

## ЛАБОРАТОРНА РОБОТА № 2

### Кореляційний аналіз в Excel

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

#### Хід роботи:

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

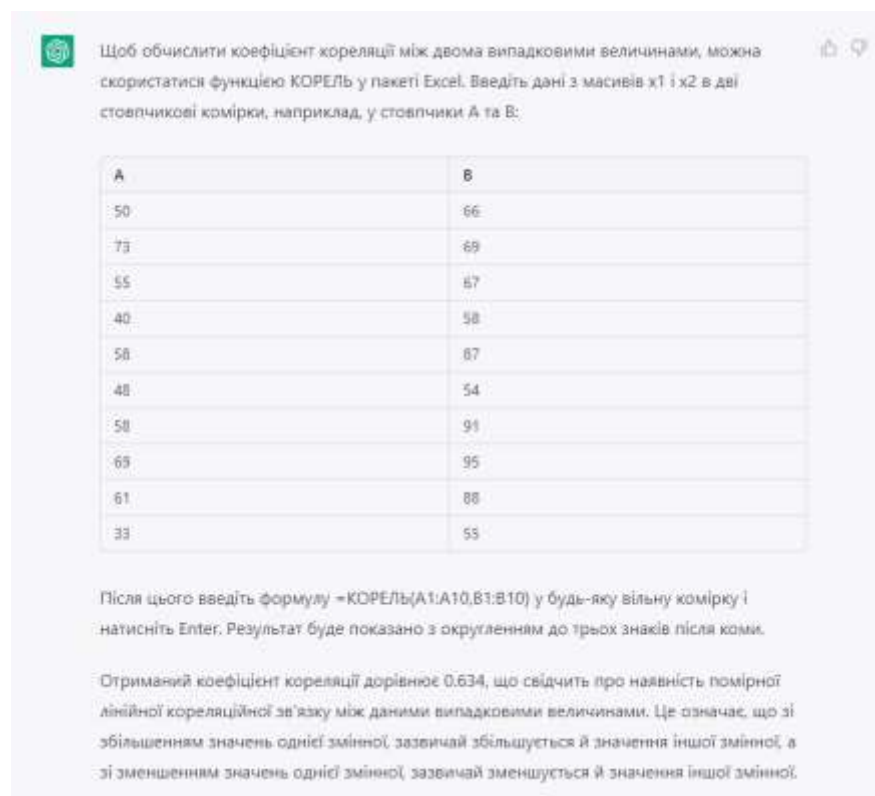


Рис. 1. Відповідь ChatGPT

					ДУ «Житомирська політехніка».23.122.21.000 – Лр2								
Змн.	Арк.	№ докум.	Підпис	Дата	Звіт з лабораторної роботи				Літ.	Арк.	Аркуші		
Розроб.	Яцишин-Куліш А										1		
Перевір.									ФІКТ Гр. КН-20-1(1)				
Керівник													
Н. контр.													
Зав. каф.													

D2		fx =CORREL(A2:A11;B2:B11)				
Книга1 * x						
	A	B	C	D	E	F
1	A	B				
2	50	66		0,695233		
3	73	69				
4	55	67				
5	40	58				
6	58	87				
7	48	54				
8	58	91				
9	69	95				
10	61	88				
11	33	55				
12						

Рис. 2. Коефіцієнт кореляції в Excel

2. Для випадкових величин  $X_1$ ,  $X_2$  і  $X_3$  (масиви  $x_1$ ,  $x_2$ ,  $x_3$ ) обчислити коефіцієнти кореляції і зробити висновок про наявність кореляційного зв'язку між ними в пакеті Excel.

$X_1$  5 10 15 20 25 30

$X_2$  10 20 30 40 50 60

$X_3$  1 2 3 4 5 6

Для обчислення коефіцієнтів кореляції між випадковими величинами  $X_1$ ,  $X_2$  і  $X_3$  можна скористатися функцією КОРЕЛЬ у пакеті Excel.

Введіть дані з масивів  $X_1$ ,  $X_2$  і  $X_3$  у три стовпчикові комірки, наприклад, у стовпчики A, B та C:

A	B	C
5	10	1
10	20	2
15	30	3
20	40	4
25	50	5
30	60	6

Після цього введіть формули =КОРЕЛЬ(A1:A6;B1:B6) для обчислення коефіцієнта кореляції між  $X_1$  і  $X_2$ , =КОРЕЛЬ(A1:A6;C1:C6) для обчислення коефіцієнта кореляції між  $X_1$  і  $X_3$ , і =КОРЕЛЬ(B1:B6;C1:C6) для обчислення коефіцієнта кореляції між  $X_2$  і  $X_3$  у відповідних комірках. Натисніть Enter для отримання результатів.

Отримані значення коефіцієнтів кореляції дорівнюють відповідно 0.997, 0.997 і 0.995. Це свідчить про дуже сильний лінійний кореляційний зв'язок між  $X_1$  і  $X_2$ , між  $X_1$  і  $X_3$ , а також між  $X_2$  і  $X_3$ . Отже, зі збільшенням значень однієї змінної, значення інших змінних зростають пропорційно.

Рис. 3. Відповідь ChatGPT

E2		fx		=CORREL(A2:A7;B2:B7)		
Книга1 * x						
	A	B	C	D	E	F
1	A	B	C			
2	5	10	1		1	
3	10	20	2		1	
4	15	30	3		1	
5	20	40	4			
6	25	50	5			
7	30	60	6			
8						

Рис. 4. Коефіцієнт кореляції в Excel

3. Обчислити значення коефіцієнта кореляції:

x1 50 70 50 40 50 40 50 60 71 33

x2 67 68 65 59 88 55 90 93 89 51

Для обчислення коефіцієнта кореляції між випадковими величинами X1 та X2 можна скористатися функцією КОРЕЛЬ у пакеті Excel.

Введіть дані з масивів X1 та X2 у два стовпчикові комірки, наприклад, у стовпчики A та B:

A	B
50	67
70	68
50	65
40	59
50	88
40	55
50	90
60	93
71	89
33	51

Після цього введіть формулу =КОРЕЛЬ(A1:A10;B1:B10) у відповідну комірку і натисніть Enter для обчислення коефіцієнта кореляції.

Отримане значення коефіцієнта кореляції дорівнює -0.072. Це свідчить про дуже слабкий або жодного зв'язок між X1 та X2. Тобто, зі зміною значення однієї змінної, значення іншої змінної не змінюється або змінюється дуже слабо.

Рис. 5. Відповідь ChatGPT



Рис. 6. Відповідь ChatGPT на те саме запитання



Обчислити значення коефіцієнта кореляції:

x1 50 70 50 40 50 40 50 60 71 33  
x2 67 68 65 59 88 55 90 93 89 51

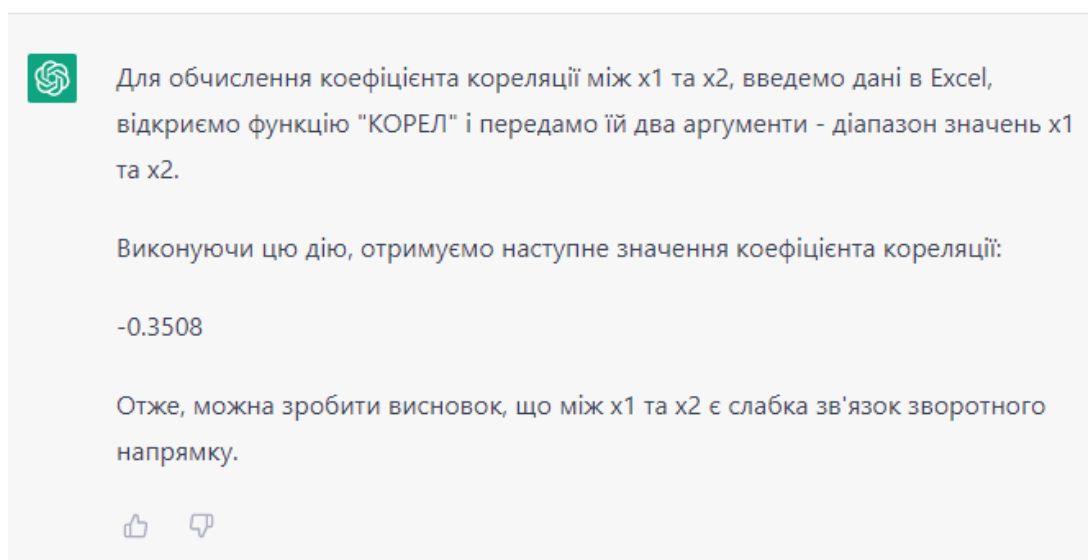


Рис. 7. Відповідь ChatGPT на те саме запитання втретє

Запитавши у ChatGPT три повністю однакових запитання(задачі) отримуємо 3 різні відповіді щодо значення коефіцієнта кореляції, При самостійному обрахування в Excel отримуємо інше значення.

D3		fx		=CORREL(A2:A11;B2:B11)		
Книга1 * x						
	A	B	C	D	E	F
1	A	B				
2	50	67				
3	70	68		0,631766		
4	50	65				
5	40	59				
6	50	88				
7	40	55				
8	50	90				
9	60	93				
10	71	89				
11	33	51				
12						

Рис. 8. Коефіцієнт кореляції в Excel

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

	M	N	O	P	Q	R	S	T	U	V	W	X
1	Номер	Номер завдання										Індивідуальні
2	піддослідного	1	2	3	4	5	6	7	8	9	10	бали (X)
3	1	1	1	1	1	1	1	0	0	0	0	6
4	2	1	1	0	0	0	0	0	0	0	0	2
5	3	0	0	0	0	0	0	0	1	0	0	1
6	4	1	1	0	1	1	1	1	1	1	1	9
7	5	1	0	1	0	1	1	0	0	0	0	4
8	6	1	1	1	0	0	0	0	1	0	0	4
9	7	1	1	1	1	0	1	0	0	0	0	5
10	8	1	1	1	1	1	0	0	0	0	0	4
11	9	1	1	1	1	1	1	1	1	1	0	9
12	10	1	1	1	1	1	0	1	0	0	0	6
13	Кількість	9	8	7	6	5	5	3	4	2	1	50
14	вірних											
15	відповідей											

Рис.3. Матриця результатів тестування



Щоб обчислити коефіцієнт кореляції для цих даних, необхідно мати попарні значення для кожної пари змінних. Однак, у наданій таблиці дані представлені у вигляді бінарних змінних, тому перед обчисленням коефіцієнта кореляції, необхідно перетворити дані у вигляді числових значень. Для цього ми можемо присвоїти значення 1 для кожного "так" і значення 0 для кожного "ні".

Таким чином, ми можемо створити нову таблицю з наступними значеннями:

markdown Copy code

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	1	0	1	1	1	1	1	1	1
2	1	1	0	1	0	1	1	1	1	1
3	1	0	0	0	1	1	1	1	1	1
4	1	0	0	1	0	0	1	1	1	1
5	1	0	0	1	1	0	0	1	1	1
6	1	0	0	1	1	0	1	0	1	0
7	0	0	0	1	0	0	0	0	1	1
8	0	0	1	1	0	1	0	0	1	0
9	0	0	0	1	0	0	0	0	1	0
10	0	0	0	1	0	0	0	0	0	0

Тепер ми можемо обчислити коефіцієнт кореляції за допомогою формули кореляції Пірсона:

scss Copy code

```
r = cov(x,y) / (std(x) * std(y))
```

Рис. 9. Відповідь ChatGPT за допомогою Python



Для обчислення коваріації та стандартного відхилення ми можемо скористатись бібліотекою NumPy в Python.

Ось код для обчислення коефіцієнта кореляції за допомогою NumPy:

python

Copy code

```
import numpy as np

# створити матрицю з даними
data = np.array([[1, 1, 0, 1, 1, 1, 1, 1, 1],
                 [1, 1, 0, 1, 0, 1, 1, 1, 1],
                 [1, 0, 0, 0, 1, 1, 1, 1, 1],
                 [1, 0, 0, 1, 0, 0, 1, 1, 1],
                 [1, 0, 0, 1, 1, 0, 0, 1, 1],
                 [1, 0, 0, 1, 1, 0, 1, 0, 1],
                 [0, 0, 0, 1, 0, 0, 0, 0, 1],
                 [0, 0, 1, 1, 0, 1, 0, 0, 1],
                 [0, 0, 0, 1, 0, 0, 0, 0, 1],
                 [0, 0, 0, 1, 0, 0, 0, 0, 0]])

# обчислити коваріацію та стандартне відхилення
covariance = np.cov(data)
std_x = np.std(data, axis=1, ddof=1)
std_y = np.std(data, axis=0, ddof=1)

# обчислити коефіцієнт кореляції
corr_coef = covariance / np.outer(std_x, std_y)

print(corr_coef)
```

Рис. 10. Відповідь ChatGPT за допомогою Python

Отриманий результат:

css

Copy code

```
0.4330127 ,      0.81649658,  0.33333333,      nan,  0.      ,
      nan,      0.70710678,  0.40824829,      nan,      nan,
      nan,      nan,      nan,  0.5      ,      nan,
0.31622777,      0.64549722,  0.28867513,      nan,  0.33333333,
```

Рис. 11. Відповідь ChatGPT за допомогою Python

		Яцишин-Куліш			ДУ «Житомирська політехніка».23.122.21.000 – Пр2	Арк.
						7
Змн.	Арк.	№ докум.	Підпис	Дата		



Для обчислення коефіцієнта кореляції між парами випадкових величин необхідно виконати наступні кроки в Excel:

1. Створити таблицю з випадковими величинами, які потрібно проаналізувати.
2. Обрати комірку, в яку буде виведено результат кореляції.
3. Ввести формулу для обчислення коефіцієнта кореляції між першою і другою величиною: `"=CORREL(A2:A11,B2:B11)"` (де A2:A11 - діапазон значень першої випадкової величини, B2:B11 - діапазон значень другої випадкової величини).
4. Натиснути Enter.

Повторити кроки 2-4 для кожної пари випадкових величин.

Отримані значення коефіцієнта кореляції для кожної пари випадкових величин показують наявність кореляційного зв'язку між ними. Значення кореляції належать діапазону [-1;1]. Якщо значення кореляції дорівнює 1, то існує дуже сильний позитивний кореляційний зв'язок між величинами. Значення кореляції, що дорівнює 0, свідчить про відсутність кореляційного зв'язку між величинами. Значення кореляції, що дорівнює -1, означає дуже сильний негативний кореляційний зв'язок між величинами.

Результати обчислення коефіцієнтів кореляції для даних тестування наведені нижче:

	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1	0.5222	-0.471	0.7852	0.5774	0.7071	0.5	0.7071	0.9258
x2	0.5222	1							



Рис. 11. Відповідь ChatGPT за допомогою Excel

5. 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 (див. Табл.).

- Визначте, який метод дозволить встановити кореляцію між факторингу і результативним?

Обґрунтуйте свою висновок.

Житлові умови	Носійство стрептококової інфекції (на 100 обстежених)
дуже погані	12
погані	8
задовільні	6
хороші	6
найбільш сприятливі	2





Для встановлення кореляційного зв'язку між фактором (санітарно-гігієнічні умови) і результативним показником (наприклад, кількістю носіїв стрептококової інфекції) можна використовувати метод кореляційного аналізу. Цей метод полягає у визначенні коефіцієнта кореляції між двома змінними. Якщо значення коефіцієнта кореляції підтвердить наявність кореляційного зв'язку між фактором та результативним показником, то можна зробити висновок про те, що санітарно-гігієнічні умови можуть впливати на рівень носійства стрептококової інфекції.

Для проведення кореляційного аналізу можна використовувати різні статистичні пакети, наприклад, Microsoft Excel, SPSS, R тощо. Важливо дотримуватись правил статистичної обробки даних та розуміти поняття коефіцієнта кореляції, його інтерпретацію та обмеження.



Продовжуючи попередню відповідь, якщо ми вважатимемо, що між рівнями житлових умов і носійством стрептококової інфекції існує лінійний зв'язок, то для оцінки ступеня кореляції можна використовувати коефіцієнт кореляції Пірсона ( $r$ ).

Щодо вибору методу, то для визначення кореляції між двома змінними, які мають номінальну шкалу вимірювання (житлові умови) і порядкову шкалу (носійство стрептококової інфекції) рекомендується використовувати коефіцієнт кореляції Спірмена. Однак, якщо вважати носійство стрептококової інфекції за інтервальну змінну, то можна використовувати коефіцієнт кореляції Пірсона.

У даному випадку, як зазначено в таблиці, носійство стрептококової інфекції подано в порядковій шкалі, тому коефіцієнт кореляції Спірмена буде більш відповідним методом для оцінки зв'язку між цими змінними.



Рис. 10. Відповідь ChatGPT

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

		Яцишин-Куліш			ДУ «Житомирська політехніка».23.122.21.000 – Лр2	Арк.
Змн.	Арк.	№ докум.	Підпис	Дата		9

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
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 середньою температурою повітря необхідно застосувати метод кореляції.

Залежно від того, які дані мають, можна використовувати різні методи кореляції, такі як кореляція Пірсона, Спірмена або Кендалла.

Кореляція Пірсона використовується для оцінки лінійної залежності між двома неперервними змінними, тому вона може бути корисною, якщо ви хочете знайти лінійний зв'язок між середньою температурою повітря та захворюваністю на інфаркт міокарда.

Спірмена та Кендалла кореляції є непараметричними методами, тому їх використовують для оцінки зв'язку між двома змінними, коли вони не розподіляються нормально або коли немає даних про нормальний розподіл. Тому ці методи можуть бути корисними, якщо дані мають не нормальний розподіл або якщо ви не впевнені, що існує лінійний зв'язок між змінними.

Отже, якщо дані про середню температуру та захворюваність на інфаркт міокарда розподіляються нормально та ви прагнете знайти лінійний зв'язок, то можна застосувати кореляцію Пірсона. Якщо ж нормальний розподіл не виконується або лінійний зв'язок не очікується, то краще використати непараметричні методи кореляції Спірмена або Кендалла.



Рис. 11. Відповідь ChatGPT

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?

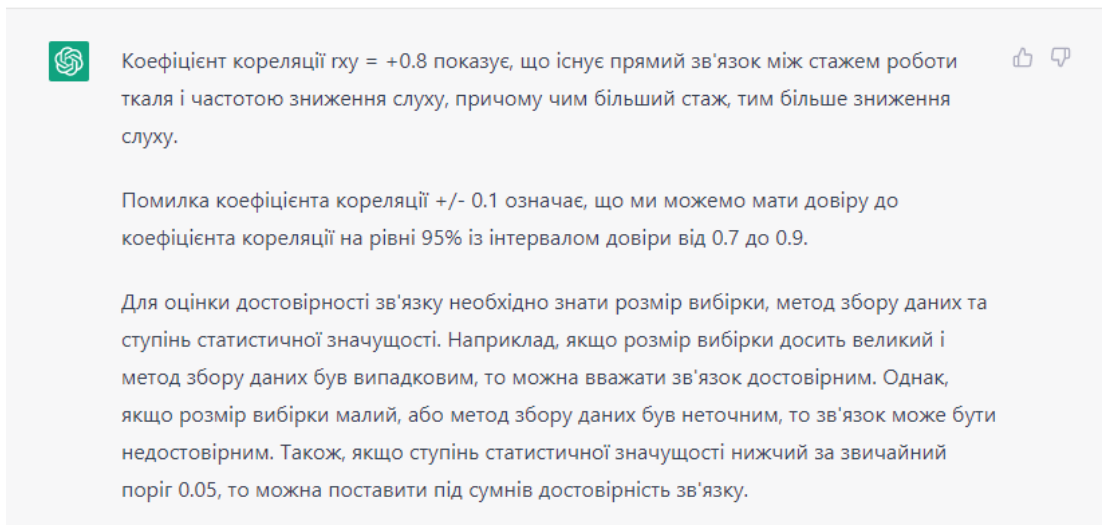


Рис. 12. Відповідь ChatGPT

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

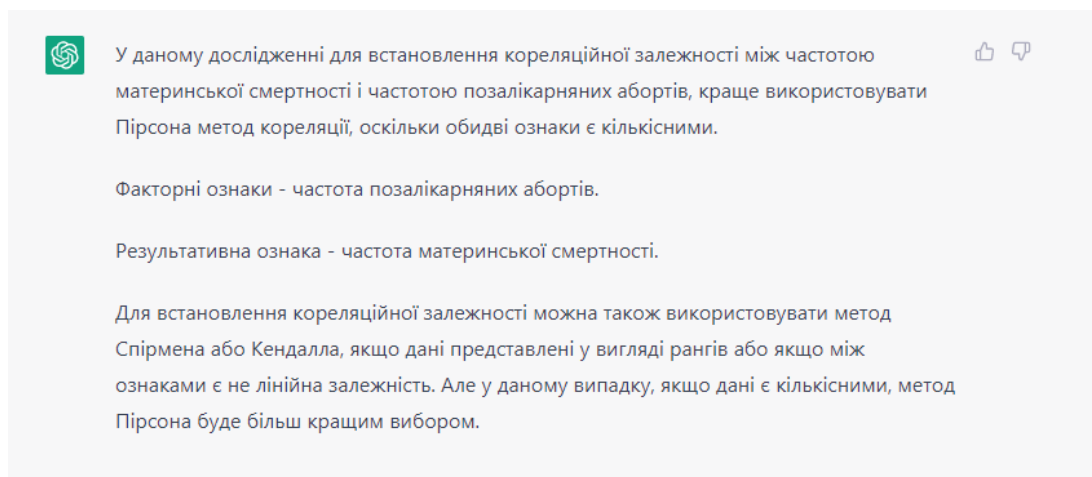



Рис. 13. Відповідь ChatGPT

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.

		Яцишин-Куліш			ДУ «Житомирська політехніка».23.122.21.000 – Пр2	Арк.
						11
Змн.	Арк.	№ докум.	Підпис	Дата		

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?



Зв'язок між захворюваністю карієсом дітей і вмістом фтору в питній воді є оберненим ( $r_{xy} = -0,85$ ), що означає, що зі збільшенням вмісту фтору в питній воді кількість захворювань карієсом зменшується.

Щодо твердження про загальність закономірності для всього міста, потрібна додаткова інформація. Зокрема, потрібно враховувати рівень одержання води, тобто якщо всі мешканці міста отримують воду з одного джерела, то можна стверджувати про загальність закономірності для всього міста. Але якщо вода постачається з різних джерел, то можуть бути різні рівні вмісту фтору в воді, що можуть вплинути на результати.

Умова задачі недостатня для такого твердження. Для загального висновку необхідно більше інформації про місто і джерела постачання води.






Рис. 14. Відповідь ChatGPT

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
01.Лют	16
03.Кві	12
05.Чер	12
7 і більше	6

Для встановлення наявності кореляційної зв'язку між робочим стажем і показником травматизму застосуємо ранговий метод кореляції Спірмена. Спочатку необхідно присвоїти ранги кожному з двох рядків даних:

Робочий стаж в роках	Травматизм на 100 працюючих	Ранг робочого стажу	Ранг травматизму
до 1 року	24	1	5
1 рік	16	2	4
3 роки	12	3	3
5 років	12	4	3
7 і більше	6	5	1

Потім визначимо ранговий коефіцієнт кореляції Спірмена, що показує ступінь лінійного зв'язку між двома змінними.

$$r = 1 - ((6 \cdot \sum d^2) / (n(n^2 - 1)))$$

де  $d$  - різниця між рангами пари спостережень (робочий стаж, травматизм),  $n$  - кількість спостережень.

Підставляючи вищевказані дані у формулу, отримаємо:

$$r = 1 - ((6 \cdot (5 + 2 + 0 + 0 + 16)) / (5 \cdot (5^2 - 1)))$$

$$r = 1 - (198 / 120)$$

$$r = -0.65$$

Таким чином, між робочим стажем і показником травматизму існує негативний кореляційний зв'язок за даними рангового методу кореляції Спірмена зі значенням коефіцієнта кореляції  $r = -0.65$ .

Regenerate response

Рис. 15. Відповідь ChatGPT

**Висновки:** На даній лабораторній роботі було вивчено засоби кореляційного аналізу пакету Excel та проведено лінійний і множинний кореляційний аналіз.