# Кореляційний аналіз

**Змістовий модуль 5**

# Вступ

Визначення кореляції та кореляційного аналізу

Значення кореляційного аналізу в біомедичних дослідженнях

Коефіцієнти кореляції (Пірсона, Спірмена, Кендалла)

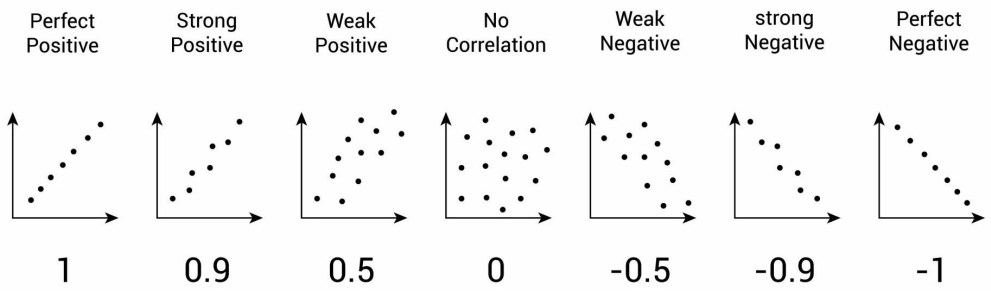
Візуалізація кореляційних зв’язків

Інтерпретація результатів аналізу та помилкові висновки

**Кореляційний аналіз** — це статистичний метод, що використовується для оцінки **сили** та **напряму** зв'язку між двома або більше кількісними змінними. Його мета полягає у виявленні того, чи існує статистично значуща асоціація між змінними, яка може бути як прямою (позитивною), так і оберненою (негативною).

В біомедичних дослідженнях кореляційний аналіз дозволяє, наприклад, досліджувати зв'язок між фізіологічними показниками або між факторами ризику та результатами захворювань. Важливо зазначити, що кореляція не обов’язково означає причинно-наслідковий зв'язок, а лише відображає рівень спільної варіації між змінними.

# Типи та сила кореляційних зв'язків




| Positive correlation | | Negative correlation | |
|---|---|---|---|
| $r = 1$ | perfect positive correlation | $r = -1$ | perfect negative correlation |
| $0.95 \leqslant r < 1$ | very strong positive correlation | $-1 < r \leqslant -0.95$ | very strong negative correlation |
| $0.87 \leqslant r < 0.95$ | strong positive correlation | $-0.95 < r \leqslant -0.87$ | strong negative correlation |
| $0.5 \leqslant r < 0.87$ | moderate positive correlation | $-0.87 < r \leqslant -0.5$ | moderate negative correlation |
| $0.1 \leqslant r < 0.5$ | weak positive correlation | $-0.5 < r \leqslant -0.1$ | weak negative correlation |
| $0 \leqslant r < 0.1$ | no correlation | $-0.1 < r \leqslant 0$ | no correlation |

# Приклади застосуваання кореляційного аналізу в біологічних дослідженнях

**Епідеміологія**: Аналіз зв’язку між рівнем забруднення повітря та частотою респіраторних захворювань. Кореляція допомагає зрозуміти, чи існує значуща асоціація між концентрацією шкідливих речовин у повітрі та рівнем захворюваності в різних регіонах.

**Генетика**: Дослідження взаємозв'язків між експресією генів і певними фізіологічними ознаками, наприклад, пошуком кореляції між експресією генів і рівнем холестерину. Це допомагає виявити гени, які можуть впливати на виникнення чи розвиток певних хвороб.

**Фармакологія**: Вивчення кореляції між дозою препарату та його ефективністю або частотою побічних ефектів. Наприклад, аналіз того, як різні дози лікарського засобу корелюють із показниками серцевого ритму або артеріального тиску у пацієнтів.

**Фізіологія**: Оцінка зв’язку між параметрами життєдіяльності, такими як частота серцевих скорочень та рівень споживання кисню ($VO_2$ max) у спортсменів. Це дозволяє визначити фізіологічні фактори, які корелюють з рівнем витривалості.

**Екологія**: Дослідження впливу кліматичних факторів на популяції рослин і тварин. Наприклад, оцінка кореляції між температурою середовища і кількістю певного виду комах у різні пори року для розуміння екологічних тенденцій.

# Методи обчислення кореляції

Коефіцієнт Пірсона: формула, переваги та недоліки

Коефіцієнт Спірмена та умови застосування

Коефіцієнт Кендалла

# Коефіцієнт кореляції Пірсона

Коефіцієнт кореляції Пірсона (rr) – це статистичний показник, що вимірює лінійний зв'язок між двома кількісними змінними. Він приймає значення від -1 до +1, де:

- r= +1 означає ідеальну пряму (позитивну) кореляцію,
- r= −1 означає ідеальну обернену (негативну) кореляцію,
- r=0 означає відсутність лінійного зв’язку між змінними.

Формула для обчислення коефіцієнта кореляції Пірсона:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

де:

$x_i$ і $y_i$ – значення змінних X та Y,

$\bar{x_i}$ і $\bar{y_i}$ – середні значення змінних X та Y,

чисельник представляє коваріацію між змінними X та Y, а знаменник – добуток стандартних відхилень цих змінних.

## Переваги коефіцієнта кореляції Пірсона

**Простота**: Легко обчислюється і зрозумілий для інтерпретації.

**Широке використання**: Застосовується в багатьох галузях, зокрема у біологічних та медичних дослідженнях.

**Інтуїтивна інтерпретація**: Дає змогу швидко оцінити силу та напрямок зв’язку між змінними.

## Недоліки коефіцієнта кореляції Пірсона

**Чутливість до аномальних даних**: Викиди можуть суттєво вплинути на значення коефіцієнта, що призводить до викривлення результатів.

**Лінійність зв’язку**: коефіцієнт Пірсона оцінює тільки лінійні зв'язки. Якщо зв'язок між змінними є нелінійним,обрахований коефіцієнт кореляції Пірсона може бути невірним.

**Не визначає причинно-наслідковий зв'язок**: Висока кореляція не означає, що одна змінна є причиною змін іншої, що потребує додаткових аналізів для підтвердження каузальності.

# Коефіцієнт кореляції Спірмена

Коефіцієнт кореляції Спірмена (ρ або rsrs) – це непараметричний показник, що оцінює монотонний зв'язок між двома змінними. На відміну від коефіцієнта Пірсона, який оцінює лінійну залежність, коефіцієнт Спірмена використовується для оцінки зв’язків, які можуть бути нелінійними, але є монотонними.

Коефіцієнт Спірмена обчислюється за ранговими значеннями змінних X та Y. Формула для обчилення коефіцієнта кореляції Спірмена:

$$rs = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

де:

- $d_i$ – різниця між рангами для кожної пари значень $x_i$ і $y_i$,
- n – кількість спостережень.

Якщо є однакові значення (повтори), вони займають середнє значення їхніх позицій у ранжуванні.

# Умови застосування коефіцієнта кореляції Спірмена

**Монотонний зв'язок**: Підходить для даних, які мають монотонну залежність, тобто коли зі збільшенням значення однієї змінної значення іншої також збільшується або зменшується, але не обов’язково лінійно.

**Непараметричний аналіз**: Оскільки даний критерій базується на рангах, його можна використовувати для даних, що не відповідають нормальному розподілу. Це робить його застосування можливим для обробки змінних з великими викидами або таких, що мають асиметричний розподіл.

**Порядкові або кількісні дані**: Коефіцієнт Спірмена може використовуватися для порядкових змінних або кількісних даних, попередньо перетворених у порядкову шкалу.

# Приклади використання

- Аналіз зв’язку між порядковими змінними, наприклад, рівнем стресу (низький, середній, високий) та рівнем фізичної активності.
- Виявлення кореляції між показниками, які мають нелінійний, але монотонний зв’язок, наприклад вік і витривалість у спортсменів.
- Дослідження зв’язку між рівнем освіти і доходом, де дані можуть мати багато аномальних значень.

Таким чином, коефіцієнт кореляції Спірмена є гнучким інструментом для дослідження зв'язків, що не відповідають лінійній залежності або не є кількісними, але мають певний порядок.

# Коефіцієнт кореляції Кендалла

Коефіцієнт кореляції Кендалла (τ) – це непараметричний показник, який вимірює монотонний зв'язок між двома змінними. Він оцінює узгодженість (або ступінь відповідності) рангових порядків між двома змінними, тобто наскільки пари значень узгоджуються або не узгоджуються в послідовності ранжування.

Коефіцієнт кореляції Кендалла (τ) визначається як:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

де:

C – кількість пар, що узгоджуються, тобто мають однаковий порядок для обох змінних,

D – кількість пар, що не узгоджуються, тобто змінні мають різний порядок у ранжуванні,

n – кількість спостережень.

# Тлумачення та умови застосування

$\tau = +1$: Ідеальна пряма монотонна залежність, всі пари узгоджені.

$\tau = -1$: Ідеальна обернена монотонна залежність, всі пари не узгоджені.

$\tau = 0$: Відсутність монотонної залежності.

## Умови застосування

- **Монотонний зв'язок**: Коефіцієнт Кендалла використовується для даних, що мають монотонний зв'язок (не обов'язково лінійний).
- **Непараметричний характер даних**: Підходить для даних, що не відповідають нормальному розподілу і можуть мати викиди.
- **Порядкові або кількісні дані**: Може застосовуватися для порядкових даних або кількісних даних, перетворених у ранжовану шкалу.

## Переваги коефіцієнта кореляції Кендалла

**Чутливість до узгодженості**: Дає більш точну оцінку при малих вибірках і вказує на узгодженість порядків між змінними.

**Стійкість до викидів**: Менш чутливий до викидів, порівняно з коефіцієнтом Пірсона.

## Недоліки коефіцієнта кореляції Кендалла

**Важчий у розрахунку**: Для великих вибірок розрахунок **т** може бути доволі складним.

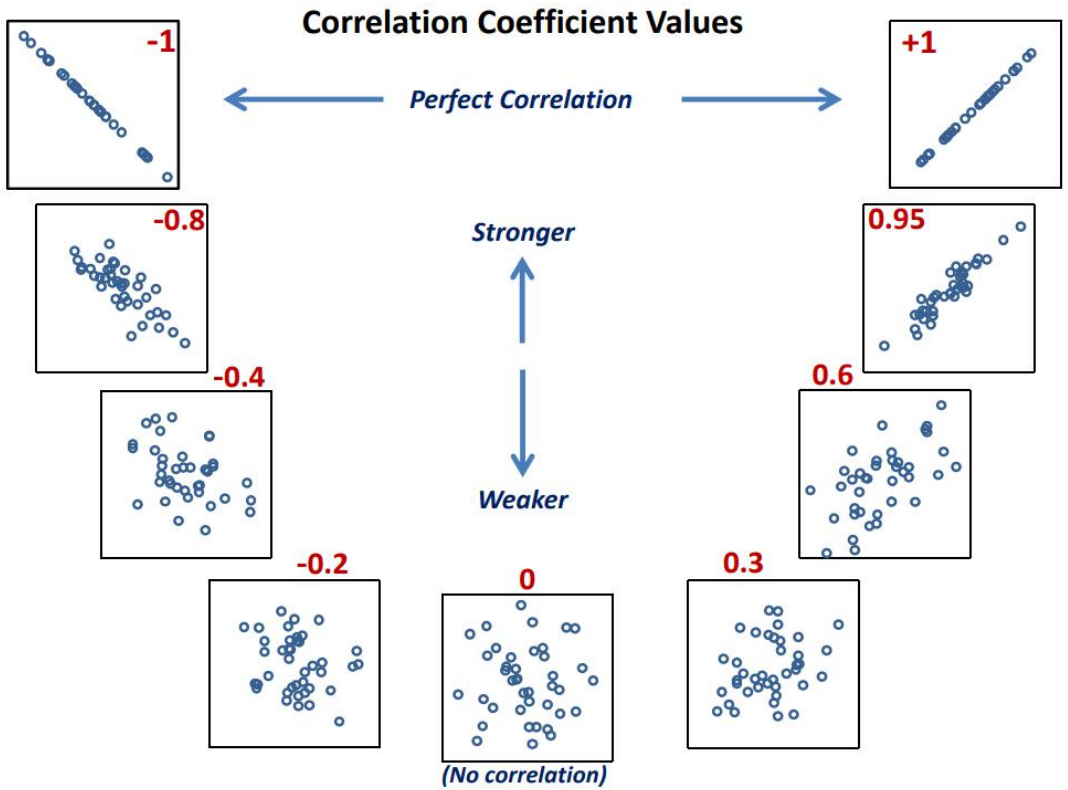
**Вимагає ранжування**: Потребує чіткого порядкового або рангового характеру даних для правильної інтерпретації.

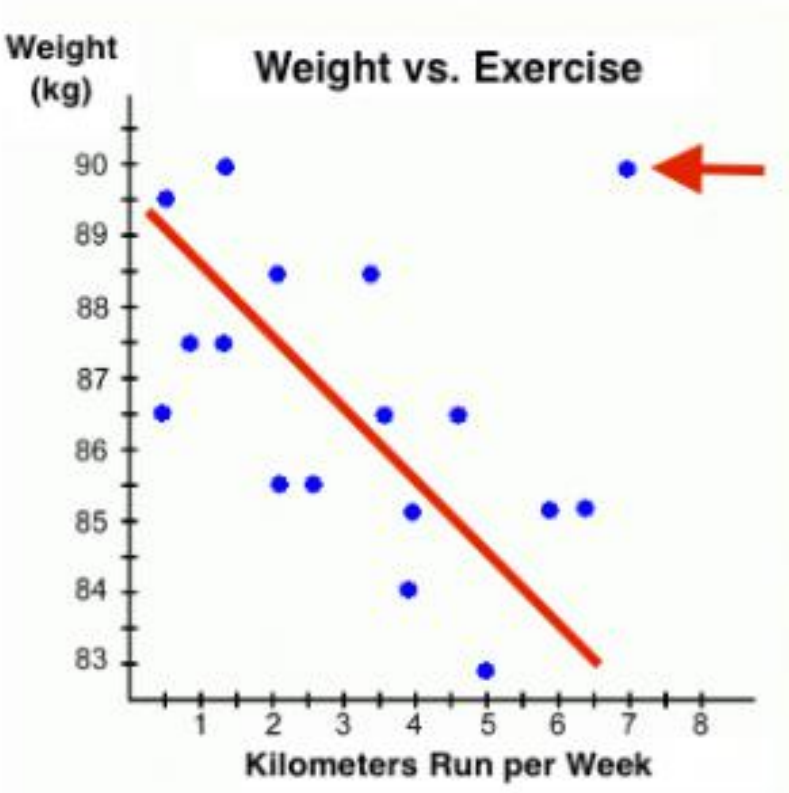
# Візуалізація кореляції

Способи графічного відображення: діаграми розсіювання, кореляційні матриці

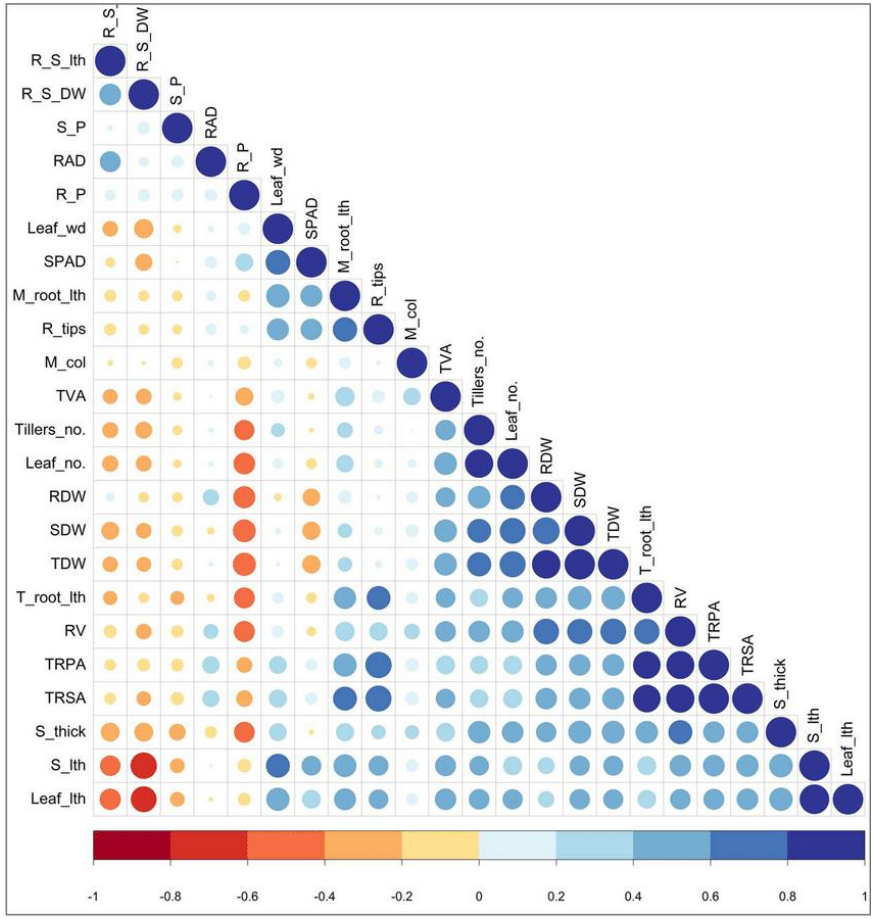
Використання теплових карт для представлення зв’язків

# Діаграми розсіювання (scatter plot)



# Кореляційні матриці (correlation matrix)

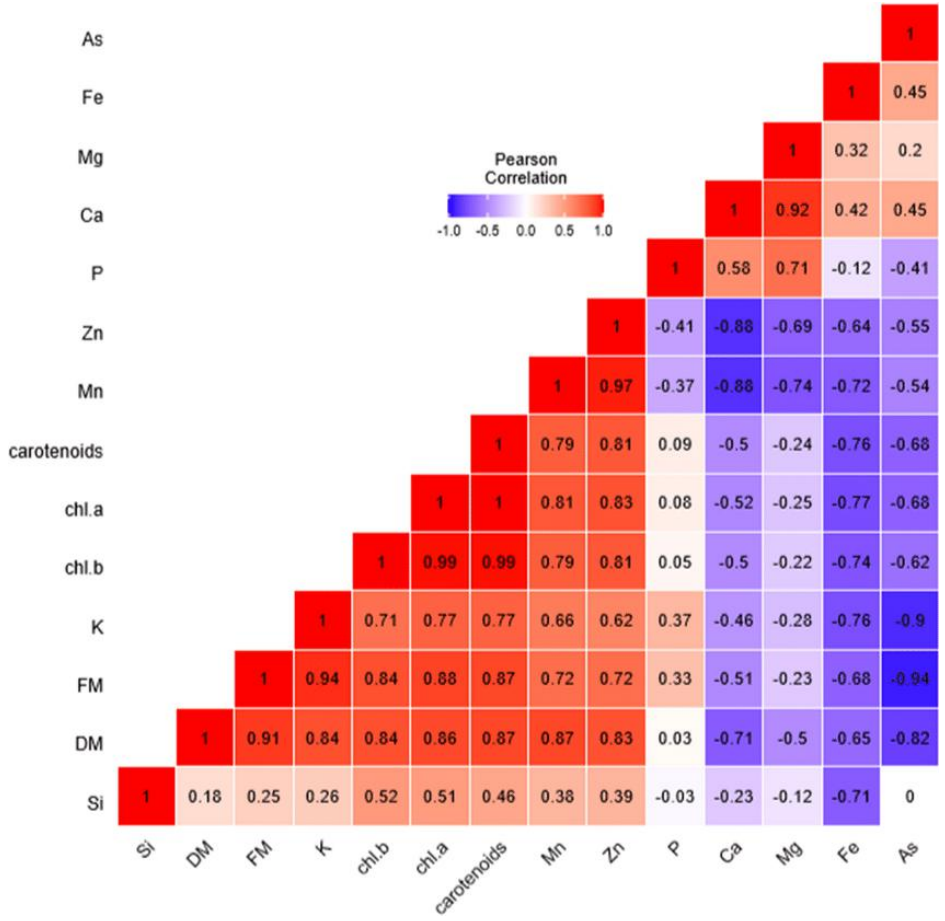



**Графік "Кореляційна матриця"** – це візуалізація кореляційної матриці, яка дозволяє легко оцінити силу і напрямок зв'язків між усіма змінними у наборі даних. Цей тип графіка зазвичай представлений у вигляді теплової карти або масиву кругів, кольорів чи числових значень, які позначають кореляційні коефіцієнти між змінними.

Змінні розміщуються вздовж горизонтальної і вертикальної осей, створюючи квадратну сітку. Кожна клітинка матриці представляє кореляцію між парою змінних.
Відображення коефіцієнтів має вигляд кругів чи квадратів різних розмірів. Чим більший або яскравіший символ, тим вищий рівень кореляції.

Додатково до кольорів або розмірів часто вказують числові значення кореляційних коефіцієнтів для точнішої інтерпретації.

# Теплові карти (heatmap)




**Властивості та особливості графіку**

- Симетрія: Як і сама кореляційна матриця, графік є симетричним щодо діагоналі.
- Діагональ: Коефіцієнти на діагоналі зазвичай позначаються як 1, оскільки це кореляція змінних із собою.
- Швидкий огляд зв'язків: Дає змогу одразу побачити, які змінні мають сильний позитивний чи негативний зв'язок, і визначити групи змінних з подібними характеристиками.

# Інтерпретація результатів

Як інтерпретувати значення коефіцієнтів кореляції

Ліміти та помилки при трактуванні даних

# Інтерпретація результатів

## 1. Сила зв’язку

**|r| = 0: Відсутність кореляції. Змінні не пов'язані лінійним чином.**

**0 < |r| < 0.3: Слабка кореляція. Зв’язок є, але незначний.**

**0.3 ≤ |r| < 0.7: Помірна кореляція. Зв’язок помітний, але не надто сильний.**

**0.7 ≤ |r| < 1: Сильна кореляція. Змінні мають чіткий зв’язок.**

**|r| = 1: Ідеальна кореляція. Зв’язок максимальний: позитивний або негативний.**

## 2. Напрямок зв’язку

r > 0: Позитивна кореляція. Коли значення однієї змінної збільшується, значення іншої також зростає. Наприклад, зростання температури може корелювати із зростанням швидкості реакції.

r < 0: Негативна кореляція. Коли значення однієї змінної збільшується, значення іншої зменшується. Наприклад, зі збільшенням фізичної активності може знижуватися частота серцевих захворювань.

**3. Обачність із причиново-наслідковими висновками**: Кореляція не вказує на причинність, тобто зв’язок між змінними ще не означає, що одна з них викликає зміни в іншій.

**4. Потрібно враховувати контекст дослідження**: Сила кореляції, яку вважають значущою, залежить від певної галузі. У соціальних науках слабка кореляція може бути прийнятною, а в фізиці або біології зазвичай очікують вищих коефіцієнтів.

**5. Врахування вибірки та змінних**: Завжди потрібно оцінювати кореляцію у контексті наявних даних і особливостей вибірки, оскільки аномальні дані можуть спотворити результати.

# Приклади можливих хибних інтерпретацій

**1. Сплутування кореляції з причинністю**:

**Невірна інтерпретація**: Якщо дві змінні корелюють, це означає, що одна змінна викликає зміну іншої.

**Приклад**: Виявлено кореляцію між кількістю морозива, проданого за день, та кількістю випадків утоплення. Однак це не означає, що продаж морозива спричиняє утоплення. Скоріш за все, обидва явища пов'язані із третьою змінною – температурою повітря.

**2. Ігнорування впливу сторонніх змінних (сплутуюча змінна):**

**Невірна інтерпретація**: Кореляція між двома змінними обов'язково означає прямий зв'язок між ними.

**Приклад**: Існує кореляція між рівнем стресу і випадками серцевих захворювань. Однак, крім стресу, є інші чинники, такі як генетика, спосіб життя, дієта, які також впливають на серцеві захворювання.

**3 Кореляція, спричинена мультиколінеарністю:**

**Невірна інтерпретація**: **Якщо змінні мають високу кореляцію, то їхній взаємозв'язок справжній**.

**Приклад**: У регресійному аналізі виявлено високу кореляцію між кількістю годин навчання та середньою успішністю. Однак при цьому корелюють ще й кількість годин сну та навчання, що може спричиняти мультиколінеарність, а отже, й викривлення реальної залежності.

**Неадекватне масштабування даних (наприклад, кореляція між середніми значеннями)**:

**Невірна інтерпретація**: Використання середніх значень змінних для розрахунку кореляції дає справжнє уявлення про зв'язок.

**Приклад**: При порівнянні успішності учнів у різних школах виявлено кореляцію між середнім балом з математики і середнім балом з фізики в кожній школі. Однак це може викривити результати, оскільки кореляція між середніми значеннями є більш високою, ніж між індивідуальними даними.