# Лекція 5.
# СТАТИСТИЧНІ МЕТОДИ ВИМІРЮВАННЯ І АНАЛІЗУ ВЗАЄМОЗВ'ЯЗКІВ

**План**

1. Види взаємозв'язків. Загальні прийоми виявлення наявності зв'язку.
2. Лінійний однофакторний кореляційно-регресійний аналіз.
3. Багатомірний аналіз. Непараметричні методи дослідження взаємозв'язків.

1. Усі явища і процеси, що існують в природі та суспільстві взаємопов′язані й взаємообумовлені, тому дослідження об′єктивних зав'язків між ними – **найважливіше завдання статистичного аналізу.** У складному переплетінні всеохоплюючого взаємозв′язку будь-яке явище є наслідком дії певної множини причин і водночас – причиною інших явищ. Але причина сама по собі не визначає наслідку, останній залежить також від умов, у яких діє причина. Вивчаючи закономірності зв′язку, причину та умови об′єднують в одне поняття "**фактор**". Відповідно ознаки, які характеризують фактори (тобто зумовлюють зміни інших, пов′язаних із ними ознак) називаються **факторними** (незалежними) чи просто **факторами**. А ті ознаки, що характеризують наслідки (тобто змінюються під дією факторних ознак) є **результативними** (вислідними).

Залежність між ознаками може проявлятися у функціональній або стохастичній формі.

**Функціональний вид зв'язку** характеризується повною відповідністю між зміною факторної ознаки й зміною результатної величини. Тобто кожному можливому значенню факторної ознаки ($\boldsymbol{x}$) відповідає одне і тільки одне чітко визначене значення результативної (вислідної) ознаки ($\boldsymbol{y}$). Завдяки цьому функціональну залежність можна описати математичними формулами. Такий зв′язок маємо у фізичних, хімічних процесах –залежність довжини ртутного стовпчика від температури навколишнього середовища. Функціональні зв'язки притаманні переважно природним і технічним системам. У екологічних процесах – це зв′язок між елементами розрахункових формул показників – адитивний (a + b + c) або мультиплікативний (a = bc, c = a/b), а також залежність середніх величин від структури сукупності. Функціональні залежності також вивчають точні науки –математика, фізика, хімія.

**Стохастичний (статистичний) вид зв'язку** передбачає, що кожному значенню факторної ознаки відповідає певна множина значень результативної ознаки. Тобто причинна залежність проявляється не в кожному окремому випадку, а в загальному, при великій кількості спостережень. Отже на відміну від функціональних, стохастичні зв′язки неоднозначні. Такі зв′язки виявляються як узгодженість варіації двох чи більше ознак. Наприклад, залежність між рівнем кваліфікації та продуктивністю праці, залежність між кольором очей та кольором волосся тощо.

Різновидом стохастичного зв′язку є **кореляційний зв′язок**, при якому зі зміною факторної ознаки змінюється середнє значення результативної ознаки. Термін "кореляція" означає співвідношення, відповідність. Цей вид зв'язків найчастіше використовують у дослідженнях екологічних явищ, для яких характерно, що поряд з істотними факторами, які формують рівень результативної ознаки, на неї впливає багато інших неврахованих і випадкових факторів. Наприклад, залежність захворюваності від екологічного стану довкілля. На забруднених радіонуклідами територіях стан здоров′я мешканців коливається від "тяжко хворого" до "практично здорового", проте в середньому у таких регіонах, порівняно з екологічно чистими, інтенсивність захворювання значно вища.

На відміну від функціональної залежності, кореляційний зв'язок є неповним (відзначають лише певне співвідношення між причиною і наслідком), оскільки залежність між функцією ($\hat{\boldsymbol{y}}$) і аргументом ($\boldsymbol{x}$) у кожній ситуації перебуває під впливом інших факторів. Кореляційні зв'язки проявляються тільки у масових явищах! З їх допомогою встановлюється тенденція змін результативної ознаки при зміні величини факторної.

Кореляційна залежність може встановлюватися для пари показників (парна кореляція) або для декількох показників (множинна) кореляція. Для виявлення наявності чи відсутності кореляційного зв'язку, використовують ряд специфічних методів–елементарні прийоми, такі як:

- порівняння паралельних рядів даних;
- аналітичне групування (побудова групових та кореляційних таблиць);
- графічне зображення кореляційного поля; а також дисперсійний та кореляційно-регресійний аналіз.

**Порівняння паралельних рядів** є найпростішим із перелічених прийомів і полягає у співставленні ряду значень факторної ознаки та ряду відповідних значень результативної ознаки. Значення фактора розташовують у ранжированому (зростаючому) порядку, а потім простежують співвідношення й напрямок зміни величини результативної ознаки. Наприклад, є умовні дані витрат на рекламу (факторна ознака) та кількість замовників (результативна ознака) поліграфічних фірм:

| № п/п фірми | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Рекл. витрати (ум. грош. од.) | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 10 | 10 | 10 | 10 |
| Кількість замовн. (чол.) | 800 | 850 | 720 | 850 | 800 | 880 | 950 | 820 | 900 | 1000 | 920 | 1060 |

Тут можна побачити, що в цілому для усієї сукупності фірм збільшення затрат на рекламу приводить до збільшення кількості замовників, хоча в окремих випадках наявність такої залежності може й не убачатися. Якщо зіставити дані по фірмах із номерами 2 та 5 або 7 та 11 побачимо зворотне співвідношення. Це пояснюється тим, що в кожному окремому випадку, кількість замовників залежить не тільки від розміру затрат фірми на рекламу, а й від інших факторів.

Отже у тих випадках, коли зростання величини факторної ознаки тягне за собою зростання величини результативної ознаки, говорять про можливу наявність **прямого кореляційного зв'язку**. Якщо ж із збільшенням факторної ознаки величина результативної ознаки має тенденцію до зменшення, то можна припустити, що існує **зворотній кореляційний зв'язок** між ознаками.

Зверніть увагу, що наявність великої кількості значень результативної ознаки, які відповідають одному й тому ж значенню ознаки-фактора, ускладнює сприйняття таких паралельних рядів особливо при великій кількості одиниць досліджуваної сукупності. У таких випадках для встановлення факту наявності зв'язку доцільніше скористатися статистичними таблицями.

**Аналітичне групування** належить до найважливіших методів дослідження взаємозв'язків. Виконується побудовою групових статистичних таблиць – усі спостереження поділяють на групи за величиною факторної ознаки і для кожної групи обчислюють середні значення результативної ознаки. Порівнюючи зміни середніх, виявляють характер зв'язку. У наведеному вище прикладі факторна ознака представлена трьома варіантами повторюваних значень, отже маємо групування:

| Групи поліграфічних фірм за рекламними затратами (умовн. грош. од.) | Кількість фірм у групі | Середня кількість замовників фірм даної групи (чол.) |
|---|---|---|
| 8 | 3 | 790 |
| 9 | 5 | 860 |
| 10 | 4 | 1980 |

Порівнявши середні значення результативної ознаки по групах, можна припустити (але не стверджувати) наявність прямого кореляційного зв'язку між досліджуваними ознаками. Отже більший внесок у рекламу може сприяти збільшенню кількості замовників.

Іншим можливим прийомом виявлення зв'язку є **використання кореляційних таблиць**. В таких таблицях факторну ознаку розташовують, як правило, у рядках, а результативну – у графах (стовпчиках). Числа у клітинках перетину означають частоту повторення даної комбінації значень x та y. Для кожного рядка розраховують середнє значення результативної ознаки, порівнюючи які і, проводять аналіз.

Слід звернути увагу і на розташування частот по діагоналі таблиці. Якщо частоти розміщені на діагоналі з лівого верхнього до правого нижнього кута (більшим значенням фактора відповідають більші значення результату) – можна припустити наявність прямої кореляційної залежності, якщо ж навпаки, з правого верхнього до лівого нижнього (більшим значенням фактора відповідають менші значення результату) – припускають наявність зворотного зв'язку між ознаками.

**Графічно** взаємозв'язок двох ознак зображується за допомогою поля кореляції. В прямокутній системі координат на вісі абсцис відкладаються значення факторної ознаки, а на вісі ординат – результативної і отримують точковий графік, який називають "полем кореляції".

За характером розміщення точок можна судити про напрям і силу зв'язку:

 точки розташовані хаотично – це свідчить про відсутність тісних зв'язків;

 сконцентровані навколо діагоналі від нижнього лівого кута координат до верхнього правого – це щільний прямий зв'язок;

 сконцентровані навколо діагоналі від верхнього лівого кута координат до правого нижнього – це зворотній зв'язок між досліджуваними ознаками.

Якщо на такий графік нанести середні значення результативної ознаки і з'єднати відрізками відповідні точки, отримаємо **емпіричну лінію зв'язку**, яка відображує форму зв'язку–лінійну (означає рівномірну зміну залежних ознак) чи криволінійну (нерівномірну).

Розглянуті прийоми характеризують лише загальні риси зв'язку, його тенденцію. Визначити внесок кожного з факторів, а також тісноту зв'язку дозволяють методи кореляційно-регресійного та дисперсійного аналізу.

2. **Кореляційний аналіз** має своїм завданням кількісно визначити та оцінити тісноту (силу) статистичного зв'язку між двома ознаками. Він не встановлює причин залежності між досліджуваними ознаками, а лише виявляє наявність самої залежності, її силу та напрям. Наприклад, необхідно з'ясувати, в якій мірі травмування водія в автомобільній аварії пов'язане з відсутністю засобів безпеки. Наскільки сильною є кореляція між цими подіями?

**Регресійний аналіз** полягає у визначенні аналітичного виразу кореляційного зв'язку – опису виду і параметрів функції зв'язку (регресійної моделі). Термін "регресія" (від лат. regredior – повертаюсь)означає повернення до середньої. За числом факторних ознак , які входять у регресійну модель, розрізняють одно- та багатофакторні моделі.

Важливими умовами правильного практичного застосування кореляційно-регресійного аналізу являються:

 однорідність та нормальний характер розподілу одиниць, які підлягають вивченню методами кореляційно-регресійного аналізу;

 достатня кількість спостережень;

 незалежність один від одного факторів, які виділені для дослідження.

На практиці іноді виникають відхилення від означених передумов, але це зовсім не означає відмови від використання кореляційно-регресійних методів аналізу в дослідженнях.

Отже в основі **кореляційно-регресійного аналізу** лежить припущення про те, що залежність між значенням факторної ознаки ($x$) і середнім значенням результативної ($\hat{y}$) може бути представлена у вигляді функції

$$\hat{y} = f(x),$$

лінії регресії $x$ на $\hat{y}$, яка називається рівнянням простої парної регресії – **однофакторною регресійною моделлю**.

Якщо залежні ознаки змінюються більш-менш рівномірно – емпірична лінія зв'язку (лінія групових середніх) наближається до прямої – зв'язок між ними можна описати за допомогою лінійної функції

$$\hat{y} = a + bx,$$

де $a$ і $b$ – параметри **лінійного регресійного рівняння**.

Параметр ***b* – коефіцієнт регресії**, розглядається як ефект впливу $x$ на $\hat{y}$. Він показує, на скільки одиниць в середньому змінюється результативна ознака $\hat{y}$ зі зміною факторної ознаки $x$ на одиницю. При прямому зв'язку між залежними ознаками $b$ – величина додатна, при оберненому – від'ємна.

Параметр ***a* – вільний член** рівняння регресії, це значення $\hat{y}$ при $x = 0$. Якщо межі варіації не містять нуля, то цей параметр має лише розрахункове значення, тобто показує усереднений вплив на результативну ознаку неврахованих або не виділених для дослідження факторів.

Рівняння регресії відбиває закон зв'язку між $x$ та $\hat{y}$ не для окремих елементів сукупності, а для сукупності в цілому. Закон, який абстрагує вплив інших факторів, виходить із принципу "за інших однакових умов".

Наступним кроком є визначення параметрів рівняння зв'язку **методом найменших квадратів**, основною умовою якого є мінімізація суми (S) квадратів відхилень емпіричних (фактичних) значень результативної ознаки ($y$) від теоретичних (розрахункових, обчислених за рівнянням регресії) ($\hat{y}$):

$$S = \sum (y - \hat{y})^2 \Rightarrow \min.$$

Розглядаючи S в якості функції параметрів $a$ і $b$ та, виконавши математичні перетворення (диференціювання) отримаємо можливість знайти оцінки цих параметрів. Для їх обчислення складають і розв'язують систему нормальних рівнянь для прямого зв'язку, де n–обсяг досліджуваної сукупності (число одиниць спостереження):

$$\sum y = na + b\sum x,$$
$$\sum xy = a\sum x + b\sum x^2.$$

Отже для визначення шуканих параметрів регресійного рівняння ($a$ і $b$) необхідно обчислити за фактичними даними такі величини:

n, $\sum y$, $\sum x$, $\sum xy$, $\sum x^2$.

На практиці часто дослідження проводять за великою кількістю спостережень. Тому вихідні дані зручніше надавати у зведеній кореляційній таблиці.

Визначивши параметри регресійного рівняння слід оцінити їх значимість, оскільки вони відіграють важливу роль у прогнозуванні показників. Значимість коефіцієнта регресії ($b$) оцінюють за допомогою ***t*–критерію Стьюдента** – фактичні дані підставляють у формулу критерію і обчислюють його розрахункове значення. Потім порівнюють з критичним (табличним). Якщо $t_{розр.} > t_{табл.}$, то коефіцієнт регресії ($b$) вважається значимим.

Необхідно також оцінити і адекватність регресійної моделі, тобто можливість надійного прогнозування середніх значень результативної ознаки за даними значеннями факторної ознаки. Для оцінки надійності застосовують ***F* – критерій Фішера** і метод перевірки гіпотез.
Якщо $F_{розр.} > F_{табл.}$, то гіпотеза про значущість рівняння приймається.

Щоб знайти теоретичні **прогнозні значення** результативної ознаки, необхідно підставити в отримане рівняння регресії конкретні значення факторної ознаки. Так одержують прогноз показника ($\hat{y}$) за умов збереження загальної тенденції розвитку явища у часі по даним за минуле, і екстраполяції здобутих залежностей на майбутнє. Правильність розрахунку перевіряється рівністю сумарних теоретичних та емпіричних значень результативної ознаки при їх зіставленні.

Слід звернути увагу на те, що моделям, які побудовані на основі рівнянь регресії, притаманні слабкі екстраполяційні властивості (поширення кількісних характеристик і висновків на іншу сукупність, інший час, за межі вивчаємої сукупності). Вони не відображують тенденцій розвитку суспільних процесів і використовуються для побудови короткочасних прогнозів. Інтерпретація моделей регресії дозволяє виявити лише резерви розвитку досліджуваних явищ.

Для кількісної оцінки **сили зв′язку** (узгодженості варіацій взаємозв′язаних ознак) статистика використовує низку коефіцієнтів із такими спільними властивостями:

- за відсутності будь-якого зв′язку значення коефіцієнта наближається до нуля;
- при функціональному зв′язку – до одиниці;
- за наявності кореляційного зв′язку коефіцієнт виражається дробом (частіше десятковим), який за абсолютною величиною тим більший, чим тісніший зв′язок.

Найпоширенішим є **лінійний коефіцієнт кореляції Пірсона (*r*)** – характеризує тісноту і напрям зв′язку між двома ознаками, що корелюють у випадку наявності між ними лінійної залежності. Змінюється в межах: $-1 \le r \le 1$. При прямому зв′язку $r$ – величина додатна, при зворотному – від′ємна. Знаки коефіцієнтів кореляції і регресії співпадають.

Для одержання висновків про практичне застосування побудованої регресійної моделі значенням коефіцієнта надається якісна оцінка, яка визначається за шкалою Чеддока:

| Значення коеф. *r* | $\lvert\pm 0{,}1\rvert - \lvert\pm 0{,}3\rvert$ | $\lvert\pm 0{,}3\rvert - \lvert\pm 0{,}5\rvert$ | $\lvert\pm 0{,}5\rvert - \lvert\pm 0{,}7\rvert$ | $\lvert\pm 0{,}7\rvert - \lvert\pm 0{,}9\rvert$ | $\lvert\pm 0{,}9\rvert - \lvert\pm 0{,}99\rvert$ |
|---|---|---|---|---|---|
| Зв′язок | Практично відсутній | Слабкий | Помірний | Сильний (щільний) | Дуже сильний |

До інтерпретації отриманих коефіцієнтів кореляції слід підходити особливо обережно у разі незначного обсягу досліджуваної (вибіркової) сукупності. Тому необхідною є перевірка істотності кореляційного зв′язку, яка виконується за допомогою **t**-критерію Стьюдента і ґрунтується на порівнянні емпіричних значень ($t$) з критичними, які могли б виникнути за відсутності зв′язку. При цьому висувається й перевіряється гіпотеза ($H_0$) про рівність лінійного коефіцієнта кореляції нулю [$H_0 : r = 0$]. Якщо фактичне значення критерію перевищує критичне, то зв′язок між ознаками не випадковий.

3. Більш загальним завданням регресійного аналізу є з′ясування і одержання залежності між групою незалежних факторів ($x_1, x_2, \ldots, x_m$) і показником (результативною ознакою $\hat{y}$). Це може бути зміна курсу долара до гривні залежно від часу, рівня емісії, процентної ставки НБУ та інших факторів. Або зміна рівня злочинності в Україні залежно від зміни факторів: рівня безробіття, заробітної плати, рівня алкоголізму і наркоманії тощо. Таке вивчення зв′язку між ознаками має назву множинної (багатофакторної) регресії і описуються узагальненою регресійною моделлю

$\hat{y}=f(x_1, x_2, \ldots, x_m)$

та багатофакторною лінійною моделлю

$\hat{y}=a+b_1x_1+b_2x_2+\ldots+b_nx_n$,

де $(b_1, b_2, \ldots, b_m)$ – параметри моделі, які необхідно оцінити.

Передумовою застосування множинного аналізу є відсутність функціонального зв'язку між факторами. Такий аналіз надає можливість оцінити вплив на досліджуваний результативний показник кожного із введених у модель факторів при фіксованому положенні на середньому рівні інших факторів. Модель є придатною для практичного використання лише тоді, коли є адекватною, тобто статистично значимими є усі змінні!

Проведення множинного аналізу на практиці доцільно виконувати з використанням спеціальних пакетів прикладних програм, залишаючи за дослідником визначення виду моделі та інтерпретацію результатів моделювання.

Розглянуті вище методи вимірювання взаємозв'язків між ознаками називають **параметричними**, оскільки вони базуються на використанні середніх величин і дисперсій, які є основними параметрами розподілу. Зрозуміло, що параметричні методи не можна застосовувати, якщо ознаки не піддаються кількісному виміру (є атрибутивними) або не виконується припущення про нормальний розподіл результативної ознаки (як кількісної так і якісної) для сукупностей незначного обсягу. В таких випадках застосовують **непараметричні методи дослідження взаємозв'язків**, які:

- не вимагають числового вираження значень ознак;
- не вимагають обчислення параметрів розподілу;
- не вимагають інформації про розподіл ознак в сукупності.

Але непараметричні методи забезпечують лише оцінку щільності зв'язку та перевірку його істотності і не дають змогу представити зв'язок аналітично.

В основі обчислення щільності зв'язку між атрибутивними ознаками лежить побудова таблиць співзалежності (взаємного спряження), у яких представлені комбінаційні розподіли сукупностей за факторною ознакою – по рядках, та результативною – по графах. Найбільш поширеними є таблиці 2×2 :

| Ознака | $A$ | Не $A$ | $\sum B$ |
|---|---|---|---|
| $B$ | $a$ | $b$ | $a+b$ |
| Не $B$ | $c$ | $d$ | $c+d$ |
| $\sum A$ | $a+c$ | $b+d$ | $a+b+c+d$ |

Для вимірювання щільності зв'язку між двома альтернативними ознаками використовують **коефіцієнт асоціації ($K_A$)** та **коефіцієнт контингенції ($K_K$)**:

$$K_A = \frac{ad-bc}{ad+bc}$$

$$K_K = \frac{ad-bc}{\sqrt{(a+b)\cdot(b+d)\cdot(a+c)\cdot(c+d)}}.$$

Зв′язок вважається підтвердженим, якщо $K_A \geq 0,5$ чи $K_K \geq 0,3$(!). Наприклад, проаналізувати успішність студентів залежно від статі, виділивши дві групи: студенти, що склали іспит, та студенти, що не склали іспит.

| Стать | Кількість студентів | | Разом |
|---|---|---|---|
| | Склали | Не склали | |
| Дівчата | a = 25 | b = 2 | a + b = 27 |
| Хлопці | c = 20 | d = 3 | c + d = 23 |
| Разом | a + c = 45 | b + d = 5 | 50 |

$K_A = 0,09 < 0,5$, а $K_K = 0,00025 < 0,3$, тобто між статтю та успішністю студентів зв′язок надто незначний, практично відсутній. Отриманий висновок можна вважати справедливим, тому, що істотними факторами успішності є не стать, а здібності, зацікавленість, організованість, відвідування занять, кількість годин самостійної роботи та інше.

Коли кожна з якісних ознак складається більше ніж із двох груп(неквадратні таблиці), то для визначення щільності зв′язку застосовують **коефіцієнт взаємної спряженості Пірсона–Чупрова (*C*)**, який теж приймає значення від 0 до 1.

Якщо одна із взаємопов’язаних ознак має кількісний вираз, а друга – альтернативний, то показником щільності є **бісеріальний коефіцієнт кореляції** (бісерія означає дві серії). Наприклад, залежність рівня доходів (кількісна ознака) від рівня освіти (атрибутивна).

Оцінка всіх непараметричних показників здійснюється через *t*-критерій Стьюдента.

Для обробки даних соціологічних досліджень (анкет), медичних обстежень, рейтингів, експертних оцінок тощо, тобто там, де ознаки вимірюються за допомогою номінальної та порядкової шкали (наприклад, “стать”, “соціоекономічний статус сім’ї”, “діагноз” і т.ін.) часто застосовують **методи рангової кореляції**.

В основу непараметричних методів рангової кореляції покладено принцип нумерації значень статистичного ряду.

Кожній одиниці сукупності надається порядковий номер за величиною значення окремої ознаки – **ранг** (натуральне число 1, 2, 3, ...). Ранжування, тобто процедура упорядкування об′єктів вивчення на основі віддавання переваг, проводиться за кожною ознакою окремо. При ранжуванні значень факторної і результативної ознак слід використовувати один принцип – або від менших значень до більших, або від більших до менших! Кількість рангів дорівнює обсягу сукупності. Зі збільшенням обсягу ступінь “розпізнаваності” елементів зменшується, тому рангові оцінки щільності зв′язку доцільно використовувати для сукупностей невеликого обсягу.

До рангових оцінок щільності належать **коефіцієнт кореляції рангів Спірмена (*ρ*)** – базується на основі різниці рангів (d) факторної і результативної ознак для кожної одиниці сукупності та **Кендала (τ)**. Ці коефіцієнти мають ті самі властивості, що і лінійний коефіцієнт кореляції – змінюються в тих же межах, оцінюють щільність зв′язку та вказують його напрям. Зв′язок між ознаками можна визнати статистично істотним, якщо значення коефіцієнтів рангової кореляції Спірмена і Кендала більше 0,5.

Наприклад, експерти оцінили технічний та фінансовий стан семи підприємств видавничої галузі в балах за певними критеріями: 1-й ранг – найменшому значенню ознаки, останній – найбільшому (умовні дані):

| № підпр-ва | Експертні оцінки | | Ранги | | $d = R_x - R_y$ | $d^2$ |
|---|---|---|---|---|---|---|
| | Технічний стан (x) | Фінансовий стан (y) | $R_x$ | $R_y$ | | |
| 1 | 27 | 26 | 1 | 2 | -1 | 1 |
| 2 | 30 | 25 | 2 | 1 | 1 | 1 |
| 3 | 38 | 30 | 6 | 4 | 2 | 4 |
| 4 | 36 | 32 | 5 | 5 | 0 | 0 |
| 5 | 33 | 28 | 3 | 3 | 0 | 0 |
| 6 | 42 | 37 | 7 | 7 | 0 | 0 |
| 7 | 35 | 33 | 4 | 6 | -2 | 4 |
| Разом | x | x | x | x | x | 10 |

$$p = 1 - \frac{6\sum d_i^2}{n \cdot \left(n^2 - 1\right)} = 0.821,$$

0.821>0.5, що свідчить про наявність прямого зв′язку між технічним і фінансовим станом підприємств видавничої спрямованості.