СТАТИСТИЧНІ МЕТОДИ ВИВЧЕННЯ ВЗАЄМОЗВ’ЯЗКІВ

Поняття про кореляцію

Кореляція - залежність між випадковими величинами, що не має строго функціонального характеру, за якої зміна однієї з випадкових величин приводить до зміни математичного сподівання іншої

При вивченні взаємозв'язку і взаємодії між явищами одні ознаки виступають як **фактори**, що зумовлюють зміни інших ознак. Ознаки цієї першої групи називають **ознаками-факторами (факторними ознаками)**; а ознаки, які є результатом впливу всіх цих факторів, називають **результативними ознаками**. В процесі дослідження необхідно встановити, яка з ознак є факторною, а якою є результативною

Приклади

У галузі комп'ютерних наук вибір факторних та результативних ознак є критичним етапом при побудові моделей машинного навчання, тестуванні продуктивності та управлінні IT-проєктами. Відповідно до принципів статистичного вивчення взаємозв'язків, **факторні ознаки** (X) - це причини або умови, а **результативні** (Y) - це наслідки, що змінюються під їхнім впливом

1. Тестування продуктивності. При аналізі роботи сервера важливо розуміти, що навантаження зазвичай є першопричиною технічних затримок.

Факторна ознака (X): Кількість одночасних запитів до сервера

Результативна ознака (Y): Час відгуку системи

Саме зростання кількості запитів (X) зумовлює збільшення затримок (Y), а не навпаки

2. Тестування систем на відмовостійкість

У процесі стрес-тестування розподілених систем (наприклад, мікросервісів) аналізують, як технічні обмеження впливають на стабільність платформи

Факторна ознака (X): Відсоток втрати пакетів у мережі. Ця ознака є зовнішнім фактором або умовою середовища, яка контролюється (штучно створюється) під час експерименту

Результативна ознака (Y): Кількість невдалих транзакцій. Ця ознака виникає як прямий наслідок нестабільного зв'язку; вона демонструє реакцію системи на фактор впливу

Нестабільність мережевого з'єднання (X) виступає **фактором**, що зумовлює зниження якості обслуговування. **Результатом** (Y) є зростання кількості помилок, які фіксує моніторингова система. Саме погіршення зв'язку спричиняє помилки, а не навпаки, що дозволяє чітко розмежувати причину та наслідок

3. Навчання нейронних мереж

При налаштуванні моделей штучного інтелекту гіперпараметри виступають факторами впливу

Факторна ознака (X): Швидкість навчання

Результативна ознака (Y): Точність моделі або значення функції втрат

Зміна швидкості навчання (X) безпосередньо зумовлює те, наскільки швидко та якісно модель зійдеться до оптимального результату (Y)

Функціональна залежність

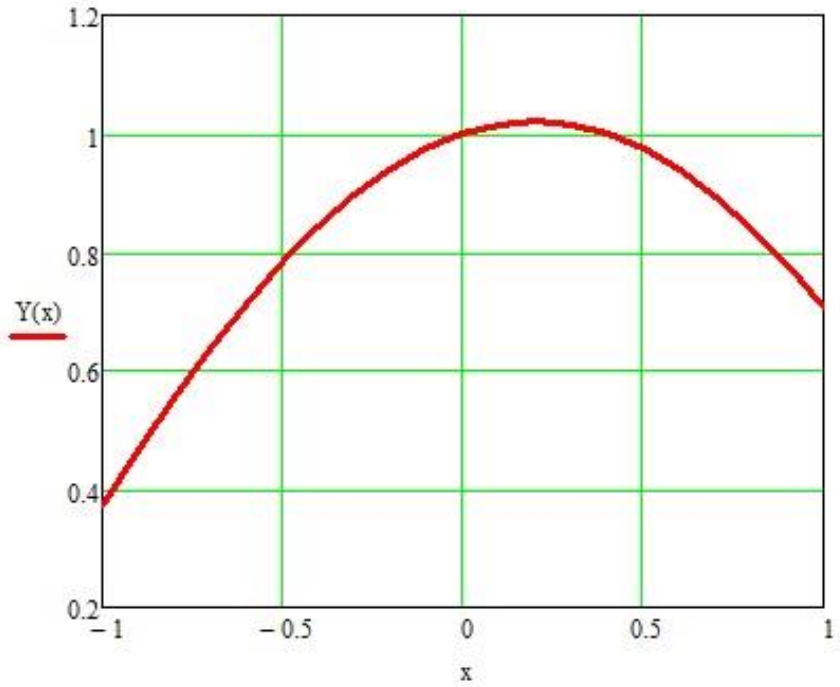
Функціональна залежність змінної *Y* від змінної *X* є залежністю виду *Y = f(X)*, де кожне допустиме значення *X* зіставляється за деяким правилом з **єдиним** можливим значенням змінної *Y*

Геометричне представлення – графік функції

$x := -1, -0.9 .. 1$

$Y(x) := \cos(x) + \frac{\sin(x)}{5}$

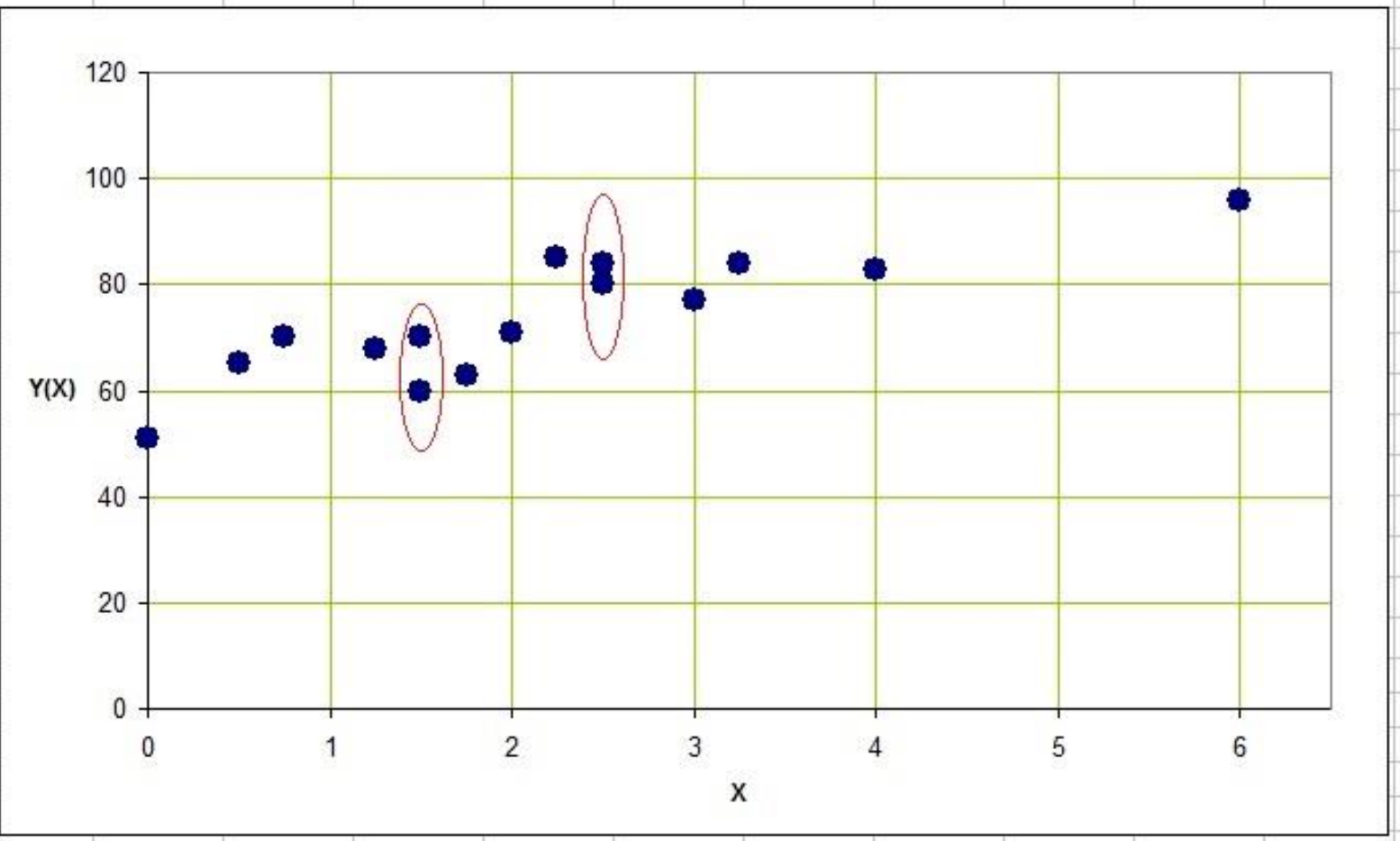
x =	Y(x) =
-1	0.372
-0.9	0.465
-0.8	0.553
-0.7	0.636
-0.6	0.712
-0.5	0.782
-0.4	0.843
-0.3	0.896
-0.2	0.94
-0.1	0.975
0	1
0.1	1.015
0.2	1.02
0.3	1.014
0.4	0.999
0.5	0.973
0.6	0.938
0.7	0.894
0.8	0.84
0.9	0.778
1	0.709



Кореляційна залежність

Кореляційна залежність - це тип статистичного зв'язку, при якому кожне значення *X* відповідає ряду значень *Y*

X	Y(X)
0	51
0,5	65
0,75	70
1,25	68
1,5	60
1,5	70
1,75	63
2	71
2,25	85
2,5	80
2,5	84
3	77
3,25	84
4	83
6	96



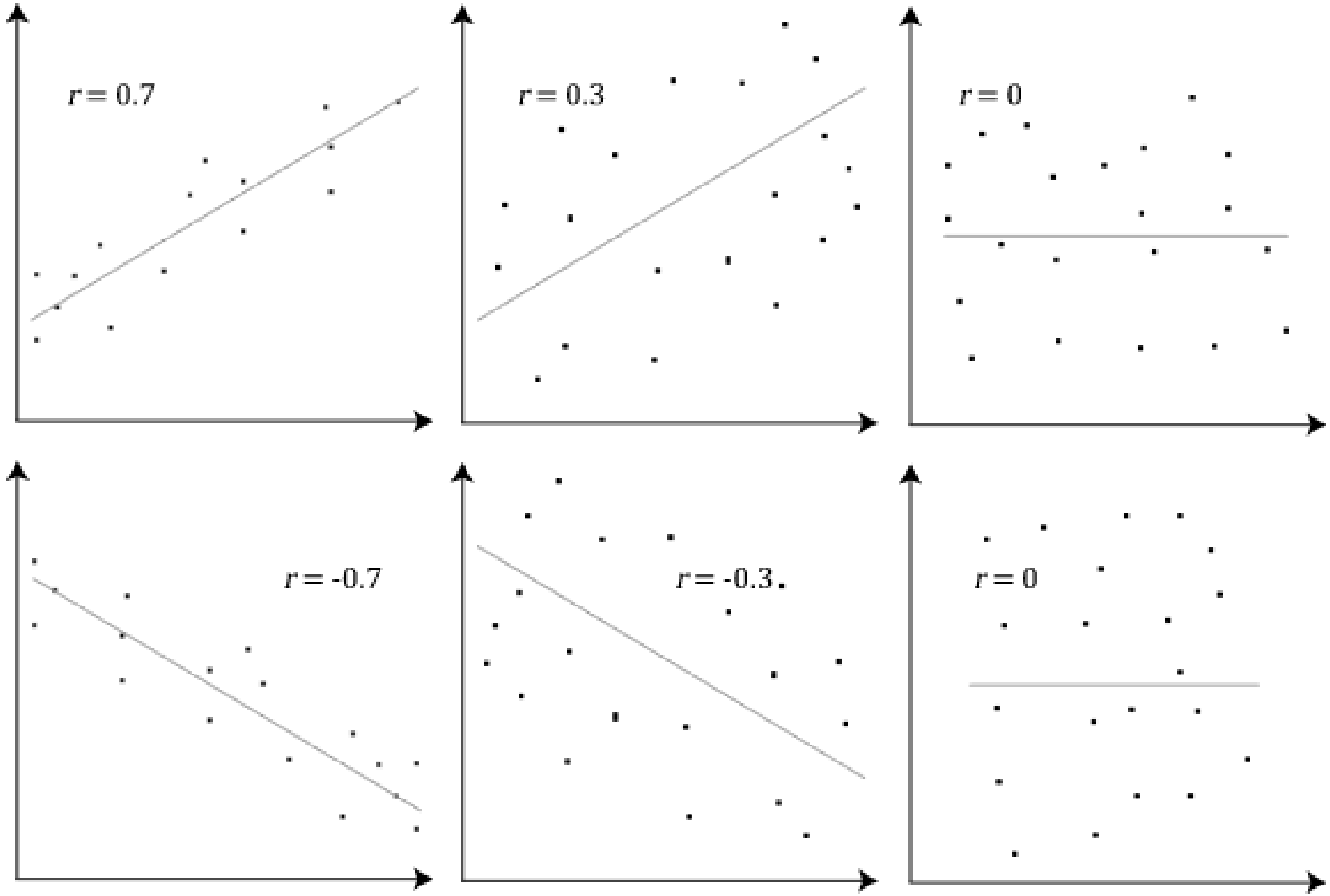
Коефіцієнт кореляції

Коефіцієнт кореляції - це кількісна міра тісноти зв'язку між параметрами *X* і *Y*

Формула розрахунку **коефіцієнта кореляції Пірсона**:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}.$$

Властивості коефіцієнта кореляції



Властивості коефіцієнта кореляції

Значення коефіцієнта кореляції	Характер зв'язку
До | ± 0,3 |	Майже відсутній
| ± 0,3 | - | ± 0,5 |	Слабкий
| ± 0,5 | - | ± 0,7 |	Помірний
| ± 0,7 | - | ± 1,0 |	Сильний

Поняття регресійного аналізу

Регресійний аналіз - це розділ математичної статистики, що вивчає зв'язок між однією або кількома незалежними змінними (факторами) та залежною змінною (результатом)

Мета аналізу - визначення форми зв'язку та розрахунок параметрів математичного рівняння, яке найкраще описує залежність між змінними

Застосування в IT та комп'ютерних науках

Прогнозування навантаження: Наприклад, розрахунок необхідної кількості оперативної пам'яті сервера (Y) залежно від кількості одночасних користувачів (X)

Machine Learning: Навчання моделей з вчителем для передбачення числових значень (ціни, температури, часу виконання запиту)

Оцінка проєктів: Прогнозування фінальної вартості розробки на основі поточної динаміки витрат у методі освоєного обсягу (EVM)

Регресійний аналіз

Лінійна регресія - це зв'язок між ознаками, графіком якої може бути деяка пряма лінія

$$Y=kx+b$$

У MathCAD для обчислення параметрів k і b використовуються функції:

slope(x,y) та intercept(x,y)

Лінійна регресія

DEPENDENT VARIABLE (y) - залежна змінна; те, що потрібно пояснити

INDEPENDENT VARIABLE (x) - незалежна змінна; фактор, який, можливо, може впливати на залежну змінну

LINE OF BEST FIT - лінія найкращої відповідності (регресійна пряма) - якщо є дані лише по ***x***, ця лінія дає найкращу оцінку значення ***y***. Якщо відповідність сильна і немає серйозних викидів, ***x*** можна використовувати як прогноз для ***y***

DATA POINT - точка даних; окреме спостереження

OUTLIER – викид; значення, яке варто вивчити окремо

R^2 = 0.77 - коефіцієнт детермінації; означає, що 77% варіації ***y*** пояснюється змінною ***x***. Значення нижче приблизно 0.30 (30%) означає, що вони навряд чи пов'язані. Вище 0.95 (95%) - вони практично ідентичні

95% CONFIDENCE BAND - 95% довірча смуга; якщо точка даних випадає за межі цих ліній, то на 95% є якась особлива причина, з якої цей результат значно кращий або гірший за інші - такий «викид» варто дослідити