Лабораторна робота
Кореляційний аналіз експериментальних даних у середовищі Microsoft Excel

Мета роботи: навчитись визначати тип та тісноту взаємозв'язку між випадковими змінними.

Основні задачі роботи:

1. ознайомлення з поняттям кореляційного аналізу;
2. вивчення методів оцінки взаємозв'язку між змінними;
3. набуття практичних навичок виконання кореляційного аналізу в програмному середовищі Microsoft Excel;
4. формування навичок інтерпретації результатів статистичного аналізу.

Теоретичні відомості

Кореляція – це статистичний показник, який характеризує ступінь взаємозв'язку між двома змінними.

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

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

Наприклад: температура тіла та частота серцевих скорочень; артеріальний тиск та вік пацієнта; тривалість фізичного навантаження та пульс.

Кореляція дозволяє визначити:

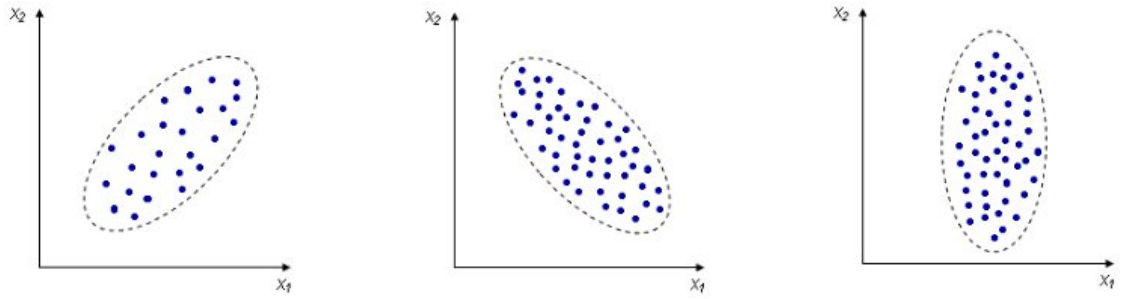
- чи існує зв'язок між величинами;
- напрямок цього зв'язку;
- силу взаємозалежності.

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. Істотні, в даному аспекті, фактори використовують далі в регресійному аналізі;
- виявлення невідомих причинних зв'язків.

Існують різні види зв'язку між змінними:

1. Прямий причинно-наслідковий зв'язок (рис. 1.1, а).
2. Зворотній причинно-наслідковий зв'язок (рис. 1.1, б).
3. Зв'язок викликаний однією або декількома прихованими змінними.
4. Зв'язку немає, залежність, що спостерігається випадкова (рис. 1, в).



а) прямий зв’язок б) зворотній зв’язок в) відсутність зв’язку

Рис. 1.1. Варіанти взаємозв’язку між випадковими змінними

Взаємозв’язок між змінними чисельно характеризується за допомогою коефіцієнту кореляції *r*. Коефіцієнт *r* є випадковою величиною, оскільки обчислюється з випадкових величин. Це лінійний коефіцієнт кореляції, який показує *лінійний* взаємозв’язок між двома змінними і коливається в межах від -1 до 1 (табл. 1.1). За відсутності лінійного зв’язку значення *r* буде близьким до 0.

Таблиця 1.1

Лінійний коефіцієнт кореляції

Значення *r*	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

Знак коефіцієнта:

r > 0 – пряма залежність;

r < 0 – обернена залежність.

Кореляційний аналіз може виконуватися з використанням методу Пірсона або рангового методу Спірмена.

Метод Пірсона застосуємо для розрахунків, які вимагають точного визначення сили, що існує між змінними. Досліджувані з його допомогою ознаки повинні виражатися тільки кількісно. Коефіцієнт кореляції обчислюється за формулою:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}. \qquad (1.1)$$

Коефіцієнт ***рангової кореляції Спірмена*** дозволяє статистично встановити наявність зв’язку між явищами. Його розрахунок передбачає встановлення для кожної ознаки порядкового номера – рангу. Ранг може бути

зростаючим або спадаючим. Для застосування методу Спірмена або рангової кореляції немає жорстких вимог у вираженні ознак – воно може бути, як кількісним, так і атрибутивним (якісним). Даний метод не встановлює точну силу зв'язку і має орієнтовний характер:

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad r = 1 - \frac{6\sum d^2}{n(n^2-1)}, \tag{1.2}$$

де n – кількість ранжованих ознак;

d – різниця між рангами за двома змінними;

Σ (d^2) – сума квадратів різниць рангів.

Оцінка значущості коефіцієнту кореляції відбувається шляхом розрахунку значення *p*, ґрунтуючись на перевірках двох гіпотез:

Основна гіпотеза H0: ρ = 0;
Альтернативна гіпотезаH1: ρ ≠ 0.

Основна гіпотеза стверджує, що кореляції не існує між ознаками *x* та *y*в генеральній сукупності. Альтернативна гіпотеза стверджує, що кореляція між ознаками *x* та *y* генеральної сукупності значима. Коли основна гіпотеза відкидається на певному рівні значущості, це означає, що існує значуща відмінність між значенням *r* та 0. Коли основна гіпотеза приймається, це означає, що значення *r* не сильно відрізняється від 0 і є випадковим.

Завдання роботи

Виконати кореляційний аналіз між наборами експериментальних даних. Використати данні дата сету, що на минулій практиці.

1. Побудувати діаграми розсіювання між змінними за допомогою конструктора діаграм та зробити попередній висновок про наявність залежності.

Рис. 1.2. Конструктор діаграм

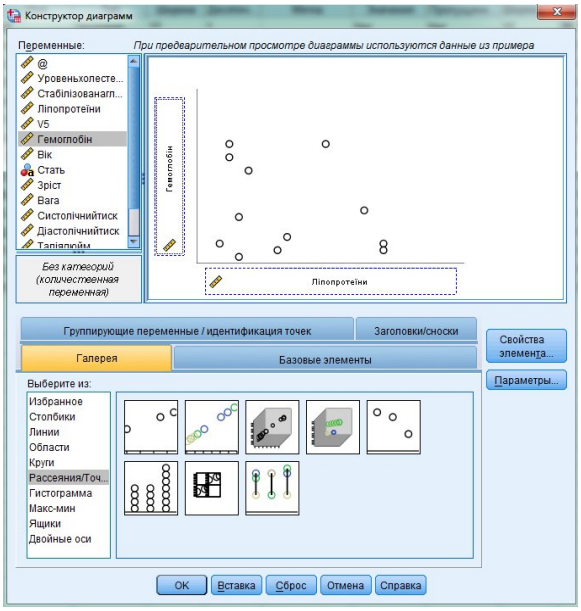


Рис. 1.3. Діалогове вікно «Конструктор діаграм»

2. Провести кореляційний аналіз даних.
Оберіть в меню: *Аналіз > Кореляції*

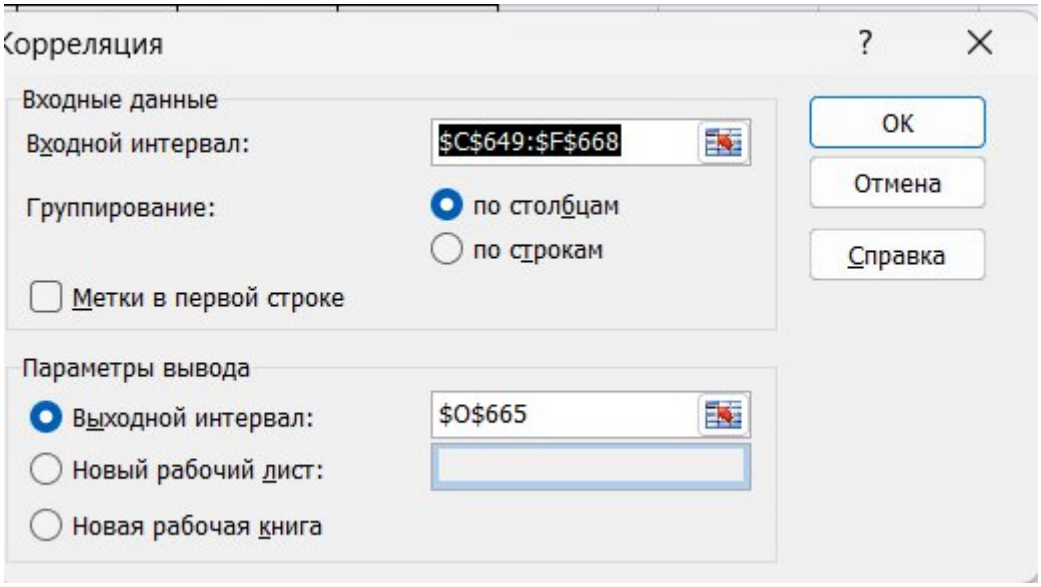


Рис. 1.4. Діалогове вікно «Кореляції»

	Столбец 1	*Столбец 2*	*Столбец 3*	*Столбец 4*
Столбец 1	1			
Столбец 2	0,356447	1		
Столбец 3	0,197891	0,05362	1	
Столбец 4	0,274176	0,351348	0,437913	1

Рис. 1.5. Результат Кореляції

Виберіть дві, або більше, числові змінні. Також доступні наступні параметри:

Коефіцієнти кореляції. Для кількісних нормально розподілених змінних виберіть коефіцієнт кореляції Пірсона. Якщо дані не розподілені нормально або

мають впорядковані категорії (є ранговими), виберіть *«Тау-b Кендалла»* або *Спірмена*, які вимірюють зв'язок між рангами.

Критерій значущості. Можна вибрати двосторонній або односторонній критерій. Якщо напрямок зв'язку відомо заздалегідь, виберіть *«Односторонній»*. В іншому випадку виберіть *«Двосторонній»*.

Відмітити значущі кореляції. Коефіцієнти кореляції, значимі на рівні 0.05, будуть позначені однією зірочкою, а значущі на рівні 0.01 – двома зірочками.

Параметри. Для кореляції Пірсона є можливість обрати один або обидва з наступних пунктів:

Середні значення і стандартні відхилення виводяться для кожної змінної, а також число спостережень без пропущених значень. Пропущені значення обробляються для кожної змінної окремо, незалежно від установки, обраної в панелі *«Пропущені значення»*.

Суми перехресних добутків відхилень і коваріацій виводяться для кожної пари змінних. Сума перехресних добутків відхилень дорівнює сумі добутків змінних, скоригованих за середнім. Це чисельник у формулі коефіцієнта кореляції Пірсона. *Коваріація* – це ненормована міра зв'язку між двома змінними, яка дорівнює сумі перехресних добутків відхилень, поділеній на *N-1*.

Пропущені значення. Ви можете вибрати один з наступних варіантів:

Виключати попарно спостереження з пропущеними значеннями однієї або обох змінних пари, для яких обчислюється коефіцієнт кореляції (виключаються з аналізу). Оскільки в обчисленнях кожного коефіцієнта беруть участь всі спостереження без пропущених значень для даної пари змінних, то в кожному обчисленні використовується максимум доступної інформації. Це може привести до того, що набір коефіцієнтів буде вирахувано для різного числа спостережень.

Виключити повністю спостереження з пропущеними значеннями. Для будь-якої змінної спостереження виключаються з обчислень всіх кореляцій.

Визначити залежність функцією CORREL.

У будь-якій вільній комірці ввести формулу: =CORREL(B2:B11;C2:C11. Натиснути Enter.

Приклад отриманого значення: **r = 0,98**

Інтерпретація: між параметрами існує дуже сильний прямий зв'язок.

Аналіз результатів

Після виконання аналізу необхідно визначити:

- чи існує кореляція;
- її напрямок;
- силу зв'язку;
- практичне значення отриманого результату.

Приклад висновку:

Отриманий коефіцієнт кореляції свідчить про сильну пряму залежність між температурою тіла та частотою серцевих скорочень.

Зміст звіту

Звіт повинен містити:

1. Назву роботи;
2. Мету;
3. Теоретичні відомості;
4. Таблицю даних;
5. Розрахунок;
6. Діаграму;
7. Аналіз;
8. Висновок.

Контрольні питання

1. Що таке кореляція?
2. Які значення може приймати коефіцієнт Пірсона?
3. Чим відрізняється пряма та обернена кореляція?
4. Для чого застосовується кореляційний аналіз?
5. Яка функція в Excel використовується для обчислення коефіцієнта кореляції?
6. Що означає значення r = 0?
7. Як інтерпретується r = -0,85?
8. Що показує коефіцієнт детермінації R^2?
9. Які задачі вирішує кореляційний аналіз?
10. Як знайти коефіцієнт кореляції?
11. Які типи коефіцієнтів кореляції існують?
12. Як проаналізувати значення коефіцієнта кореляції на значимість?