

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ»
КАФЕДРА ПРИКЛАДНОЇ МАТЕМАТИКИ І ІНФОРМАТИКИ**



**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**

Покровськ-2022

УДК 519.233

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 / Укл.: І.А. Назарова. - Покровськ: ДонНТУ, 2022. - 44с.

Методичні вказівки містять коротку довідку з теоретичного матеріалу, завдання та індивідуальні варіанти до виконання лабораторних робіт з курсу "Статистичні методи оцінювання".

Укладач: Назарова І.А., к.т.н., доцент кафедри прикладної математики і інформатики

Рецензент Ковальов С.О., к.т.н., доцент кафедри комп'ютерної інженерії

Відповідальний за випуск: О.А. Дмитрієва, професор, д.т.н., завідувач кафедрою "Прикладна математика і інформатика"

Розглянуто на засіданні кафедри прикладної математики і інформатики,
протокол №5 від 23.05.22р.

Затверджено навчально-методичним відділом ДВНЗ «ДонНТУ»,
протокол № 10 від 31.05.2022р.

ДВНЗ «ДонНТУ», 2022р.

ЗМІСТ

Вступ.....	4
Лабораторна робота №1 Кореляційний аналіз даних.....	6
1.1 Теоретична довідка.....	6
1.2 Завдання до лабораторної роботи.....	11
1.3 Варіанти завдань до лабораторної роботи	11
1.4 Питання для самоконтролю.....	18
Лабораторна робота № 2 Парна лінійна модель регресії.....	19
2.1 Теоретична довідка.....	19
2.2 Завдання до лабораторної роботи.....	22
2.3 Варіанти завдань до лабораторної роботи	22
2.4 Питання для самоконтролю.....	23
Лабораторна робота № 3 Рангова кореляція. Непараметричні методи аналізу взаємозв'язку.....	24
3.1 Теоретична довідка.....	24
3.2 Завдання до лабораторної роботи.....	28
3.3 Варіанти завдань до лабораторної роботи.....	28
3.4 Питання для самоконтролю.....	28
Лабораторна робота № 4 Дисперсійний аналіз даних	29
4.1 Теоретична довідка.....	29
4.2 Завдання до лабораторної роботи	32
4.3 Варіанти завдань до лабораторної роботи.....	33
4.4 Питання для самоконтролю.....	43
Список використаної літератури.....	44

ВСТУП

Методичні вказівки і завдання складено до курсу “Статистичні методи оцінювання” навчального плану магістрів та призначено для студентів, що навчаються за спеціальністю 121 Інженерія програмного забезпечення галузі знань 12 Інформаційні технології денної форми навчання.

При дослідженні динамічних процесів довільної природи в багатьох випадках важливо визначити чи є взаємозв'язок між параметрами, що описують цей процес. В більшості випадків між величинами існують такі залежності, коли кожному значенню однієї змінної відповідає множина можливих значень іншої змінної. Така залежність отримала назву ймовірнісної, статистичної або стохастичної залежності. Для вивчення та опису такого зв'язку розроблено методи, що отримали назву кореляційно-регресійний аналіз. Основним завданням кореляційного аналізу є встановлення зв'язку між випадковими змінними та оцінка його тісноти. Як і при перевірці статистичних гіпотез, кореляційні методи можуть бути параметричними та непараметричними, так звана рангова кореляція. На відміну від цього регресійний аналіз встановлює форму зв'язку, дає його аналітичний вираз та оцінює якість отриманої моделі. Дисперсійний аналіз є одним з основних методів статистичного оцінювання результатів експерименту.

Дисперсійний аналіз дає змогу встановити, наскільки вибіркові показники зв'язку результативного і факторних ознак достатні для поширення одержаних за вибіркою даних на генеральну сукупність. Достойнством цього методу є те, що він дає досить надійні висновки за вибірками невеликого обсягу.

Отже, методичні вказівки та завдання до курсу «Статистичні методи оцінювання» містять три розділи, що присвячені основам кореляційного, регресійного та дисперсійного аналізу.

Методичні вказівки побудовано таким чином, що студент повторно ознайомлюється з теоретичними відомостями відповідно до зазначеної теми, може проглянути приклад використання вказаного методу для розв'язання поставленої задачі та проаналізувати одержані результати. Для оцінки засвоєння теорії та ступеню оволодіння практичними навичками при вивченні даної дисципліни студенти можуть скористуватись питаннями для самоконтролю, що забезпечить повторення та закріплення пройденого матеріалу.

ЛАБОРАТОРНА РОБОТА №1

КОРЕЛЯЦІЙНИЙ АНАЛІЗ ДАНИХ

Ціль роботи: засвоєння теоретичного матеріалу та набуття практичних навичок у застосуванні засад кореляційного параметричного двовимірного аналізу.

1.1 Теоретична довідка

Кореляційний аналіз – метод, що дозволяє досліджувати залежність між декількома випадковими величинами (ВВ).

Метою кореляційного аналізу є виявлення оцінки сили зв'язку між випадковими величинами (ознаками), які характеризують певний реальний процес або об'єкт.

Завдання кореляційного аналізу:

- вимірювання ступеня зв'язності (тісноти, сили, строгості, інтенсивності) двох і більше явищ;
- відбір факторів, що мають найбільш істотний вплив на результативну ознаку, на підставі вимірювання ступеня зв'язності між явищами.

Існують різні види зв'язку між змінними:

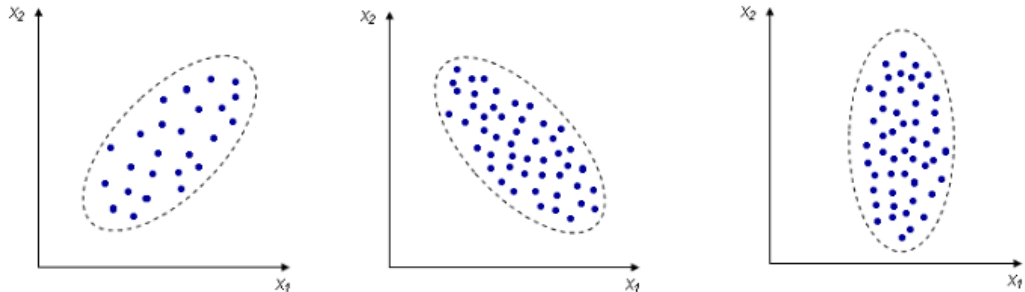
1. Прямий причинно-наслідковий зв'язок (рис. 1.1, а).
2. Зворотний причинно-наслідковий зв'язок (рис. 1.1, б).
3. Зв'язку немає, залежність, що спостерігається випадкова (рис. 1.1, в).

Коваріація і коефіцієнт кореляції

Перші початкові і другі центральні моменти неповністю характеризують систему ВВ, вони не висловлюють ступінь залежності її складових.

Особливу роль, як характеристика системи, має другий змішаний центральний момент:

$$\mu_{11} = M[\dot{X} \cdot \dot{Y}] = M[(X - m_X) \cdot (Y - m_Y)].$$



а) прямий зв'язок б) зворотний зв'язок в) відсутність зв'язку

Рис. 1.1. Варіанти взаємозв'язку між випадковими змінними

Коваріація (кореляційний момент) K_{xy} випадкових величин X і Y – математичне очікування добутку відхилень цих величин від своїх математичних очікувань: $K_{xy} = \text{cov}(X, Y)$, $K_{xy} = K_{yx}$.

Для ДВВ:
$$K_{xy} = \sum_i \sum_j (x_i - m_x) \cdot (y_j - m_y) \cdot p_{ij}.$$

Для БВВ:
$$K_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x) \cdot (y - m_y) \cdot f(x, y) dx dy.$$

Зміст кореляційного моменту: K_{xy} описує зв'язок між X і Y та розбіг навколо точки (m_x, m_y) .

Властивості коваріації

- 1) Коваріація двох незалежних ВВ дорівнює 0.
- 2) Якщо кореляційний момент двох ВВ відрізняється від 0, це ознака наявності залежності між ними.

3) Коваріація двох ВВ дорівнює математичному очікуванню їх добутку мінус добуток математичних очікувань кожної з величин:

$$K_{xy} = M[X \cdot Y] - m_x \cdot m_y.$$

4) Коваріація двох ВВ за абсолютною величиною не перевищує добуток їх середніх квадратичних відхилень:

$$|K_{xy}| \leq \sigma_x \cdot \sigma_y.$$

Коваріація величина розмірна, її розмір визначається добутком розмірів ВВ X і Y .

Коефіцієнт кореляції двох ВВ – відношення їх коваріації до добутку середніх квадратичних відхилень цих величин:

$$r_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y}, \quad r_{xy} = r_{yx}.$$

Властивості коефіцієнта кореляції

- 1) Коефіцієнт кореляції приймає значення на відрізку $[-1; 1]$.
- 2) Якщо ВВ незалежні, то їх коефіцієнт кореляції дорівнює 0. ВВ – некорельовані, якщо їх коефіцієнт кореляції дорівнює 0.
- 3) Якщо ВВ незалежні отже вони некорельовані. Зворотне, в загальному випадку, невірно. З не корельованості ВВ ще не слідує їх незалежність.
- 4) Якщо коефіцієнт кореляції дорівнює $r_{xy} = 1$, то між цими ВВ існує лінійна функціональна пряма залежність.

Якщо $r_{xy} = -1$, то між цими ВВ існує лінійна функціональна зворотна залежність.

5) В загальному випадку, коли величини X і Y пов'язані довільною ймовірнісною залежністю, то:

- якщо $r_{xy} > 0$ – позитивна кореляція (прямий зв'язок);
- якщо $r_{xy} < 0$ – негативна кореляція (зворотний зв'язок).

Величина коефіцієнта кореляції та рівень зв'язку наведені у табл. 1.1.

Таблиця 1.1 Лінійний коефіцієнт кореляції

Значення r_{xy}	Рівень зв'язку між змінними
0,75 – 1,00	дуже високий позитивний
0,50 – 0,74	високий позитивний
0,25 – 0,49	середній позитивний
0,00 – 0,24	слабкий позитивний
0,00 – -0,24	слабкий негативний
-0,25 – -0,49	середній негативний
-0,50 – -0,74	високий негативний
-0,75 – -1,00	дуже високий негативний

Оцінки числових характеристик системи ДВВ

Нехай є результати незалежних дослідів над системою 2 випадкових величин (X, Y) : $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$.

Потрібно знайти оцінки для числових характеристик системи: $m_x^*, m_y^*, D_x^*, D_y^*, K_{xy}^*, r_{xy}^*$ або $\tilde{m}_x, \tilde{m}_y, \tilde{D}_x, \tilde{D}_y, \tilde{K}_{xy}, \tilde{r}_{xy}$.

Незміщена оцінка математичного очікування:

$$\tilde{m}_x = \frac{\sum_{i=1}^n x_i}{n}; \quad \tilde{m}_y = \frac{\sum_{i=1}^n y_i}{n}.$$

Оцінки елементів кореляційної матриці:

$$\tilde{D}_x = \frac{\sum_{i=1}^n (x_i - \tilde{m}_x)^2}{n-1}; \quad \tilde{D}_y = \frac{\sum_{i=1}^n (y_i - \tilde{m}_y)^2}{n-1};$$

$$\tilde{K}_{xy} = \frac{\sum_{i=1}^n (x_i - \tilde{m}_x) \cdot (y_i - \tilde{m}_y)}{n-1}.$$

Для обчислень зручніше користуватися такими формулами:

$$D_x^* = \alpha_2^*[X] - (m_x^*)^2; \quad D_y^* = \alpha_2^*[Y] - (m_y^*)^2;$$

$$K_{xy}^* = a_{1,1}^*[X, Y] - m_x^* m_y^*,$$

$$\alpha_2^*[X] = \frac{\sum_{i=1}^n x_i^2}{n}; \quad \alpha_2^*[Y] = \frac{\sum_{i=1}^n y_i^2}{n}; \quad \alpha_{1,1}^*[X, Y] = \frac{\sum_{i=1}^n x_i \cdot y_i}{n}.$$

Незміщені оцінки:

$$\tilde{D}_x = D_x^* \cdot \frac{n}{n-1}; \quad \tilde{D}_y = D_y^* \cdot \frac{n}{n-1};$$

$$\tilde{K}_{xy} = K_{xy}^* \cdot \frac{n}{n-1}; \quad \tilde{r}_{xy} = \frac{\tilde{K}_{xy}}{\tilde{\sigma}_x \cdot \tilde{\sigma}_y}.$$

Перевірка гіпотези про значущість вибіркового коефіцієнта кореляції

Нехай обчислене вибіркове значення коефіцієнта кореляції, значення r_{xy}^* або $\tilde{r}_{xy}$ відмінно від нуля.

Виникає питання, пояснюється це існуванням лінійного кореляційного зв'язку між величинами X і Y або є наслідком випадковості відбору змінних в вибірку.

1) Нульова гіпотеза $H_0 : r_{xy} = 0$.

2) Альтернативна $H_1 : r_{xy} \neq 0$.

3) Статистика критерію:
$$t^* = \frac{r_{xy}^* \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^{*2}}}$$

При справедливості нульової гіпотези статистика критерію має розподіл Стюдента з $k = n - 2$ ступенями свободи.

Нерівність для критичної області, двосторонній критерій:

$$|t^*| > t_{1-\frac{\alpha}{2}}(n-2).$$

Нульова гіпотеза відкидається, величини є корельованими.

1.2 Завдання до лабораторної роботи

За індивідуальним варіантом вихідних даних провести статистичний аналіз двовимірної вибірки для двох випадкових X та Y величин у наступній послідовності:

а) побудувати графік розсіювання в координатних осях за результатами вимірів для двох випадкових величин;

б) визначити точкові оцінки математичного очікування та дисперсій для кожної із ВВ, що входять до системи;

в) визначити статистичні оцінки коваріації та коефіцієнта кореляції;

г) з'ясувати, чи є випадкові величини корельованими, залежними, оцінити ступінь залежності ВВ, перевірити значущість вибіркового коефіцієнта кореляції.

1.3 Варіанти завдань до лабораторної роботи

№1. Дослідити залежність барометричного тиску від висоти (при однаковій температурі).

Н (м)	100	500	1000	1100	1200	1400	1500	1600	1750	2000
Р (мм рт. ст.)	700	670	640	595	504	363	310	267	200	100

№2 . Дослідити зв'язок між масою тіла (X) та кількістю гемоглобіну в крові (Y).

Маса тіла (кг)	18	17.7	19	18	19	22	21	21	20	30
Гемоглобін	70	74	72	80	77	80	80	89	76	86

№3. Дослідити залежність обсягу продаж (тис. мільйонів доларів) від чистого доходу (мільйони доларів) за 20 фірмами США.

X	8.9	8.4	7.4	7.2	7.0	6.1	5.9	5.8	5.5	4.8
	4.4	4.2	4.2	4.1	3.8	3.8	3.6	3.5	3.3	3.2
Y	441	278	456	934	89	611	770	53	234	217
	454	291	321	51	111	102	356	150	237	151

№4. Дослідити залежність між ціною на товар і прибутком торгового підприємства.

Номер підприємства	1	2	3	4	5	6	7	8	9	10	11	12
Прибуток торгового підприємства, Y, тис. грн.	12	20	19	22	27	29	16	17	34	35	31	27
Ціна товару X, грн.	5	10	7	12	14	18	7	11	45	42	36	28

№5. Дослідити залежність між величинами Y та X. Дана інформація по залежності рівня безробіття від рівня інфляції в Україні за 2000-2007рр.

Рік	2000	2001	2002	2003	2004	2005	2006	2007
Рівень безробіття Y, %	4,3	3,2	3,51	3,9	3,82	3,14	2,51	2,9
Рівень інфляції X, %	25,8	6,1	-0,6	8,2	12,3	10,3	12,5	16,6

№6. Дослідити залежність діяльності підприємства мобільного зв'язку України на основі наступних даних.

Рік	2000	2001	2002	2003	2004	2005	2006	2007
Кількість абонентів, млн. осіб	24,3	23,2	23,51	23,9	23,82	23,14	22,51	22,9

№7. Досліджується якість навчання в університетах США, залежно від витрат. В таблиці **Y** – показник якості університету, **X1** – середня зарплата (тис. дол.), **X2** – число (%) доцентів та професорів (з довідника «Хто є хто в Америці»). Дослідити залежність **Y(X1)**.

Y	31.6	32.6	38.8	34.0	34.8	35.1	16.3	16.7	22.2	28.0	28.6
	29.8	35.6	36.0	36.3	36.5	36.5	36.6	36.6	36.7	38.8	36.9
X1	13.2	13.5	13.4	12.7	14.0	12.4	12.4	14.4	15.4	13.1	15.1
	13.0	11.3	12.3	12.6	10.8	12.3	11.1	10.8	9.8	10.9	10.5
X2	47.1	23.3	14.9	12.7	19.0	10.8	21.0	49.7	21.2	6.5	32.1
	12.0	12.0	10.3	18.8	12.0	23.9	10.6	10.9	10.5	15.5	8.3

№8. За умовами попереднього завдання дослідити залежність **Y(X2)**.

№9. За умовами завдання 7 дослідити залежність **X1(X2)**.

№10. У таблиці: **Y** – число розлучень на 1000 жителів у 20 штатах США, **X1** – середній дохід на родину (тис. дол.), **X2** - % міського населення штату. Дослідити залежність **Y(X1)**.

Y	1.2	1.1	0.4	2.4	2.7	2.1	1.2	1.5	1.9	4.6
	3.6	3.9	4.0	6.7	3.0	2.4	1.2	3.3	3.2	3.1
X1	4.9	6.3	6.4	6.2	5.8	6.2	4.2	4.9	5.0	4.6
	4.9	5.2	5.9	5.8	5.4	5.9	4.9	6.2	5.9	6.7
X2	38.5	83.6	85.4	73.4	62.4	73.4	39.3	54.3	55.8	62.9
	75.0	47.5	56.8	73.7	65.7	74.9	51.3	68.1	62.2	86.4

№11. За умовами попереднього завдання дослідити залежність **Y(X2)**.

№12. За умовами завдання 10 дослідити залежність **X1(X2)**.

№13. Дослідити зв'язок між обсягом виробництва та ціною вишні (дані за 16 років з Agricultural Statistics), **X** – у 1000 т, **Y** – у доларах за 1т.

X	204	200	168	239	192	218	185	266	276	150	344	248
	200	198	228	278								
Y	207	174	228	208	225	243	227	217	163	345	154	165
	299	325	294	188								

№14. Дана вибірка з 20 плавок сталі, що вивчається з погляду вмісту сірки та витрат енергії в металі. Чи насправді енергетичні витрати визначаються % вмістом сірки у металі? **X** – процентний вміст сірки; **Y** – витрати енергії.

X	0.006	0.001	0.012	0.014	0.03	0.022	0.021	0.02	0.028	0.025
	0.016	0.01	0.009	0.017	0.018	0.023	0.015	0.023	0.009	0.011
Y	2.46	2.43	2.43	2.48	2.69	2.48	2.63	2.72	2.72	2.64
	2.68	2.35	2.46	2.54	2.58	2.6	2.42	2.51	2.46	2.37

№15. Чи існує якийсь зв'язок між оцінками 19 студентів з 2-х іспитів, **X1** – сума балів з 1-го іспиту, **X2** – з 2-го.

X1	32	71	35	31	42	101	76	44	102	67	21	48	45	57	72	35	87	50	72
X2	68	57	17	52	17	81	62	43	118	70	40	30	39	80	79	64	77	36	89

№16. У таблицю занесені результати лабораторного аналізу зразків сланцевих порід на вміст двоокису кремнію (**SiO₂**) та двоокису алюмінію (**Al₂O₃**) (в умовних одиницях). Чи є між змістом зв'язок?

Al₂O₃	SiO₂	
	17.2	57.8
19.9	53.3	54.6
17.1	57.9	54.8
15.5	54.0	51.7
17.6	52.6	61.1
16.3	53.8	62.3
17.2	53.6	52.2
15.8	51.5	49.2
15.0	54.0	53.9
14.4	50.4	60.0
15.3	53.0	56.2
16.6	53.3	55.2
14.9	51.6	

№17. За даними 10 міст США: **X** – чисельність населення (дес. тисяч),
Y – кількість телевізійних точок.

	Денвер	Сан-Антоніо	Канзас-Сіті	Сіетл	Цинциннаті	Буффало	Новий Орлеан	Мілуокі	Хьюстон	Нью Йорк
X	45	46	47	48	51	58	59	65	67	802
Y	12	12	29	25	38	35	16	43	22	345

№18. Межа витривалості сталі при вигині **Y**(Н/мм) оцінюється на підставі іншої її характеристики – межі пружності при крученні **X** (Н/мм).
Наведено результати дослідних зразків 12 марок.

X	51	67	84	81	101	109	71	97	109	51	105	89
Y	25	30	43	44	57	58	43	46	62	45	55	45

№19. Дослідити залежності параметрів за групою чоловіків одного віку, **X** – зріст, **Y** – вага, **Z** – розмір взуття. Виявити існування та ступінь залежності показників **Y(X)**.

X – ріст	Y – вага	Z – розмір взуття
165	72.9	40
171	48.4	41
182	66.3	44
165	64.0	40
183	62.7	45
180	76	44
183	72.8	42
166	50.6	42
173	52.3	40
184	68.6	42
168	52.6	41
164	72.8	39
170	61.6	38
174	66.8	41
172	56.5	41

№20. За умовами попереднього завдання дослідити залежність $Z(X)$.

№21. За умовами завдання 19 дослідити залежність $Z(Y)$.

№22. Дослідити залежність роздрібного товарообігу (млн. руб.) магазинів (Y) від середньооблікового числа працівників (X).

X	73	85	102	115	122	126	134	147	150	152	160	168	170	180	200	208	210
Y	0.5	0.7	0.9	1.1	1.4	1.4	1.7	1.9	1.9	1.9	2	2.2	2.2	2.3	2.6	3	3

№23. Дослідити залежність оцінки за тестами для читання та арифметики за групою з 16 першокласників.

Читання	43	58	45	53	37	58	55	61	46	64	46	62	60	65	38	56
Арифметика	32	25	28	30	22	25	22	20	20	30	21	28	34	28	23	24

№24. Дослідити залежність індексу роздрібних цін на харчові товари від року.

Рік	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Індекс цін	100	101	113	115	113	113	111	112	115	120
Індекс виробництва	64	75	81	84	91	85	85	99	100	93

№25. За умовами попереднього завдання дослідити залежність індексу промислового виробництва від року.

№26. У таблиці наведено **Z** – значення врожайності; **X** – вартість внесених добрив; **Y** – сума витрат, пов'язаних із зрошенням. Дані зібрані з полів, що у однакових кліматичних умовах. Дослідити залежність **Z(X)**.

Z	13.3	12.2	12.4	16.4	22.0	17.2	21.8	22.4	24.8	26.7	19.5	30.1	26.7
X	4.2	4.0	5.1	5.0	8.0	6.2	4.8	5.0	6.0	4.9	8.0	8.0	7.0
Y	4.6	5.2	3.8	3.1	5.0	5.1	6.2	7.5	7.2	4.8	3.8	8.0	7.5

№27. За умовами попереднього завдання дослідити залежність **Z(X)**.

X: цукор у крові (в мг%)	100	105	99	97	100	102	98	103	102	95	96	101	104	99
Y: білок	6.8	5.7	8.3	6.7	6.4	7.2	6.0	4.5	5.2	8.8	6.9	5.6	8.4	6.5
Z: Гемо-глобін	12	11.3	14.7	11.4	11.5	12.7	13.0	13.4	16.0	13.6	14.0	13.8	12	11

№28. За умовами попереднього завдання дослідити залежність **Y(X)**.

№29. За умовами завдання 27 дослідити залежність **Z(Y)**.

№30. Дослідити залежність врожайності пшениці на сусідніх полях за наступними роками.

Пшениця	Рік
20.1	1926
23.6	1927
26.3	1928
19.9	1929
16.7	1930
23.2	1931
31.4	1932
33.5	1933
28.2	1934
35.3	1935
29.3	1936
30.5	1937

1.4 Питання для самоконтролю

1. Сутність кореляційного аналізу?
2. Основні завдання кореляційного аналізу?
3. Початковий і центральний моменти.
4. Коваріація і коефіцієнт кореляції.
5. Властивості коваріації.
6. Властивості коефіцієнта кореляції.
7. Вибіркові коваріація і коефіцієнт кореляції.
8. Значущість коефіцієнта кореляції.

ЛІТЕРАТУРА: [1, 4-5,7-8].

ЛАБОРАТОРНА РОБОТА №2

РЕГРЕСІЙНИЙ АНАЛІЗ. ПАРНА ЛІНІЙНА МОДЕЛЬ РЕГРЕСІЇ

Ціль роботи: засвоєння теоретичного матеріалу та набуття практичних навичок у побудові двовимірної парної лінійної регресійної моделі.

2.1 Теоретична довідка

Регресійний аналіз – статистичний метод, що присвячений методам аналізу залежності однієї величини від іншої. На відміну від кореляційного аналізу не з'ясовує чи істотний зв'язок, а займається пошуком моделі цього зв'язку, вираженої у функції регресії:

$$y = f(\bar{x}, \bar{b}),$$

де y – залежна змінна, $\bar{x} = (x_1, x_2, \dots, x_k)$ – в загальному випадку вектор незалежних змінних, $\bar{b} = (b_1, b_2, \dots, b_k)$ – невідомі параметри регресії.

Види регресії

В залежності від числа факторів, які включено у рівняння регресії, розрізняють:

- просту (парну) регресію,
- множинну регресію.

Парна регресія – регресія між двома змінними y та x , $y = f(x, \bar{b})$, де

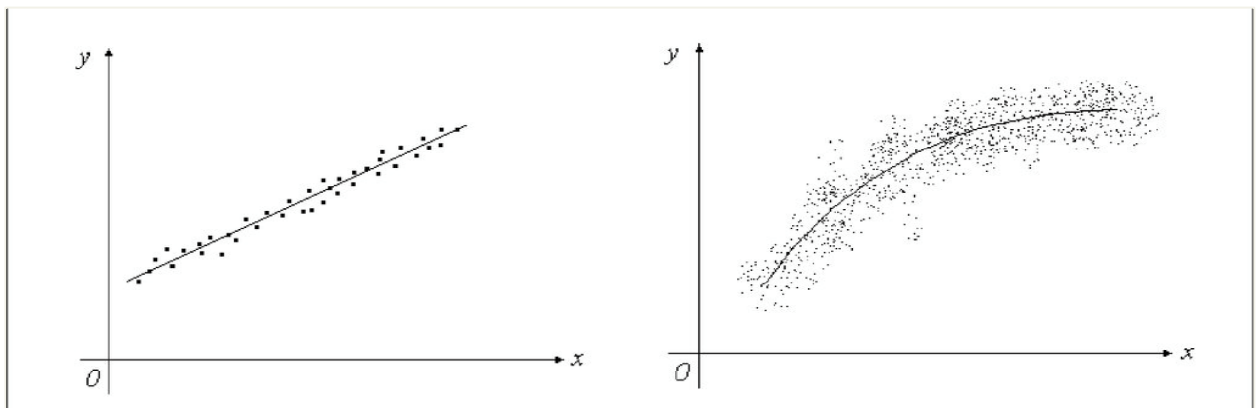
- 1) y – залежна змінна (результативна ознака),
- 2) x – незалежна змінна (факторна ознака).

Парна регресія є достатньою, якщо існує домінуючий фактор, який і використовується як незалежна змінна.

Множинна регресія – це модель виду: $y = f(x_1, x_2, \dots, x_k; \bar{b})$, де y – результативна ознака, $x_1, x_2, \dots, x_k$ – факторні ознаки.

За видом функції f розрізняють лінійні та нелінійні моделі. Найбільш простою формою залежності в математиці є лінійна, тому в регресійному аналізі найбільш популярні лінійні моделі.

У практиці статистичних досліджень найчастіше використовують лінійні залежності або залежності, що зводяться до лінійних, шляхом перетворення змінних. Параметри рівняння регресії підбираються методом найменших квадратів.



Лінійна, парна або проста регресія

Нехай є сукупність експериментальних значень: $(x_i, y_i), i = 1, \dots, n$.

За методом найменших квадратів (МНК) потрібно підібрати такі значення параметрів a і b лінійної функції $y = ax + b$, які доставляють мінімум функції $Q(a, b)$:

$$Q(a, b) = \sum_{i=1}^n [y_i - (ax_i + b)]^2 \rightarrow \min.$$

Застосовуючи необхідні умови екстремуму для функції багатьох змінних, і, вирішуючи рівняння щодо a і b , отримаємо:

$$\begin{cases} a = \frac{\alpha_{1,1}^*[X, Y] - m_x^* \cdot m_y^*}{\alpha_2^*[X] - (m_x^*)^2} \\ b = m_y^* - a \cdot m_x^* \end{cases} \text{ або } \begin{cases} a = \frac{K_{xy}^*}{D_x^*} \\ b = m_y^* - a \cdot m_x^* \end{cases} \text{ або } \begin{cases} a = \frac{r_{xy}^* \cdot \sigma_y^*}{\sigma_x^*} \\ b = m_y^* - \frac{r_{xy}^* \cdot m_x^*}{\sigma_x^*} \cdot \sigma_y^* \end{cases}.$$

Лінійна залежність між x і y має вигляд:

$$y_x = ax + b = \frac{K_{xy}^*}{D_x^*} \cdot x + m_y^* - \frac{K_{xy}^*}{D_x^*} \cdot m_x^*,$$

$$y_x = \underbrace{\frac{r_{xy}^* \cdot \sigma_y^*}{\sigma_x^*}}_a \cdot x + \underbrace{\left(m_y^* - \frac{r_{xy}^* \cdot m_x^*}{\sigma_x^*} \cdot \sigma_y^* \right)}_b,$$

$$y_x = \frac{K_{xy}^*}{D_x^*} \cdot x + m_y^* - \frac{K_{xy}^*}{D_x^*} \cdot m_x^* \text{ або } y_x - m_y^* = \frac{K_{xy}^*}{D_x^*} \cdot (x - m_x^*).$$

a – вибірковий коефіцієнт регресії Y за X , який показує, на скільки одиниць в середньому змінюється Y при збільшенні X на 1 одиницю виміру (кожна з ВВ вимірюється в своїх одиницях виміру).

Оцінка коефіцієнта кореляції r_{xy}^* – є середня геометрична вибірових коефіцієнтів регресії в лінійних рівняннях регресії y_x та x_y , і має їх загальний знак.

Нехай a_{y_x} – вибірковий коефіцієнт регресії в рівнянні $y_x = a_{y_x} \cdot x + b_1$, а a_{x_y} – вибірковий коефіцієнт регресії в $x_y = a_{x_y} \cdot y + b_2$, тоді:

$$r_{xy}^{*2} = a_{y_x} \cdot a_{x_y}$$

$$r_{xy}^* = \pm \sqrt{a_{y_x} \cdot a_{x_y}}.$$

Модель регресії може бути побудована як за індивідуальними значеннями ознаки, так і за згрупованими даними. Для виявлення зв'язку між ознаками за досить великою кількістю випробувань потрібна кореляційна таблиця.

Для запису кореляційної таблиці парного зв'язку статистичні дані необхідно попередньо згрупувати за обома ознаками (X і Y), групуючи, наприклад, за рядками – факторну ознаку, а за стовпцями – результативну.

$y_j \backslash x_i$	1	...	j	...	l
1					
...					
i			n_{ij}		
...					
k					

Кореляційна таблиця дає загальне уявлення про напрям зв'язку. Якщо обидві ознаки розташовуються в порядку зростання, а частоти зосереджені навколо головної діагоналі, то можна судити про існування прямого зв'язку; а якщо навколо допоміжної діагоналі – зворотного.

Про тісноту зв'язку між ознаками можна судити за кучністю розташування частот навколо діагоналі. Чим частоти ближче до неї, тим зв'язок тісніше. Якщо в розташуванні частот немає системності, то можна судити про відсутність зв'язку.

Розрахункові формули:

$$m_X^* = \sum_{i=1}^k x_i' \cdot p_i^* = \frac{1}{n} \sum_{i=1}^k x_i' \cdot m_i ,$$

$$D_X^* = \sum_{i=1}^k (x_i')^2 \cdot p_i^* - (m_X^*)^2 ,$$

$$r_{XY}^* = (\sum_{j=1}^l \sum_{i=1}^k n_{ij} \cdot x_i' \cdot y_j' - n \cdot m_X^* \cdot m_Y^*) / (n \cdot \sigma_X^* \cdot \sigma_Y^*) ,$$

2.2 Завдання до лабораторної роботи

За вхідними даними провести статистичний регресійний аналіз двовимірної вибірки для двох випадкових величин X і Y в наступній послідовності. Побудувати діаграму розбігу в координатних осях за результатами вимірювань. Методом найменших квадратів знайти два лінійних рівняння регресії, прийнявши в якості незалежної змінної X , а

пояснювальної змінної Y , і, навпаки. Побудувати графічне зображення отриманих рівнянь регресії в одних координатних осях. Розрахувати середню помилку апроксимації для кожної моделі.

2.3 Варіанти завдань до лабораторної роботи

Індивідуальні варіанти завдань до лабораторної роботи №2 взяти з лабораторної роботи №1.

2.4 Питання для самоконтролю

1. Основні положення регресійного аналізу.
2. Метод найменших квадратів.
3. Кореляційне поле.
4. Система нормальних рівнянь для парної лінійної регресії.
5. Висновок рівнянь для регресійній моделі парної лінійної регресії на основі методу найменших квадратів.
6. Вибіркові коефіцієнти регресії.
7. Взаємозв'язок вибірових коефіцієнтів регресії з коефіцієнтом кореляції.

ЛІТЕРАТУРА: [1-2, 4-5,8].

ЛАБОРАТОРНА РОБОТА № 3
РАНГОВА КОРЕЛЯЦІЯ. НЕПАРАМЕТРИЧНІ ПОКАЗНИКИ
ВЗАЄМОЗВ'ЯЗКУ

Ціль роботи: засвоєння теоретичного матеріалу та набуття практичних навичок у застосуванні засад кореляційного непараметричного двовимірної аналізу або рангової кореляції.

3.1 Теоретична довідка

Рангова кореляція

В аналізі соціально-економічних явищ часто доводиться вдаватися до різних, елементарних показників, що характеризують напрямок і тісноту зв'язку. Дані коефіцієнти обчислюються за умови, що досліджувані ознаки підпорядковуються різним законам розподілу або для якісних ознак, які можуть бути ранжовані. Такі методи отримали назву непараметричних або рангової кореляції. До них відносять коефіцієнти Фехнера, Спірмена та Кендала.

Ранжування – процедура упорядкування об'єктів вивчення, яка виконується на основі важливості, значущості, переваги.

Ранг – порядковий номер значень ознаки, розташованих в порядку зростання або зменшення їх величин.

Якщо значення ознаки мають однакову кількісну оцінку, то ранг всіх цих значень приймається рівним середньої арифметичної від відповідних номерів місць, які їх визначають. Дані ранги називають зв'язаними або пов'язаними.

Коефіцієнт Фехнера оцінює зв'язок на основі порівняння знаків відхилень значень ознак від середніх арифметичних:

«-» – значення ознаки менше середньої;

«+» – значення ознаки більше середньої.

Збіг знаків за обома ознаками означає узгоджену варіацію, розбіжність – порушення узгодженості.

Коефіцієнт Фехнера розраховується, як:

$$K_{\Phi} = \frac{C - H}{C + H}$$

де C – кількість збігів знаків за обома ознаками, H – кількість розбіжностей, $0 \leq |K_{\Phi}| \leq 1$.

Якщо $K_{\Phi} = 1$ – повна узгоджена пряма мінливість;

$K_{\Phi} = 0$ – мінливість повністю не узгоджується;

$K_{\Phi} = -1$ – повна узгоджена зворотна мінливість.

Коефіцієнт кореляції рангів Спірмена

Для розрахунку коефіцієнта кореляції рангів Спірмена значення $BB\ X$ і Y нумеруються (кожен окремо) в порядку зростання (спадання) від 1 до n , тобто їм присвоюється певний ранг (r_i і s_i) – порядковий номер в ряду. Якщо зустрічається кілька однакових значень (X або Y), то кожному значенню присвоюється ранг, рівний приватному від ділення суми рангів, що припадають на ці значення, на число цих рівних значень (середнім арифметичним рангів величин, якби вони були різні). Такі ранги називають **зв'язаними**.

Різниця рангів: $d_i = r_i - s_i, i = \overline{1, n}$.

Ступінь тісноти зв'язку між досліджуваними ознаками за коефіцієнтом Спірмена:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

Коефіцієнт кореляції рангів Спірмена належить відрізку $[-1, 1]$.

1) Якщо ранги всіх об'єктів рівні, то $\rho = 1 \Rightarrow$ повний прямий зв'язок.

2) Якщо ранги об'єктів за двома змінним розташовані в зворотному порядку, то $\rho = -1 \Rightarrow$ повний зворотний зв'язок.

3) У всіх інших випадках то $|\rho| < 1$.

При існуванні зв'язаних рангів коефіцієнт кореляції Спірмена обчислюється:

$$\rho = 1 - \frac{\sum_{i=1}^n d_i^2}{\frac{1}{6}n(n^2-1) - (T_r + T_s)},$$

де $T_r = \frac{1}{12} \sum_{i=1}^{m_r} (t_r^3 - t_r)_i$, $T_s = \frac{1}{12} \sum_{i=1}^{m_s} (t_s^3 - t_s)_i$, m_r, m_s – число груп зв'язаних рангів; t_r, t_s – число рангів, що входять в групу зв'язаних рангів для змінних X і Y .

Перевірка значущості коефіцієнта Спірмена

Основна і альтернативна гіпотези:

$$H_0: \rho = 0.$$

$$H_1: \rho \neq 0.$$

Обмеження: $n > 10$.

$$\text{Статистика критерію: } t = \frac{\rho \sqrt{n-2}}{\sqrt{1-\rho^2}}.$$

Коефіцієнт Спірмена значущий на рівні значущості α , якщо виконується нерівність:

$$- \text{ для двостороннього критерію: } |t^*| > t_{1-\frac{\alpha}{2}}(n-2);$$

$$- \text{ для правостороннього критерію: } |t^*| > t_{1-\alpha}(n-2)$$

Коефіцієнт рангової кореляції Кендала

Коефіцієнт рангової кореляції Кендала (τ) - використовується для вимірювання взаємозв'язку між якісними і кількісними ознаками, характеризує однорідні об'єкти і ранжирування за першим принципом.

Розрахунок рангового коефіцієнта Кендала без зв'язкових рангів здійснюється за формулою:

$$\tau = \frac{2S}{n(n-1)},$$

де n – число спостережень;

S – сума різниць між числом послідовностей і числом інверсій за другою ознакою.

Розрахунок даного коефіцієнта виконується в наступній послідовності:

- 1) значення x ранжуються в порядку зростання або зменшення;
- 2) значення y розташовуються в порядку, відповідному значенням;
- 3) для кожного рангу визначається число наступних за ним значень рангів, що перевищують його величину, підсумовуючи отримані числа, визначаємо величину P , з огляду на її зі знаком «+»;
- 4) для кожного рангу визначається число наступних за ним рангів менших його величини, сумарна величина позначається через Q і враховується зі знаком «-»;
- 5) визначити S як алгебраїчну суму по всіх членах ряду:

$$S = P + Q.$$

Якщо в досліджуваній сукупності є зв'язкові ранги, то

$$\tau = \frac{S}{\sqrt{\left[\frac{n(n-1)}{2} - T_r\right] \cdot \left[\frac{n(n-1)}{2} - T_s\right]}},$$

$$\text{де } T_r = \frac{1}{2} \cdot \sum_{i=1}^{m_r} (t_r^2 - t_r)_i, T_s = \frac{1}{2} \cdot \sum_{i=1}^{m_s} (t_s^2 - t_s)_i \quad \text{визначається}$$

аналогічно, як у випадку коефіцієнта Спірмена.

При перевірці значущості τ коефіцієнта Кендала виходять з того, що в разі справедливості нульової гіпотези про відсутність кореляційної зв'язку між змінними ($n > 10$) має наближено нормальний закон розподілу

$$\text{з } m_\tau = 0 \text{ і } \sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

$$\text{Статистика } t = \tau \cdot \sqrt{\frac{9n(n-1)}{2(2n+5)}}.$$

H_0 відкидаємо, тобто t – значущий, якщо на рівні значущості α , де $t^* > t_{1-\alpha/2}$.

3.2 Завдання до лабораторної роботи

За вхідними даними провести статистичний кореляційний аналіз двовимірної вибірки для двох випадкових величин X і Y на основі непараметричних показників кореляції: рангових коефіцієнтів Фехнера, Спірмена та Кендалла. Перевірити їх значущість та порівняти з результатами параметричного кореляційного аналізу, що проведений у лабораторній роботі №1.

3.3 Варіанти завдань до лабораторної роботи

Індивідуальні варіанти завдань до лабораторної роботи №3 взяти з лабораторної роботи №1.

3.4 Питання для самоконтролю

1. Сутність непараметричних методів оцінки взаємозв'язку та особливості їх застосування.
2. Ранг та ранжування. Зв'язані ранги.
3. Коефіцієнт рангової кореляції Фехнера. Особливості застосування.
4. Коефіцієнт рангової кореляції Спірмена. Особливості застосування. Перевірка значущості коефіцієнту.
5. Коефіцієнт рангової кореляції Кендала. Особливості застосування. Перевірка значущості коефіцієнту.

ЛІТЕРАТУРА: [1-3, 4-5, 8].

ЛАБОРАТОРНА РОБОТА №4

ДИСПЕРСІЙНИЙ АНАЛІЗ

Ціль роботи: засвоєння теоретичного матеріалу та набуття практичних навичок у застосуванні засад дисперсійного однофакторного аналізу.

4.1 Теоретична довідка

Попередні методи статистичного аналізу даних використовують для порівняння двох об'єктів. У практиці досліджень часто виникають завдання, що стосуються групи об'єктів (наборів спостережуваних даних). Одним з методів для таких завдань є **дисперсійний аналіз** – статистичний метод виявлення на досліджувану випадкову величину (параметр) одночасної дії одного або декількох факторів. Дія деякого фактора на складну систему, спричинює мінливість його властивостей. Фактор може бути відомий або й невідомий, природного або штучного походження, як от: умови експерименту, методика вимірювань і опрацювання тощо.

За кількістю оцінюваних факторів дисперсійний аналіз поділяють на одно-, дво- та багатofакторний. Кожен фактор може бути дискретною чи неперервною випадковою змінною, яку розділяють на декілька сталих рівнів (градацій, інтервалів). Якщо кількість вимірювань (проб, даних) на всіх рівнях кожного з факторів однакова, то дисперсійний аналіз називають рівномірним, інакше – нерівномірним.

В основі дисперсійного аналізу є такий принцип (факт з математичної статистики): якщо на випадкову величину діють взаємно незалежні фактори **A, B, ...,** то загальна дисперсія дорівнює сумі дисперсій, зумовлених дією окремо кожного з факторів:

$$\sigma^2 = \sigma_A^2 + \sigma_B^2 + \dots$$

Цей метод ґрунтується на розділенні загальної дисперсії σ^2 на складові, що відповідають впливу різних джерел мінливості (дисперсія, зумовлена дією факторів, σ_1^2 і залишкова дисперсія σ_1^2), а застосовувані критерії дають змогу одночасно вивчати відмінності як у середніх значеннях, так і в дисперсіях.

Однофакторний дисперсійний аналіз

На практиці часто виникає необхідність узагальнення задачі перевірки суттєвості відмінності вибірових середніх для більш ніж двох сукупностей ($m > 2$). Нехай є m незалежних, нормально розподілених ВВ ($X_1, X_2, \dots, X_m$). Передбачається, що всі вони мають одне і теж стандартне відхилення σ і центри розподілу цих величин ($m_1, m_2, \dots, m_m$) в загальному випадку різні. Нехай над кожною ВВ проводиться серія з n спостережень.

Дані i -тої серії позначимо: $X_{i1}, X_{i2}, \dots, X_{in}, i = \overline{1, m}$.

Спираючись на статистичні дані, перевірити основну гіпотезу:

$$H_0: m_1 = m_2 = \dots = m_m.$$

Маємо, $m \cdot n$ – вимірювань x_{ij} , де i – номер ВВ, j – номер досвіду, $i = \overline{1, m}; j = \overline{1, n}$. Таблиця результатів вимірювань має вигляд:

$$\begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} = \|x_{ij}\|, i = \overline{1, m}; j = \overline{1, n}.$$

Кожна ВВ $X_{i1}, X_{i2}, \dots, X_{in}, i = \overline{1, m} - N(m_i, \sigma)$.

Позначимо усереднення за якимось індексом зіркою замість індексу. Середнє арифметичне з вимірів для i -й випадкової величини або групова середня для i -го рівня фактору, набуде вигляду:

$$\bar{x}_{i*} = \frac{1}{n} \sum_{j=1}^n x_{ij}.$$

Загальна середня:

$$\bar{x}_{**} = \bar{x} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n x_{ij} = \frac{1}{m} \sum_{i=1}^m \bar{x}_{i*}.$$

Загальна або повна сума квадратів відхилень окремих спостережень від загальної середньої:

$$Q = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2.$$

Сума квадратів відхилень групових середніх від загальної середньої або міжгрупова (факторна) сума квадратів відхилень (розбіг за факторами):

$$Q_1 = n \cdot \sum_{i=1}^m (\bar{x}_{i*} - \bar{x})^2.$$

Сума квадратів відхилень спостережень від групових середніх або внутрішньо групова (залишкова) сума квадратів відхилень:

$$Q_2 = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i*})^2.$$

У дисперсійному аналізі аналізуються не суми квадратів відхилень, а так звані середні квадрати, які є незміщеними оцінками відповідних дисперсій, які можна отримати, якщо поділити суму квадратів відхилень на відповідне число степенів свободи. Нагадаємо, що число степенів свободи визначається як загальне число спостережень мінус число рівнянь, що зв'язують їх.

Завдання з'ясування впливу фактору на результати експерименту: $H_0: m_1 = m_2 = \dots = m_m$ зводиться до порівняння виправлених оцінок факторної D_1^* і D_2^* залишкової дисперсій ($\sigma^2 = D$):

$$D_1^* = \frac{Q_1}{m-1}, D_2^* = \frac{Q_2}{mn-m}.$$

Порівняння двох дисперсій проводиться на основі перевірки наступної нульової гіпотези – $H_0: D_1 = D_2$.

У якості статистики критерію вибирається ВВ:

$$F = D_1/D_2,$$

що має розподіл Фішера з k_1 та k_2 степенями свободи: $k_1 = m - 1, k_2 = m(n - 1)$.

Гіпотеза H_0 відкидається, якщо фактично розраховане значення статистики $F^* = D_1^*/D_2^*$ більш ніж критичне $F_{\alpha; k_1, k_2}$, визначене на рівні значущості α при числі степенів свободи $k_1 = m - 1, k_2 = m(n - 1)$.

Гіпотеза H_0 приймається, якщо:

$$F^* = D_1^*/D_2^* \leq F_{\alpha; k_1, k_2}.$$

Стосовно до даної задачі спростування основної гіпотези означає наявність істотних відмінностей за рахунок впливу фактору на даному рівні значущості.

Зауваження. Для обчислення сум квадратів часто буває зручно використовувати такі формули:

$$Q_1 = \frac{1}{n} \sum_{i=1}^m (\sum_{j=1}^n x_{ij})^2 - \frac{1}{mn} (\sum_{i=1}^m \sum_{j=1}^n x_{ij})^2,$$

$$Q_2 = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2 - \frac{1}{n} \sum_{i=1}^m \left(\sum_{j=1}^n x_{ij} \right)^2,$$

$$Q = \sum_{i=1}^m \sum_{j=1}^n x_{ij}^2 - \frac{1}{mn} (\sum_{i=1}^m \sum_{j=1}^n x_{ij})^2.$$

тобто самі середні, взагалі кажучи, знаходити не обов'язково.

4.2 Завдання до лабораторної роботи

За індивідуальним варіантом, вирішити завдання лабораторної роботи з використанням дисперсійного аналізу та оцінити вплив

аналізованого фактору. Перевірити результат відповідним параметричним методом перевірки статистичних гіпотез.

4.3 Варіанти завдань до лабораторної роботи

№1. Методики навчання та оцінка якості знань на їх основі ознаки ефективності навчання трьома різними методиками.

Бали за тестами для трьох методик навчання		
Методика 1	Методика 2	Методика 3
28	39	41
33	52	49
42	53	56
47	54	62
48	56	63
50	58	64
50	59	65
51	63	72
60	64	77
71	77	87

№ 2. Дослідити залежність величини часу реакції від індивідуальних особливостей випробуваних за віком.

Час реакції для 4 груп			
Група 1	Група 2	Група 3	Група 4
630	690	810	830
690	660	780	650
760	780	690	720
780	740	620	770
810	810	710	870
760	830	810	830
830	690	780	650
740	660	690	720
860	830	620	770
730	840	710	870

№ 3. З'ясувати, чи залежить середнє число помилок на 1000 одиниць програмного коду від професійного досвіду розробника? Які з представлених груп відрізняються суттєво одна від одної?

Середнє число помилок для трьох груп програмістів		
Middle-developer	Junior-developer	Студент
28	39	41
33	52	49
12	53	56
17	54	62
28	56	63
40	58	64
30	59	65
21	63	72
30	64	77
11	77	87

№4. При вимірі продуктивності двох агрегатів отримано результати (кг речовини за годину роботи).

№ виміру	1	2	3	4	5	6	7	8	9	10	11	12
Агрегат А	14.1	10.1	14.7	13.7	14.0	13.5	12.7	13.4	11.7	10.3	14.2	12.5
Агрегат В	14.0	14.5	13.7	12.7	14.1	14.1	11.3	14.5	12.6	11.2	13.6	14.1

Чи можна вважати, що продуктивності агрегатів А та В однакові, припущення, що обидві вибірки отримані з нормально розподілених генеральних сукупностей?

№5. При дослідженні впливу двох типів покриття на питому провідність телевізійних трубок отримано такі результати (в умовних одиницях):

№ трубки	1	2	3	4	5	6	7	8	9	10	11	12
1 тип	6	5	12	9	10	7	6	11	4	9	14	4
2 тип	14	11	0	5	6	8	9	8	7	6	10	8

Чи можна виходячи з цих даних вважати, що тип покриття впливає граничну провідність трубок?

№6. Щоб визначити, який вплив дає температура навколишнього середовища на систематичну помилку інструменту, проведено вимірювання горизонтального кута об'єкта вранці, ($t = 10^{\circ}\text{C}$) і вдень ($t = 26^{\circ}\text{C}$). Результати вимірювань (у кутових секундах) такі:

Вранці	38.2	36.4	37.7	36.1	37.9	37.8	37.5	39.2	36.5	38.4	38.2
Вдень	39.5	38.7	37.8	38.6	39.2	39.1	38.9	37.8	37.6	37.3	37.8

Чи можна вважати, що температура довкілля впливає на систематичну помилку кутомірного інструменту?

№7. Для 10 осіб було запропоновано спеціальну дієту. Після 2-х тижневого харчування за цією дієтою маса їх тіла змінилася таким чином:

Маса до дієти	68	80	92	81	70	79	78	66	57	76	73
Маса після дієти	60	84	87	79	74	71	72	67	57	70	81

Чи можна рекомендувати цю дієту для людей, які бажають схуднути?

№8. Порівнювалися дії 2-х екстрактів вірусу тютюнової мозаїки. Для цього кожна половина листа натиралася відповідним препаратом. Число уражених місць наводиться у таблиці:

Екстракт А	20	39	43	13	28	26	17	49	36	31	14	34
Екстракт В	31	22	45	6	21	13	17	46	31	41	24	41

Чи можна вважати, що дія цих екстрактів по-різному?

№9. Вивчався вплив чорної та квітневої пари на врожай жита. Досвід тривав 11 років.

Враховувалася маса 1000 зерен у грамах. Результати досвіду:

Роки	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
по чорному пару	31.1	24	24.6	28.6	29.1	30.1	25.4	27.2	29.8	22.8	29.2
по квітневому пару	31.6	24.2	24.8	19.1	29.9	31	26.3	31.1	19.8	23.4	31.6

Чи можна вважати, що врожай жита по квітневому пару значно вище, ніж по чорному?

№10. Зміна врожайності при застосуванні одного з видів передпосівного обробітку насіння характеризується такими даними (у центнерах з гектара):

Рік	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Необроблене насіння	20	17.9	20.6	22	21.4	23.8	21.4	19.8	18.4	19.6	18.7
Оброблене насіння	22.1	18.5	19.4	22.1	21.7	24.9	21.6	20.3	18.3	18.9	19.4

Чи можна вважати, що передпосівний обробіток збільшує врожайність?

№11. Вимірювалося напруга пробою у діодів, відібраних випадковим чином із 2-х партій. Результати вимірювання (у вольтях):

1 партія	39	50	61	67	40	40	54	67	52	59	44	38	57
2 партія	60	53	42	41	40	54	63	69	48	62	56	50	53

Чи можна вважати, що у діодів 2-ї партії напруга пробою вище, ніж у 1-ої.

№12. У польових транзисторів із двох партій, виготовлених із застосуванням різної технології, вимірювався диференціальний опір каналу R_i . Результати вимірювань (у мікроомах):

Технологія А	0.01	0.02	0.12	0.3	0.29	0.15	0.21	0.18	0.27	0.3	0.18
Технологія В	0.15	0.07	0.25	0.15	0.22	0.18	0.18	0.22	0.23	0.27	0.26

Чи технологія виготовлення на величину диференціального опору каналу R_i ?

№13. За вибірками з двох партій мікросхем після операції легування полікремнію вимірювався опір. Результати вимірів:

1 партія	52.2	33	76	32.5	49.5	32.5	191.5	112.5
	52.9	114.8	33.7	69.1	112.5	48.5	16.5	-
2 партія	119	17.5	43.5	43.5	90.5	40.0	50.0	108
	2.4	16.5	7.5	96	46	48.5	50.0	-

Чи можна стверджувати, що обидві партії отримані з однієї генеральної сукупності?

№14. Вивчався вплив кобальту на збільшення маси тіла кроликів. Досвід проводився на 2-х групах тварин – дослідній та контрольній. Обидві групи містилися на одному кормовому раціоні, але дослідні кролики щодня отримували по 0.06 г хлористого кобальту на кг ваги тіла. За час досвіду спостерігалось наступне збільшення в м'ясі (за один тиждень):

Контрольні	560	580	600	420	530	490	580	470	490	640	575
Дослідні	692	700	621	640	561	680	630	530	580	420	611

Чи можна стверджувати, що добавки хлористого кобальту дійсно дають збільшення маси тіла?

№15. Двом групам піддослідних пропонувалося провести упізнання трьох накреслень цифри 5. Результати експерименту (секунди):

1 група	25	28	27	29	26	24	28	23	30	25	26	25
2 група	18	19	31	32	17	15	41	35	38	13	14	24

Чи можна вважати, що результати дослідів є різними?

№16. Протягом деякого часу добова продуктивність двох автоматів характеризується такими даними:

1 автомат	105	60	83	111	138	71	87	130	93	105	122
2 автомат	172	45	51	55	117	103	82	93	31	51	71

Перевірити гіпотезу про збіг показників результатів роботи автоматів.

№17. Для контролю налаштування двох верстатів-автоматів, що виробляють деталі по одному кресленню, визначили відхилення від номінальних розмірів у кількох деталей, виготовлених на обох верстатах. В результаті отримали дані (мкм):

Станок А	44	-14	32	8	-50	20	-35	15	10	-8	-20	5
Станок В	52	-49	61	-35	-48	18	-45	35	23	21	-59	-19

Перевірити гіпотезу про збіг результатів роботи верстатів за цими даними на рівні значущості.

№18. Контрольований розмір кількох деталей було перевірено до та після налагодження верстата.

В результаті отримані дані (мм):

До початку	36.4	37.5	36.9	38.1	35.5	37.8	38.3	36.6	38.4	37.5	37.1
Після налагодження	36.8	39.2	37.6	39.9	34.2	36.5	36.3	39.8	38.4	37.5	38.6

Чи можна вважати, що результати вимірювань до та після налагодження верстата є різними?

№19. Середня врожайність зернових культур двох районах за 1991-2001г.г. характеризується такими даними, ц/га:

Рік	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
1-й район	30	20	23	16	22	19	31	28	19	23	15
2-й район	25	34	30	28	29	22	23	17	31	17	28

Чи можна вважати, що врожайність другого району вища, ніж перший?

№20. В результаті демографічних досліджень у двох областях були отримані наступні результати за кількістю членів сім'ї, що спільно проживають:

Число членів в родині	2	3	4	5	6	7	8	9	10	11	12	13
Область 1	15	34	25	16	8	10	7	22	1	0	1	0
Область 2	17	39	20	14	10	21	19	15	1	2	0	0

Чи вірно, чи результати демографічного дослідження показали рівні результати по обох областях?

№21. При плануванні вибіркового обстеження зайнятості населення сільських районів республіки є такі дані:

Район	Чоловіки	Жінки
1	3.5	3.1
2	5.6	5.0
3	1.7	2.1
4	2.8	3.0
5	3.0	3.7
6	4.1	3.8
7	5.1	1.8
8	3.7	4.2
9	4.5	2.6
10	1.9	3.2
11	4.8	3.0

Чи можна сказати, що рівень зайнятості чоловіків та жінок однаковий?

№22. У процесі підготовки обстеження якості кондитерських виробів, що імпортуються, було проведено пробну перевірку 11 ящиків для збору інформації про варіацію їх ваги різними пакувальниками:

№ ящика	Упаковщик 1	Упаковщик 2
1	540	530
2	520	530
3	550	540
4	510	520
5	530	530
6	560	500
7	520	570
8	500	510
9	510	530
10	540	520
11	570	560

Чи однаково фасуються кондитерські вироби?

№23. На Всеукраїнському конкурсі фірм-виробників молочних продуктів одним із критеріїв вибору був термін придатності товарів. Дані двох фіналістів наведені у таблиці:

Число днів	1	2	3	4	5	6	7	8	9	10	11	12
Ласуня	3	2.5	3.5	2	4	1.5	3.0	1.0	4.0	3.5	2.5	3.0
Президент	3.5	2.5	4.0	1.5	3.0	2.0	2.5	2.5	3.0	2.5	3.5	4.0

Хто переможе у цьому конкурсі?

№24. В порядку експерименту для свиней додавали хімічні добавки, які впливають на швидкість росту тварин. Дані про результати експерименту наведено у таблиці:

До	200	210	220	250	270	240	290	230	280	260	235
Після	220	200	230	235	290	260	270	250	240	225	280

Чи вплинуло застосування добавок на вагу свиней?

№25. Із двох партій монет вартістю 5 коп. було вибрано по 12 штук, які зважували на терезах. Результати цих зважувань представлені у вигляді 2-х статистичних розподілів:

y_i , мг	9,4	9,6	9,8	10,2	9,33	9,63	9,75	10,1	9,5
n_i	5	2	3	2	2	3	1	3	2

Чи суттєва різниця між вагою монеток?

№26. Кров'яний тиск було виміряно у 11 осіб віком 40 років з одного району міста та у 11 осіб того ж віку з іншого району міста:

1 -й район	114	116	118	117	114	120	116	122	119	124	115
2 -й район	15	19	17	21	15	124	130	127	118	117	122

Чи однаковий тиск у осіб віком 40 років в різних районах міста?

№27. Протягом року вимірювалася продуктивність праці (у тис. грн/прац.) у двох будівельних фірмах:

1 -я фірма	120	150	180	210	120	140	120	240	180	270	160	170
2 -я фірма	90	150	130	170	120	210	180	160	250	170	290	90

Чи можна вважати, що на 2-ой фірмі продуктивність праці вища?

№28. Досліджувався місячний прибуток працівників у гривнях двох заводів однієї і тієї ж галузі виробництва:

1 -й завод	150	153	160	167	170	150	164	180	172	190	185
2 -й завод	140	150	160	180	162	170	180	172	160	185	200

Чи можна вважати, що прибуток працівників на 1-му заводі значно перевищує прибуток працівників 2-го заводу?

№29. Вимірювалося денне споживання олії одним мешканцем у двох регіонах країни:

1 -й регіон	35	35.2	35.4	35.7	35.6	35.8	36	35.9	34.9	35.3	35.7
2 -й регіон	35.4	35.8	35	36.2	35.2	35	35.8	36.6	36.0	35.1	37

Чи істотно відрізняється споживання олії у різних регіонах?

№30. Вимірювалася жива маса курей на 2-х птахофермі:

1-я	96.5	99.5	97.0	98.4	99.0	102.5	108.5	111.5	110.6	107	110
2-я	85.5	105.5	97.9	125.5	99.5	145.5	111.7	102.5	110.9	107	12

Чи можна вважати, що маса курей на 2-й птахофермі значно перевищує масу курей на 1-ій?

4.4 Питання для самоконтроля

1. Суть і область застосування методу дисперсійний аналіз.
2. Однофакторний і багатфакторний дисперсійний аналіз
3. Обмеження при використанні методу дисперсійний аналіз.
4. Основне співвідношення дисперсійного аналізу.
5. Внутрішньо групова дисперсія і її оцінка.
6. Міжгрупова дисперсія і її оцінка
7. Загальна дисперсія і її оцінка.

ЛІТЕРАТУРА: [1,5-8].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник для вузов. – М.: Юнити-Дана, 2000. – 543с.
2. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений – М.: Наука, 1969. – 509с.
3. Жлуктенко В.І., Наконечний С.І., Савіна С.С. Теорія ймовірностей і математична статистика: Навч.посіб. у 2-х час. – Ч. II. Математична статистика. – К.: КНЕУ, 2001. – 336с.
4. Вентцель Е.С. Теория вероятностей. – М.: Наука, 1969. – 576с.
5. Статистика /Под ред. Ионина В.Г. – М.: Инфра, 2000. – 310с.
6. Практикум з теорії ймовірностей та математичної статистики: Нав. посібник для студ. вищ. навч. закладів / За ред.. Р.К.Чорнея. –К.: МАУП, 2003.– 328с.
7. Методические указания и задания к лабораторным работам по курсу “СТАТИСТИКА, часть 2 (для студентов области знаний 0305 “Экономика и предпринимательство” / разраб.: Назарова И.А., Яценко Е.А. – Донецк: ДонНТУ, 2009. – 51 с.
8. Методичні вказівки до самостійної роботи за курсом "Статистичні методи оцінювання" для магістрів, що навчаються за спеціальністю 121 Інженерія програмного забезпечення денної форми навчання / Укл.: І.А. Назарова. - Покровськ: ДонНТУ, 2019. - 39с.