3008:15 і містити:

1. Вхідні дані.
2. Запис вхідних даних у вигляді матриць та векторів.
3. Запис формул для обчислення коефіцієнтів регресії.
4. Операції із матрицями.
5. Результати пошуку коефіцієнтів регресії за допомогою матричної алгебри та за допомогою функції Excel.
6. Формули для обчислення та результати обчислень стандартної помилки регресії та стандартних помилок коефіцієнтів регресії.
7. Результати перевірки наявності мультиколінеарності даних.
8. Опис методу варіанта усунення мультиколінеарності даних, отримане нове рівняння регресії.
9. Перевірка нового рівняння регресії на значущість коефіцієнтів (критерій Стюдента), розрахунок скоригованого коефіцієнта детермінації, перевірка значущості коефіцієнта детермінації (критерій Фішера), перевірка гіпотези про випадковість відхилень).

2.5 Контрольні питання

1. Як визначається модель множинної лінійної регресії?
2. Запишіть функціонал, який мінімізують для знаходження коефіцієнтів рівняння множинної лінійної регресії.
3. Опишіть алгоритм визначення коефіцієнтів множинної лінійної регресії за методом найменших квадратів в матричній формі.

4. Що характеризує коефіцієнт еластичності множинної регресії?

5. Наведіть формули розрахунку стандартних помилок коефіцієнтів регресії.

6. Дати визначення мультиколінеарності. Перелічити ознаки, що свідчать про мультиколінеарність даних та методи усунення мультиколінеарності даних?

7. Дати визначення та навести формулу для розрахунку часткового коефіцієнта кореляції множинної регресії.

Література [1-3]

3 Лабораторне заняття

Розробка алгоритмічної та програмної реалізації ГПВЧ

Мета заняття: формування уявлень та практичних навичок у розробці та дослідженні ефективності генераторів псевдовипадкових чисел, створення програмних імітаторів ГПВЧ.

3.1 Завдання до лабораторного заняття

1. Розробка алгоритму та програми-імітатора генератора псевдовипадкових чисел, рівномірно розподілених на відрізку $[0,1]$, за заданим варіантом. Порівняння якості розробленого ГПВЧ з вбудованим в обрану мову програмування за критерієм Пірсона, критерієм серій, частотним тестом та точковими оцінками параметрів розподілу.

2. Розробка алгоритму та програми-імітатора генератора псевдовипадкових чисел, рівномірно розподілених на відрізку $[a, b]$.

3. Розробка алгоритму та програми-імітатора генератора псевдовипадкових чисел, нормально-розподілених з параметрами $N(m, D)$.

4. Перевірка ефективності всіх побудованих генераторів за критерієм Пірсона, критерієм серій, частотним тестом та точковими оцінками параметрів розподілу. Аналіз отриманих результатів.

3.2 Варіанти індивідуальних завдань

1. В першому пункті завдання обрати метод ГПВП з таблиці 3.1 за формулою: $(N \bmod 9) + 1$, де N – номер варіанту за списком групи. Написати програму-імітатор, мову програмування вибрати самостійно.

Таблиця 3.1– Методи ГПВЧ

N	а. Метод
1	б. Серединних квадратів
2	с. Серединних добутків
3	д. Перемішування
4	е. Лінійний конгруентний генератор
5	ф. Адитивний конгруентний генератор
6	Інверсний конгруентний генератор
7	Регістр зсуву зі зворотним лінійним зв'язком. Метод М-послідовності
8	г. Комбінований метод Таусворта
9	h. П'яти параметричний метод

2. Напишіть комплексну програму моделювання вибірки випадкових чисел, параметри: математичне очікування і дисперсія для всіх нижче перерахованих розподілів дорівнюють $m = N$; $\sigma = N \bmod 3 + 1$:

- а) рівномірний на інтервалі $R(a, b)$;
- б) нормальний закон $N(m, \sigma)$.

3. Задавши параметри розподілів відповідно до варіанта, промодельуйте вибірки всіх вищевказаних розподілів та для ГПВЧ, що вбудований у використану мову програмування. Обсяг кожної вибірки прийняти $N = 1000$, кількість вибірок 10.

4. Обчисліть оцінки математичного очікування і дисперсії кожної з отриманих в п. 2-3 послідовностей випадкових чисел для наступних

обсягів вибірки $N1 = 100$, $N2 = 200$, $N3 = 500$, $N4 = 1000$ і $N5 = 10000$.

5. Порівняйте отримані вибіркові оцінки із теоретичними параметрами. Побудуйте графіки залежностей оцінок параметрів від номеру експерименту та обсягу вибірки. Оцінити відносні похибки для будь-якої однієї вибірки.

6. Для всіх вибірок різних розподілів, отриманих за пунктом 3, розрахуйте і побудуйте гістограми розподілів. Порівняйте гістограми з графіками теоретичних розподілів.

7. Виконайте перевірку ефективності ГПВЧ за частотним критерієм

3.3 Вказівки до виконання завдань

Виконання лабораторної роботи 3 потребує розробки програмного додатку генерації псевдо-випадкових чисел на довільній мові програмування, наприклад Python.

Наприклад, один з методів генерації ГПВЧ – метод перемішування. Метод працює з комірками пам'яті, де зберігаються числа, і базується на ітеративному перетворенні початкового значення. Спочатку у визначену комірку записується початкове число R_0 . Далі виконується циклічний зсув вмісту комірки: вліво на одну четверту довжини комірки, що формує число R_0^* , та вправо на одну четверту довжини комірки, що дає число R_0^{**} . Нове випадкове число R_1 отримують шляхом додавання результатів цих зсувів. У разі, якщо сума перевищує розмір комірки, зайві розряди відкидаються, щоб значення залишалось в межах допустимого діапазону. Після цього отримане число R_1 замінює початкове значення R_0 і стає новим входним значенням для наступної ітерації. Цей процес повторюється, створюючи видимо випадкову послідовність чисел. Алгоритм забезпечує псевдовипадковість завдяки

циклічним зсувам і додаванню, які ефективно перемішують початкове значення.

При спробі реалізації даного алгоритму було отримано дані, дуже схожі між собою. Для поліпшення результату алгоритм було декілька змінено. Оновлений алгоритм для генерації псевдовипадкових чисел використовує покращений підхід, що включає циклічні зсуви, побітові операції та контроль повторень для забезпечення більш якісного розподілу.

Початковим етапом є збереження стартового числа, яке виступає як початковий стан системи. У процесі генерації виконуються циклічні зсуви цього числа: вліво на третину довжини комірки та вправо на п'яту частину довжини комірки. Результати цих зсувів комбінуються шляхом побітового складання, що формує базове нове значення.

Додатково це значення обробляється через серію побітових перетворень, які включають зсуви та додаткові побітові операції. Такі маніпуляції з числом збільшують його складність та ентропію, створюючи більш випадковий результат. Щоб значення залишалось в межах розміру комірки, до нього застосовується обмеження через маскування, яке відкидає зайві біти.

Кінцевий результат нормалізується шляхом приведення отриманого числа до діапазону від 0 до 1. Завдяки цьому алгоритм забезпечує генерацію псевдовипадкових чисел із високою якістю розподілу, уникаючи закономірностей навіть при багаторазовому використанні.

Далі було зроблено порівняння розробленого генератору на основі методу перемішування та вбудованої функції генерації псевдовипадкових чисел в мові Python. Створено по дві вибірки розміром $n = 1000$ рівномірного та нормального розподілів. Для їх порівняння було визначено точкові оцінки вибірок. Їх наведено у

таблиці 3.2. За результатами порівняння визначено, що отримані вибірки згенеровані різними методами мають незначні розбіжності між собою та відповідають оцінкам відповідних параметрів рівномірного та нормального розподілів.

Таблиця 3.2 – Порівняння точкових оцінок

	Власний генератор		Вбудований	
	X	Y	X	Y
m^*	6.0235	5.9180	5.9304	6.0642
σ^*	1.9707	2.0076	2.0217	1.9939
As^*	-0.0467	0.0607	0.0388	-0.0233
E_x^*	-1.1524	0.1134	-1.2291	0.0751

Далі побудуються гістограми для кожної вибірки та відображаються послідовності значень вибірок у графічному вигляді (рис. 3.1 – 3.2) та результати генерації від обсягу вибірок (рис. 3.3-3.4).

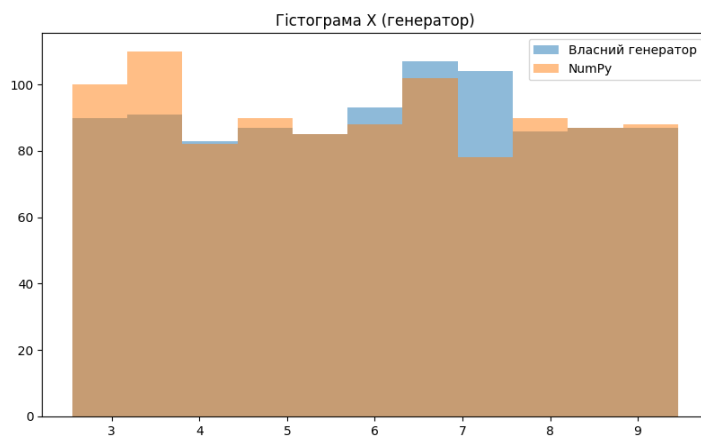


Рисунок 3.1 – Гістограми вибірок X для $n = 1000$

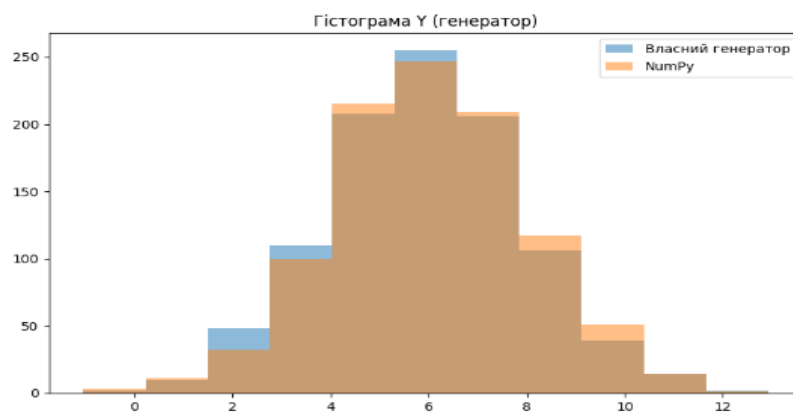


Рисунок 3.2 – Гістограми вибірок Y для $n = 1000$

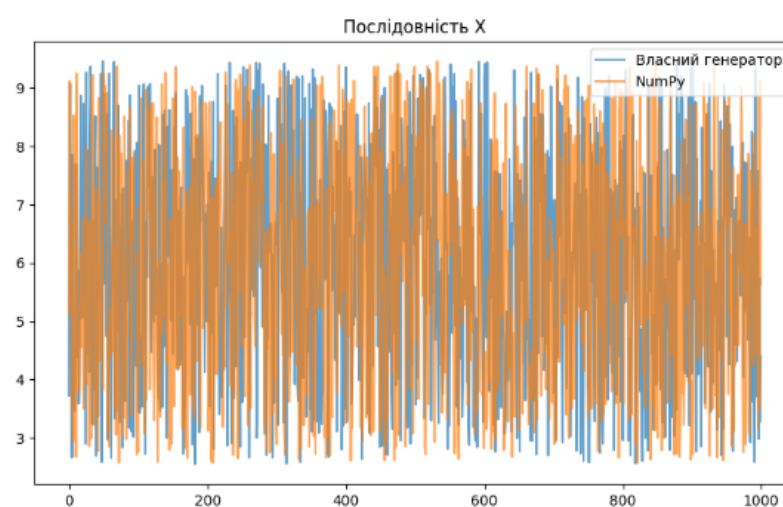


Рисунок 3.3 – Послідовності значень у вибірках X

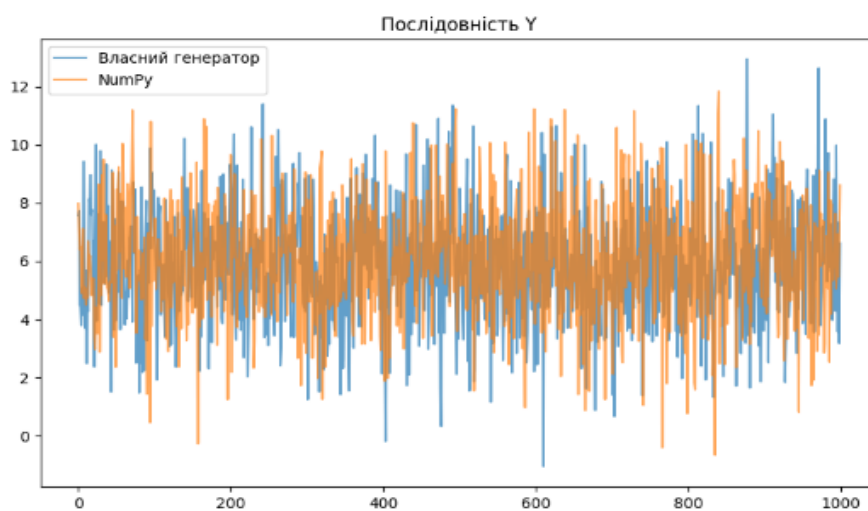


Рисунок 3.4 – Послідовності значень у вибірках Y

За отриманими вибірками було зроблено припущення про закони розподілу відповідних генеральних сукупностей: X – рівномірний та Y – нормальний. Зроблено перевірку теоретичних законів розподілу для вибірок згенерованих як з власним генератором, так і для вбудованої функції генерації випадкових чисел мови Python. Застосовано критерій згоди Пірсона.

Результати розрахунків для критерію Пірсона по кожній з вибірових сукупностей наведено в таблицях 3.2 – 3.5.

Таблиця 3.2 – Критерій Пірсона для X за власним генератором

Інтервал		m_i	p_i	np_i	$\frac{(np_i - m_i)^2}{np_i}$
2,55	3,24	103	0,1	100	0,09
3,24	3,93	91	0,1	100	0,81
3,93	4,62	96	0,1	100	0,16
4,62	5,31	96	0,1	100	0,16
5,31	6,00	93	0,1	100	0,49
6,00	6,69	120	0,1	100	4,00
6,69	7,39	109	0,1	100	0,81
7,39	8,08	105	0,1	100	0,25
8,08	8,77	94	0,1	100	0,36
8,77	9,46	93	0,1	100	0,49
Сума		1000	1	1000	7,62
χ^2		7,620		$\chi^2_{\text{крит}}$	16,9190
Розподіл рівномірний				H_0	приймається

Таблиця 3.3 – Критерій Пірсона для Y за власним генератором

Інтервал		m_i	p_i	np_i	$\frac{(np_i - m_i)^2}{np_i}$
-1,05	0,35	3	0,0025	2,52	0,0929
0,35	1,75	18	0,0161	16,14	0,2142
1,75,	3,15	51	0,0648	64,84	2,9524
3,15	4,54	171	0,1633	163,29	0,3645

4,54	5,94	266	0,258	258,03	0,2461
5,94	7,34	253	0,256	255,96	0,0341
7,34	8,74	151	0,1594	159,38	0,4402
8,74	10,14	65	0,0623	62,27	0,1199
10,14	11,53	20	0,0153	15,25	1,4782
11,53	12,93	2	0,0023	2,34	0,0493
Сума		1000	1	1000	5,9919
χ^2		5,9919		$\chi^2_{\text{крит}}$	14,0671
Розподіл нормальний				H_0	приймається

Таблиця 3.4 – Критерій Пірсона для X за вбудованим генератором

Інтервал		m_i	p_i	np_i	$\frac{(np_i - m_i)^2}{np_i}$
2,57	3,26	114	0,1	100	1,96
3,26	3,95	112	0,1	100	1,44
3,95	4,63	95	0,1	100	0,25
4,63	5,32	102	0,1	100	0,04
5,32	6,01	81	0,1	100	3,61
6,01	6,7	111	0,1	100	1,21
6,7	7,39	98	0,1	100	0,04
7,39	8,08	88	0,1	100	1,44
8,08	8,77	100	0,1	100	0,00
8,77	9,46	99	0,1	100	0,01
Сума		1000	1	1000	10
χ^2		10		$\chi^2_{\text{крит}}$	16,9190
Розподіл рівномірний H_0					приймається

Таблиця 3.5 – Критерій Пірсона для Y за вбудованим генератором

Інтервал		m_i	p_i	np_i	$\frac{(np_i - m_i)^2}{np_i}$
-0,66	0,59	4	0,0027	2,67	0,6662
0,59	1,84	17	0,0141	14,11	0,5931
1,84	3,09	40	0,051	50,98	2,3644
3,09	4,34	120	0,1259	125,92	0,2786

4,34	5,59	231	0,2127	212,7	1,574
5,59	6,84	245	0,2458	245,76	0,0023
6,84	8,09	189	0,1942	194,24	0,1413
8,09	9,34	99	0,105	105	0,3434
9,34	10,58	43	0,0388	38,81	0,4512
10,58	11,83	12	0,0098	9,81	0,4907
Сума		1000	1	1000	6,9052
χ^2		6,9052		$\chi^2_{\text{крит}}$	14,0671
Розподіл нормальний				H_0	приймається

За результатами аналізу всі гіпотези про закон розподілу було прийнято.

Для перевірки на рівномірність розподілу ГПВЧ повинен видавати близькі до наступних значення статистичних параметрів, характерних для рівномірного випадкового закону:

- 1) $m_r \approx 0.5$ – математичне очікування;
- 2) $D_r \approx 1/12$ – дисперсія;
- 3) $\sigma_r \approx 0.2887$ – середньоквадратичне відхилення.

Частотний тест дозволяє з'ясувати, скільки чисел потрапило в інтервал $(m_r - \sigma_r; m_r + \sigma_r)$ тобто $(0.5 - 0.2887; 0.5 + 0.2887)$ або, в кінцевому підсумку, $(0.2113; 0.7887)$.

3.4 Зміст звіту

Звіт до лабораторного заняття повинен бути оформлений відповідно до ДСТУ-3008:15 і містити:

- титульну сторінку;
- опис алгоритмів ГПВЧ;
- результати аналізу якості кожного з побудованих генераторів: критерії серій та Пірсона, частотний критерій, аналіз отриманих оцінок параметрів розподілу (математичне очікування, дисперсія, асиметрія та ексцес);

- результати порівняння розробленого ГПВЧ з вбудованим в обрану мову програмування;
- висновки.

3.5 Контрольні питання

1. Метод Монте-Карло, основні засади.
2. Випадкові та псевдовипадкові числа. Типи генераторів псевдовипадкових чисел (ГПВЧ).
3. Загальна характеристика алгоритмічних ГПВЧ.
4. Метод серединних квадратів.
5. Метод серединних добутків.
6. Метод перемішування.
7. Лінійний конгруентний метод (ЛКМ).
8. Мультиплікативний конгруентний метод.
9. Змішаний конгруентний метод.
10. Методи генерації нормально розподілених випадкових величин.
11. Перевірка якості роботи ГПВЧ: перевірки на рівномірність розподілу (частотний тест, критерій Пірсона, оцінки параметрів розподілу).
12. Перевірка на випадковість та статистичну незалежність, критерій серій.

Література: [4-5].

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Методичні вказівки і завдання до лабораторних робіт за курсом "Статистичні методи оцінювання" для магістрів, що навчаються за спеціальністю 121 Інженерія програмного забезпечення денної форми навчання [Електронний ресурс] / уклад. І. А. Назарова. - Луцьк: ДонНТУ, 2022. - 45 с. - Режим доступу : <http://lc.donntu.edu.ua/elcat/download3/21788>. - Назва з екрана.
2. Бідюк П. І. Прикладна статистика: навч. посіб. / П. І. Бідюк, В. Я. Данилов, О. Л. Жиров. - Київ : НТУУ «КПІ», 2023. - 186 с.
3. Диха М. В. Економетрія: навч. посіб. / М. В. Диха, В. С. Мороз. - Київ : Центр навчальної літератури, 2021. - 202 с.
4. Ковальова І. Л. Економетрія: навч. посіб. / І. Л. Ковальова. - Одеса : ОДАБА, 2019. - 423 с.
5. Генератори випадкових чисел [Електронний ресурс]. - Режим доступу : http://baklaniv.at.ua/MC/lekcija_22.pdf. - Назва з екрана.
6. Методичні вказівки і завдання до практичних робіт за курсом "Аналіз випадкових процесів" для студентів освітнього ступеню «Бакалавр» спеціальності 125 Кібербезпека денної форми навчання [Електронний ресурс] / уклад. І.А. Назарова. - Покровськ : ДВНЗ «ДонНТУ», 2021. - 79 с. - Режим доступу : <http://lc.donntu.edu.ua/elcat/download3/22360>. - Назва з екрана.