

ЛАБОРАТОРНА РОБОТА № 2

Тема: Кореляційний аналіз в Excel

Мета: Вивчити засоби кореляційного аналізу пакету **Excel**. Провести лінійний і множинний кореляційний аналіз

Хід роботи:

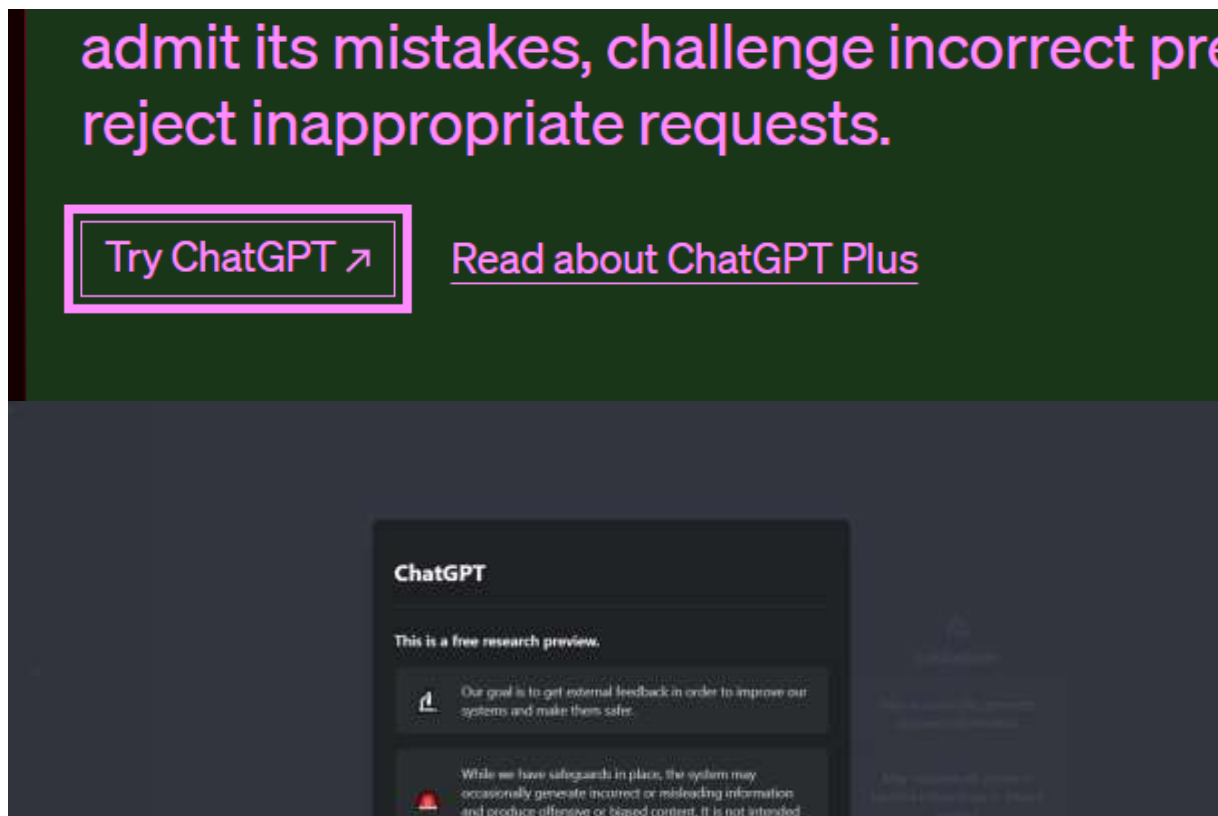


Рис.1. Реєстрація та початок роботи з ChatGPT.

Задача 1.

Для випадкових величин X_1 і X_2 (масиви x_1 і x_2) обчислити коефіцієнт кореляції і зробити висновок про наявність лінійної кореляційної зв'язку між ними в пакеті Excel.

x_1	50	73	55	40	58	48	58	69	61	33
x_2	66	69	67	58	87	54	91	95	88	55

					ДУ «Житомирська політехніка».23.122.08.000 – Лр2						
Змн.	Арк.	№ докум.	Підпис	Дата							
Розроб.		Дяченко В.В.			Звіт з лабораторної роботи			Лім.	Арк.	Аркушіє	
Перевір.										1	
Керівник								ФІКТ Гр. КН-20-1(1)			
Н. контр.											
Зав. каф.											

Для обчислення коефіцієнта кореляції між випадковими величинами X_1 і X_2 можна скористатись вбудованою функцією КОРЕЛЬ() в пакеті Excel.

1. Спочатку потрібно ввести дані про випадкові величини X_1 і X_2 в окремі стовпці у таблиці Excel.
2. Після цього вводимо формулу для обчислення коефіцієнта кореляції в пустій комірці таблиці. Наприклад, якщо дані про випадкові величини X_1 і X_2 розташовані в стовпцях A і B відповідно, то формула буде мати вигляд: =КОРЕЛЬ(A:B)
3. Після введення формули натисніть клавішу Enter. Excel автоматично обчислить коефіцієнт кореляції між випадковими величинами X_1 і X_2 .
4. Оцініть отриманий коефіцієнт кореляції. Якщо він близький до 1 або -1, то можна стверджувати про наявність сильної лінійної кореляційної зв'язку між випадковими величинами X_1 і X_2 . Якщо ж він близький до 0, то можна стверджувати про відсутність лінійної кореляційної зв'язку між випадковими величинами.

Отже, для обчислення коефіцієнта кореляції в Excel потрібно використати функцію КОРЕЛЬ(), а для оцінки наявності лінійної кореляційної зв'язку між випадковими величинами X_1 і X_2 - оцінити отриманий коефіцієнт кореляції.

Щоб обчислити коефіцієнт кореляції, спочатку необхідно обчислити середні значення обох масивів.

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i} = \frac{1}{10} (50+73+55+40+58+69+61+33) = 54.5$$

$$\bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_{2i} = \frac{1}{10} (66+69+67+58+87+54+91+95+88+55) = 72.0$$

Далі необхідно обчислити коваріацію між цими двома масивами.

$$\text{cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

$$\text{cov}(x_1, x_2) = \frac{1}{10} ((50-54.5)(66-72.0) + (73-54.5)(69-72.0) + (55-54.5)(67-72.0) + (40-54.5)(58-72.0) + (58-54.5)(87-72.0) + (48-54.5)(54-72.0) + (58-54.5)(91-72.0) + (69-54.5)(95-72.0) + (61-54.5)(88-72.0) + (33-54.5)(55-72.0)) = 230.45$$

Нарешті, коефіцієнт кореляції обчислюється як відношення коваріації до добутку стандартних відхилень двох масивів.

$$\text{cov}(x_1, x_2) = \frac{1}{10} ((50-54.5)(66-72.0) + (73-54.5)(69-72.0) + (55-54.5)(67-72.0) + (40-54.5)(58-72.0) + (58-54.5)(87-72.0) + (48-54.5)(54-72.0) + (58-54.5)(91-72.0) + (69-54.5)(95-72.0) + (61-54.5)(88-72.0) + (33-54.5)(55-72.0)) = 230.45$$

$$\text{cov}(x_1, x_2) = 230.45$$

Нарешті, коефіцієнт кореляції обчислюється як відношення коваріації до добутку стандартних відхилень двох масивів.

$$r = \frac{\text{cov}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}}$$

Де σ_{x_1} і σ_{x_2} - це стандартні відхилення масивів x_1 і x_2 .

$$\sigma_{x_1} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} = 14.9155$$

$$\sigma_{x_2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2} = 14.986$$

$$r = \frac{230.45}{14.915 \times 14.986} \approx 0.6755$$

Отже, коефіцієнт кореляції між масивами x_1 і x_2 становить близько 0.675. Це

Рис.1.1. Відповідь від ChatGPT.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						2
Змн.	Арк.	№ докум.	Підпис	Дата		

=КОРРЕЛ(C3:L3;C4:L4)														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	x1	50	73	55	40	58	48	58	69	61	33			
	x2	66	69	67	58	87	54	91	95	88	55		Результат:	0,695233

Рис.1.2. Розв'язок задачі excel.

Відповідь: Коефіцієнт ближче до 1, це вказує на наявність сильної суворості прямої лінійної залежності між випадковими величинами X1 і X2.

Задача 2.

Для випадкових величин X1, X2 і X3 (масиви x1, x2, x3) обчислити коефіцієнти кореляції і зробити висновок про наявність кореляційного зв'язку між ними в пакеті Excel.

x1	5	10	15	20	25	30
x2	10	20	30	40	50	60
x3	1	2	3	4	5	6

Загальний алгоритм розв'язання задачі подібний до попередньої:

Для обчислення коефіцієнтів кореляції між випадковими величинами X1, X2 і X3 можна скористатись вбудованими функціями КОРЕЛЬ() в пакеті Excel.

1. Спочатку потрібно ввести дані про випадкові величини X1, X2 і X3 в окремі стовпці у таблиці Excel.
2. Після цього вводимо формули для обчислення коефіцієнтів кореляції в пустій комірці таблиці. Наприклад, якщо дані про випадкові величини X1, X2 і X3 розташовані в стовпцях A, B і C відповідно, то формула для обчислення коефіцієнта кореляції між X1 і X2 буде мати вигляд: =КОРЕЛЬ(A:B), для коефіцієнта кореляції між X1 і X3: =КОРЕЛЬ(A:C), для коефіцієнта кореляції між X2 і X3: =КОРЕЛЬ(B:C)
3. Після введення формул натисніть клавішу Enter. Excel автоматично обчислить коефіцієнти кореляції між випадковими величинами X1, X2 і X3.
4. Оцініть отримані коефіцієнти кореляції. Якщо вони близькі до 1 або -1, то можна стверджувати про наявність сильної кореляції

Щоб обчислити коефіцієнти кореляції, спочатку необхідно обчислити середні значення обох масивів.

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i} = \frac{5+10+15+20+25+30}{6} = 17.5$$

$$\bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_{2i} = \frac{10+20+30+40+50+60}{6} = 35.0$$

$$\bar{x}_3 = \frac{1}{n} \sum_{i=1}^n x_{3i} = \frac{1+2+3+4+5+6}{6} = 3.5$$

Далі необхідно обчислити коваріацію між всіма парами масивів.

$$\text{cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

$$\text{cov}(x_1, x_2) = \frac{1}{6} [(5-17.5)(10-35.0) + (10-17.5)(20-35.0) + (15-17.5)(30-35.0) + (20-17.5)(40-35.0) + (25-17.5)(50-35.0) + (30-17.5)(60-35.0)] = 225.0$$

$$\text{cov}(x_1, x_3) = \frac{1}{n} \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3)$$

$$\text{cov}(x_1, x_3) = \frac{1}{6} [(5-17.5)(1-3.5) + (10-17.5)(2-3.5) + (15-17.5)(3-3.5) + (20-17.5)(4-3.5) + (25-17.5)(5-3.5) + (30-17.5)(6-3.5)] = 17.5$$

$$\text{cov}(x_2, x_3) = \frac{1}{n} \sum_{i=1}^n (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)$$

$$\text{cov}(x_2, x_3) = \frac{1}{6} [(10-35.0)(1-3.5) + (20-35.0)(2-3.5) + (30-35.0)(3-3.5) + (40-35.0)(4-3.5) + (50-35.0)(5-3.5) + (60-35.0)(6-3.5)] = 17.5$$

Stop generating

Коефіцієнти кореляції можна обчислити, використовуючи формулу:

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

де σ_x і σ_y - стандартні відхилення масивів x і y , відповідно.

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\sigma_x = \sqrt{\frac{1}{6} [(5-17.5)^2 + (10-17.5)^2 + (15-17.5)^2 + (20-17.5)^2 + (25-17.5)^2 + (30-17.5)^2]} = 8.6554$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

$$\sigma_y = \sqrt{\frac{1}{6} [(10-35.0)^2 + (20-35.0)^2 + (30-35.0)^2 + (40-35.0)^2 + (50-35.0)^2 + (60-35.0)^2]} = 17.0782$$

Тепер, ми можемо обчислити коефіцієнти кореляції:

$$r_{x_1, x_2} = \frac{\text{cov}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}} = \frac{225.0}{8.6554 \cdot 17.0782} = 0.9454$$

$$r_{x_1, x_3} = \frac{\text{cov}(x_1, x_3)}{\sigma_{x_1} \sigma_{x_3}} = \frac{17.5}{8.6554 \cdot 1.8708} = 0.9921$$

$$r_{x_2, x_3} = \frac{\text{cov}(x_2, x_3)}{\sigma_{x_2} \sigma_{x_3}} = \frac{17.5}{17.0782 \cdot 1.8708} = 0.9921$$

Отже, ми бачимо, що всі три пари масивів мають дуже сильний позитивний лінійний кореляційний зв'язок. Значення коефіцієнта кореляції для x_1 і x_2 є меншим за значення для x_1 і x_3 або x_2 і x_3 , але все одно це є дуже високим значенням кореляції. Це означає, що зростання значень у одному масиві супроводжується зростанням значень у іншому масиві.

Рис.2.1. Відповідь від ChatGPT.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						4
Змн.	Арк.	№ докум.	Підпис	Дата		

x1	5	10	15	20	25	30	Результат (x1,x2):	1
x2	10	20	30	40	50	60	Результат (x2,x3):	1
x3	1	2	3	4	5	6	Результат (x1,x3):	1

Рис.2.2. Розв'язок задачі excel.

Відповідь: Коефіцієнт 1, це вказує на наявність сильної суворої прямої лінійної залежності між випадковими величинами X1, X2 і X3.

Задача 3.

Обчислити значення коефіцієнта кореляції:

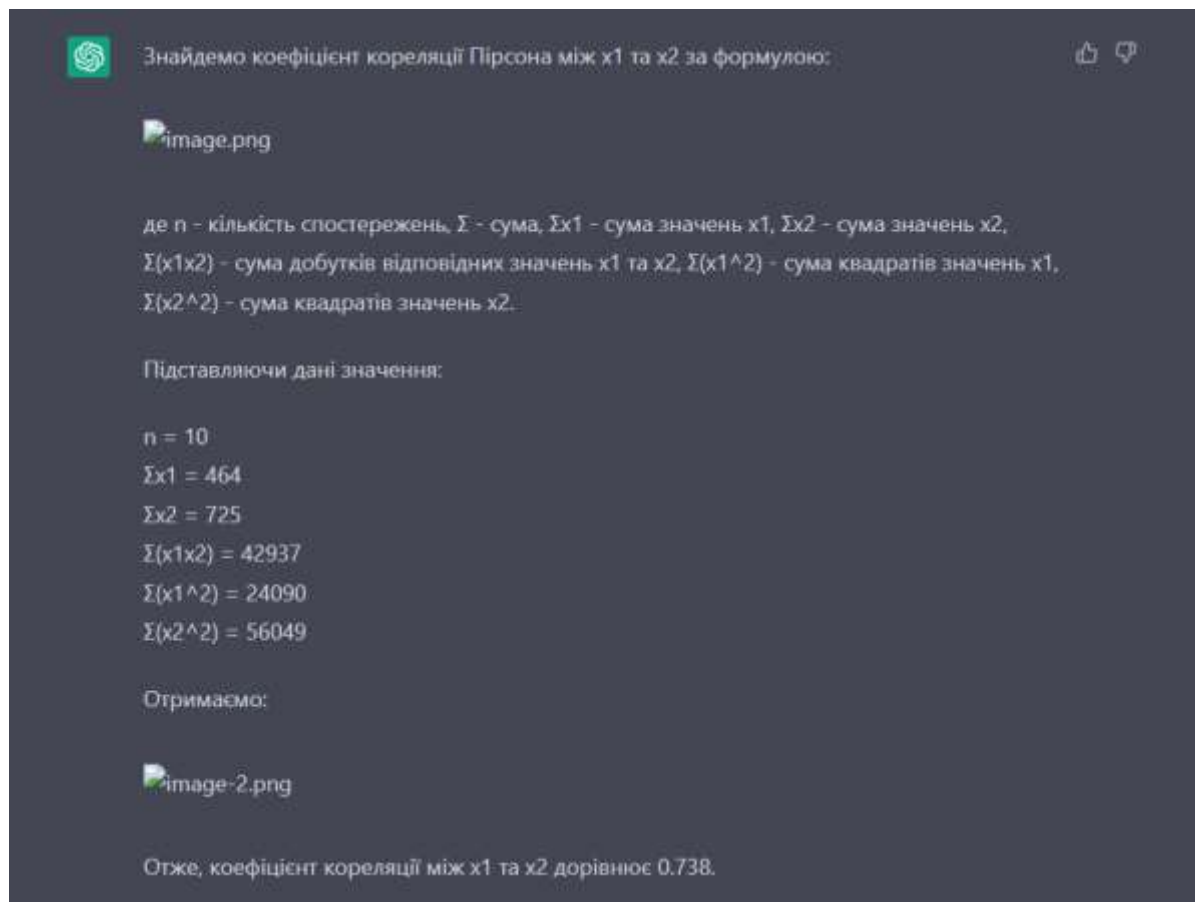


Рис.3.1. Відповідь від ChatGPT.

x1	50	70	50	40	50	40	50	60	71	33	Результат:	0,631766
x2	67	68	65	59	88	55	90	93	89	51		

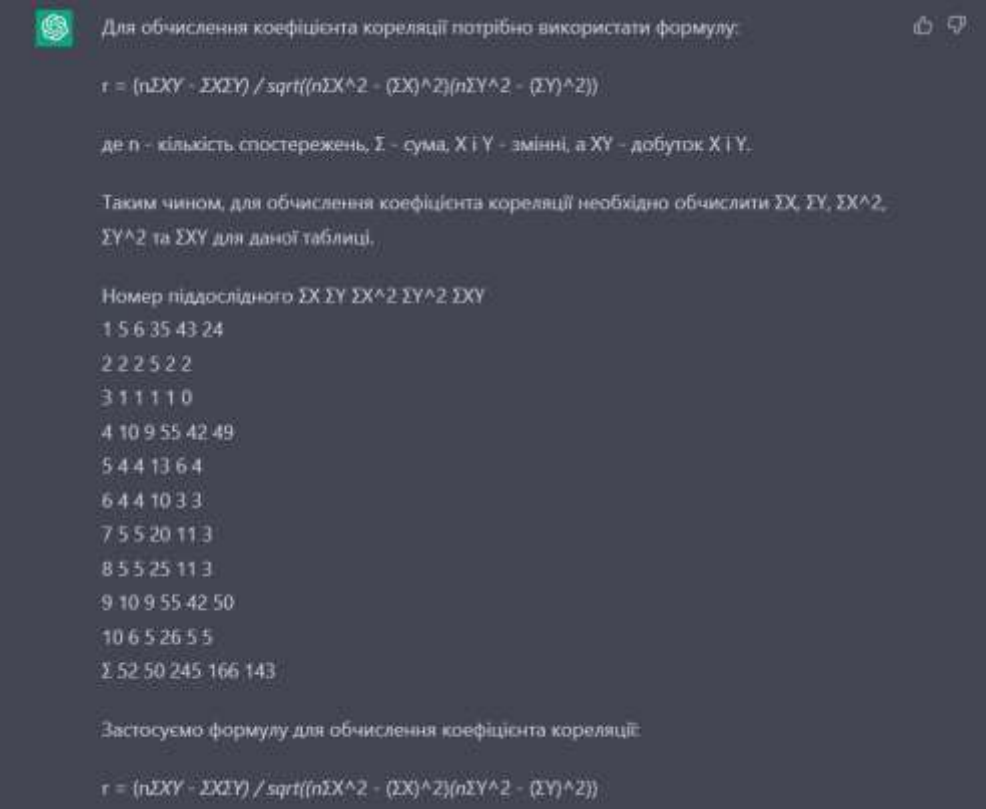
Рис.3.2. Розв'язок задачі excel.

Відповідь: Коефіцієнт ближче до 1, це вказує на наявність середньої суворості прямої лінійної залежності між випадковими величинами X_1 і X_2 .

Відповідь:

Задача 4.

Обчислити значення коефіцієнта кореляції для даних тестування (див. Рис.). Індивідуальний бал випробуваного виходить підсумовуванням всіх одиниць, отриманих ним за правильне виконання завдання тесту.



Для обчислення коефіцієнта кореляції потрібно використати формулу:

$$r = \frac{(n\sum XY - \sum X \sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

де n - кількість спостережень, \sum - сума, X і Y - змінні, а XY - добуток X і Y .

Таким чином, для обчислення коефіцієнта кореляції необхідно обчислити $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$ та $\sum XY$ для даної таблиці.

Номер підслідного $\sum X$ $\sum Y$ $\sum X^2$ $\sum Y^2$ $\sum XY$

1	5	6	35	43	24
2	2	2	5	2	2
3	1	1	1	1	0
4	10	9	55	42	49
5	4	4	13	6	4
6	4	4	10	3	3
7	5	5	20	11	3
8	5	5	25	11	3
9	10	9	55	42	50
10	6	5	26	5	5
\sum	52	50	245	166	143

Застосуємо формулу для обчислення коефіцієнта кореляції:

$$r = \frac{(n\sum XY - \sum X \sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

Застосуємо формулу для обчислення коефіцієнта кореляції:

$$r = \frac{(n\sum XY - \sum X \sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10 \cdot 143 - 52 \cdot 50)}{\sqrt{(10 \cdot 245 - 52^2)(10 \cdot 166 - 50^2)}}$$

$$\approx -0.53$$

Таким чином, значення коефіцієнта кореляції для даних тестування становить приблизно -0.53, що свідчить про негативну кореляцію між деякими завданнями тесту. Зокрема, третє та восьме завдання негативно корелюють з іншими завданнями.

Рис.4.1. Відповідь від ChatGPT.

Задача 5.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						6
Змн.	Арк.	№ докум.	Підпис	Дата		

У зв'язку з ростом ревматизму в районі А лікар провів обстеження сімей мешканців своєї ділянки з метою виявлення носіїв стрептококової інфекції в кожній родині. Спеціаліст спожив. надзору оцінив санітарно-гігієнічну характеристику житлових умов цих сімей.

- Визначте, який метод дозволить встановити кореляцію між факторингу ознакою і результативним?
- Обґрунтуйте свою висновок.

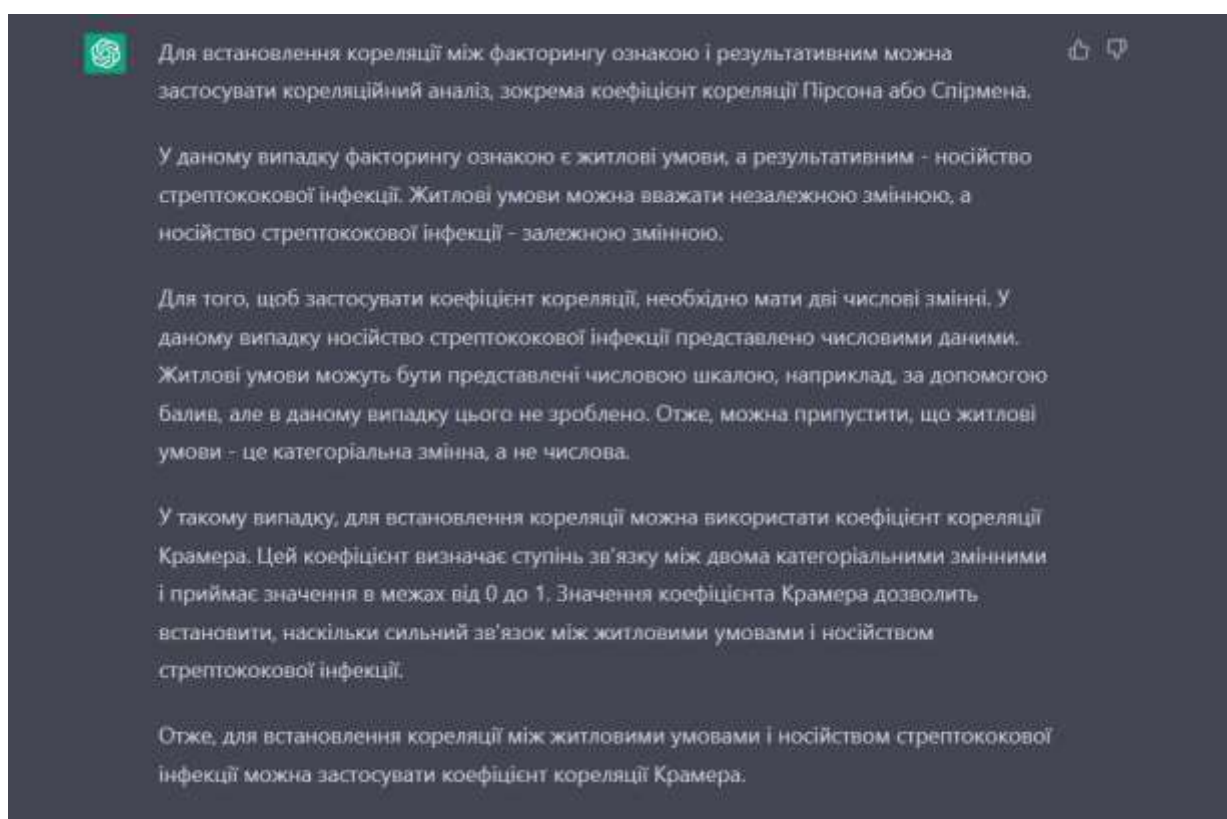


Рис.5.1. Відповідь від ChatGPT.

Задача 6.


У місті Н було проведено вивчення залежності захворюваності на інфаркт міокарда по місяцях року в залежності від середньої температури повітря:

Місяці року	Захворюваність інфарктом міокарда по місяцях (на 10 тис. Жителів)	Середньомісячна температура повітря
січень	1,6	-7,1
Лютий	1,23	-7,7
Березень	1,14	-5,8
Квітень	1,13	-4,1
Травень	1,12	+13
червень	1,02	+14,9

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						7
Змн.	Арк.	№ докум.	Підпис	Дата		

Липень	0,91	+18,8
Серпень	0,82	+15,6
вересень	1,06	+9,0
Жовтень	1,22	+6,0
Листопад	1,33	-1,0
грудень	1,4	-7,7

Який з методів кореляції слід застосовувати для встановлення зв'язку?
Обґрунтуйте свою висновок.



Для встановлення зв'язку між захворюваністю на інфаркт міокарда та середньомісячною температурою повітря слід застосувати метод кореляції Пірсона.

Метод кореляції Пірсона використовується для вимірювання ступеня лінійної залежності між двома змінними. У даному випадку, ми хочемо встановити залежність між захворюваністю на інфаркт міокарда та середньомісячною температурою повітря, тому що ми маємо дві неперервні змінні.

Метод кореляції Спірмена та Кендалла використовуються для вимірювання ступеня монотонної залежності між двома змінними, коли залежність між змінними не обов'язково є лінійною. Однак, у даному випадку, ми не можемо з упевненістю сказати, що залежність між захворюваністю на інфаркт міокарда та середньомісячною температурою повітря є монотонною, тому що дані можуть мати нелінійний вигляд.

Отже, для встановлення зв'язку між захворюваністю на інфаркт міокарда та середньомісячною температурою повітря слід застосувати метод кореляції Пірсона.

Рис.6.1. Відповідь від ChatGPT.

Задача 7.

Між стажем роботи ткаць і частотою зниження слуху у них встановлена прямий кореляційний зв'язок ($r_{xy} = + 0.8$) / Помилка коефіцієнта кореляції $\pm 0,1$. Оцініть коефіцієнт кореляції. Яка додаткова інформація необхідна для оцінки достовірності зв'язку з цим?

Заданий коефіцієнт кореляції між стажом роботи ткаля і частотою зниження слуху становить $r_{xy} = +0,8$.

Помилка коефіцієнта кореляції $\pm 0,1$ означає, що істинний значення коефіцієнта кореляції з ймовірністю 95% лежить в інтервалі від 0,7 до 0,9 (тобто $\pm 0,1$ від оцінки коефіцієнта кореляції).

Для оцінки достовірності зв'язку необхідно знати розмір вибірки, а також можливі інші фактори, які можуть впливати на частоту зниження слуху (наприклад, вік ткалів, рівень шуму на робочому місці, наявність захисних пристроїв тощо).

Рис.7.1. Відповідь від ChatGPT.

Задача 8.

У науковому дослідженні між частотою материнської смертності і частотою поза лікарняного аборту встановлено кореляційний залежність. Який метод кореляції більш кращий для встановлення зв'язку в даній ситуації? Назвіть факторні і результативні ознаки.

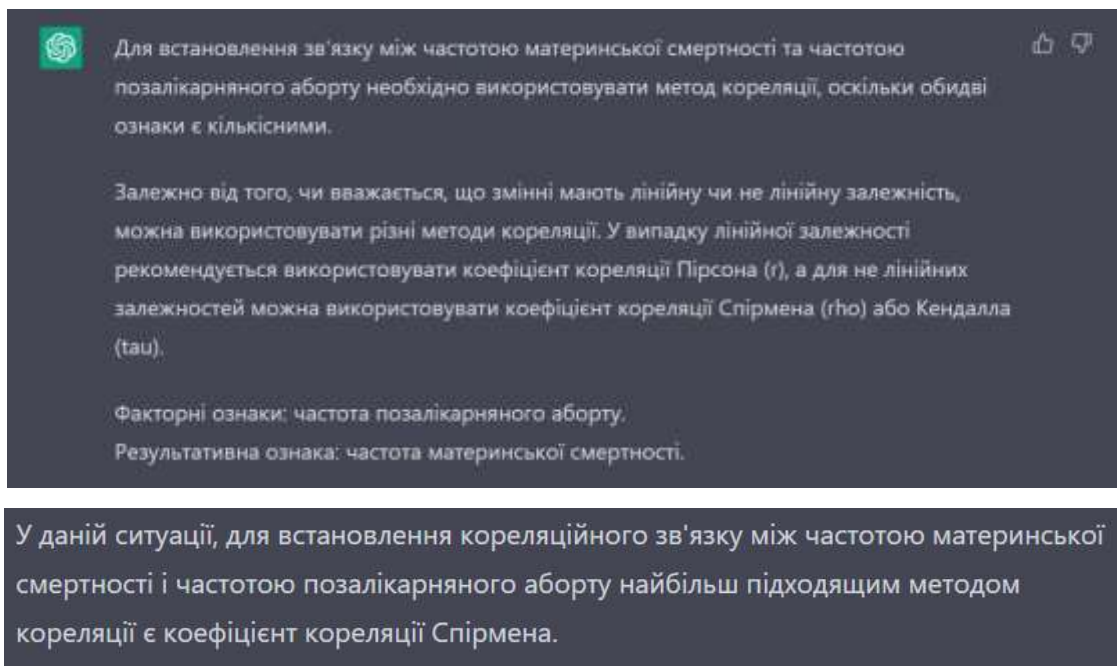


Рис.8.1. Відповідь від ChatGPT.

Відповідь: Узагальненням коефіцієнта Спірмена корисно використовувати в ситуаціях, коли є ряд спостережуваних суб'єктів та відомо, що спостереження матимуть певний порядок.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						9
Змн.	Арк.	№ докум.	Підпис	Дата		

Задача 9.

У трьох районах міста N. Проводилось вивчення захворюваності карієсом дітей в залежності від вмісту фтору в питній воді. При цьому було встановлено зв'язок ($r_{xy} = -0.85$). Оцініть силу і напрям зв'язку.

Чи можна стверджувати, що при єдиному централізованому водопостачанні ця закономірність характерна для захворюваності карієсом дітей всього міста?

Чи є умова завдання достатнім для такого твердження?

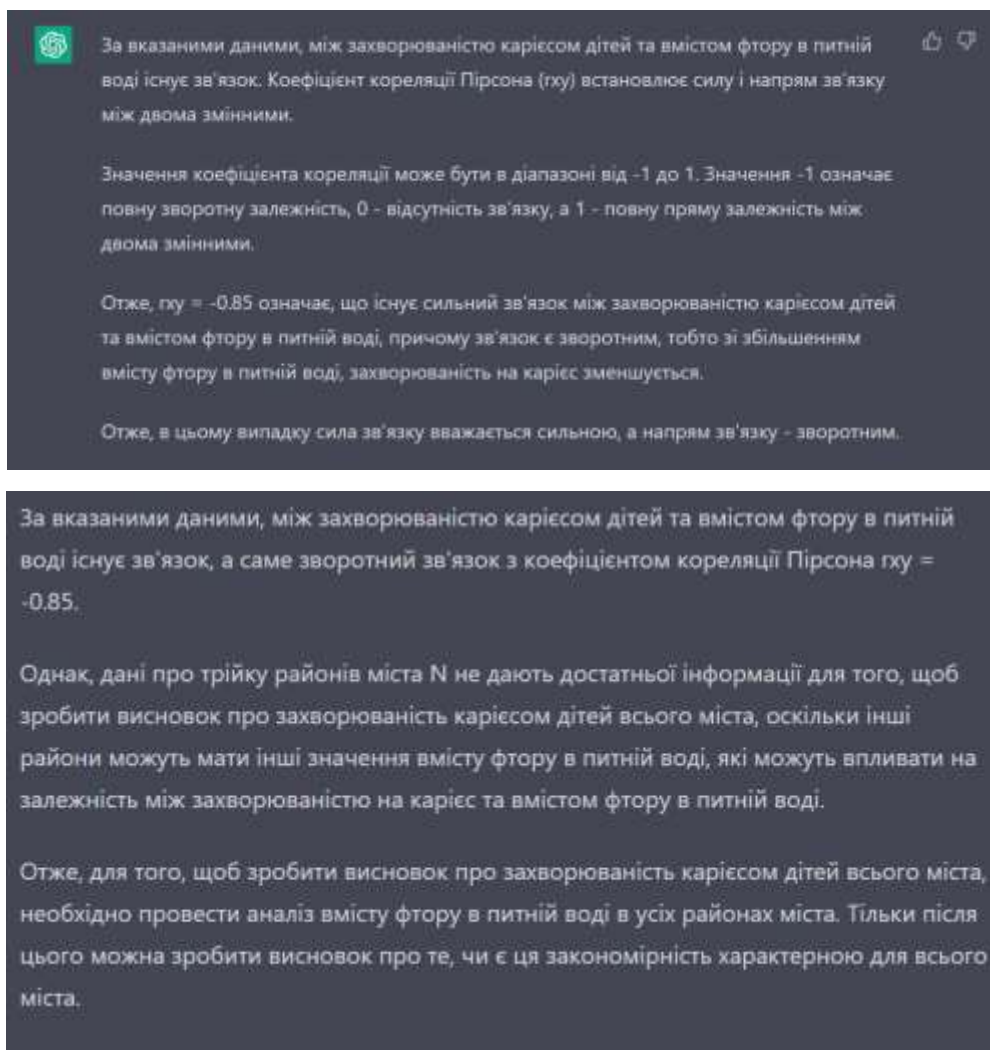


Рис.9.1. Відповідь від ChatGPT.

Відповідь: Коефіцієнт ближче до -1, це вказує на наявність сильної суворой зворотньої лінійної залежності між карієсом та вмістом фтору у воді. Але за цими даними неможливо сказати про захворюваність по всьому місту. Навіть з централізованим постачанням води. Труби та інші фактори також можуть впливати на результат.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						10
Змн.	Арк.	№ докум.	Підпис	Дата		

Задача 10.

Необхідно встановити наявність кореляційної зв'язку рангових методом між робочим стажем і показником травматизму за даними:

Робочий стаж в роках	Травматизм на 100 працюючих
до 1 року	24
1-2	16
3-4	12
5-6	12
7 і більше	6

3/3

Для встановлення наявності кореляційної зв'язку між ранговими змінними робочим стажем і показником травматизму можна використовувати кореляційний коефіцієнт Спірмена.

Спочатку необхідно присвоїти кожному значенню робочого стажу його ранг (тобто перетворити кількісні дані на рангові):

Робочий стаж в роках	Ранг
до 1 року	1
1-2	2
3-4	3
5-6	4
7 і більше	5

Потім потрібно зробити те саме для показника травматизму:

Травматизм на 100 працюючих	Ранг
24	5
16	4
12	2.5
12	2.5
6	1

Regenerate response

Звертаємо увагу, що для однакових значень показника травматизму були присвоєні середні ранги.

Після цього можна обчислити ранговий коефіцієнт кореляції Спірмена за формулою:

$$r_s = 1 - (6 * \sum d^2) / (n * (n^2 - 1))$$

де d - різниця рангів відповідних значень двох змінних, n - кількість спостережень.

Підставляємо значення і отримуємо:

$$r_s = 1 - (6 * ((0-4)^2 + (1-3)^2 + (2.5-2.5)^2 + (2.5-2.5)^2 + (4-1)^2)) / (5 * (5^2 - 1)) = 1 - (6 * 35) / (5 * 24) = 1 - 1.75 = 0.25$$

Таким чином, отримуємо значення рангового коефіцієнта кореляції Спірмена $r_s = 0.25$, що свідчить про досить слабку кореляційну залежність між робочим стажем і показником травматизму.

Рис.10.1. Відповідь від ChatGPT.

Відповідь: Коефіцієнт ближче до 0.2, це вказує на наявність слабкої суворі прямої лінійної залежності методом між робочим стажем і показником травматизму.

Висновки: Виконавши дану лабораторну роботу ми вивчили засоби кореляційного аналізу пакету **Excel**. Провесли лінійний і множинний кореляційний аналіз.

		Дяченко В.В.			ДУ «Житомирська політехніка».23.122.08.000 – Пр2	Арк.
						12
Змн.	Арк.	№ докум.	Підпис	Дата		