Лекція № 8

Статистичний аналіз даних

План

1. Статистичний аналіз, класифікація методів статистичного аналізу
2. Описова статистика
2.1 Частотний розподіл
2.2 Відсоткові показники
2.3 Заходи центральної тенденції
2.4 Міри розкиду даних
3. Методи вторинної статистичної обробки результатів
3.1 Методи порівняння елементарних статистик (параметричні та непараметричні методи)
3.2 Кореляційний аналіз
4. Роль і значення графічного методу в статистиці

1. Статистичний аналіз, класифікація методів статистичного аналізу

Статистика - це точна наука, що вивчає методи збору, аналізу і обробки даних, які описують масові дії, явища і процеси. Дані, що вивчаються в статистиці, зачіпають не окремі об'єкти, а їх сукупності.

Статистика в біології та медицині є одним з інструментів аналізу експериментальних даних і клінічних спостережень, а також мовою, за допомогою якої повідомляються отримані математичні результати.

Статистичні методи включають як прості методи, які доступні навіть непідготовленим користувачам, так і складні математичні процедури, доступні лише кваліфікованим фахівцям високого класу.

Методами статистичної обробки результатів експерименту називаються математичні прийоми, формули, способи кількісних розрахунків, за допомогою яких показники, що одержані в ході експерименту, можна узагальнювати, приводити в систему, виявляючи приховані в них закономірності.

Головна мета будь-якого статистичного методу - представити кількісні дані в систематизованій і стислій формі з тим, щоб полегшити їх розуміння.

Всі методи статистичного аналізу умовно діляться на первинні і вторинні.

ПЕРВИННІ: методи, за допомогою яких можна отримати показники, що безпосередньо відображають результати отриманих в експерименті вимірювань. Це:

- визначення середньої арифметичної,
- дисперсії,
- моди
- медіани

ВТОРИННІ - методи статистичної обробки, за допомогою яких на базі первинних даних виявляють приховані в них статистичні закономірності. Це:

- кореляційний аналіз,
- регресійний аналіз,
- факторний аналіз,
- методи порівняння первинних даних двох або декількох вибірок.

Генеральна сукупність і вибірка.

Сукупність, з якої відбирають певну частину її членів для спільного вивчення, називають **генеральною**.

Відібрана тим чи іншим способом частина генеральної сукупності отримала назву **вибіркової сукупності** або **вибірки**.

Загальну суму членів генеральної сукупності називають її **обсягом** і позначають буквою **N.**

Обсяг генеральної сукупності нічим не обмежений, тобто генеральну сукупність представляють як нескінченно велику безліч відносно однорідних одиниць або членів, що складають її зміст. Обсяг вибірки, що позначається буквою n, може бути і великим, і малим, але він не може містити менше двох одиниць.

2. Описова статистика

Описова статистика дозволяє узагальнювати первинні результати, отримані при спостереженні або в експерименті. Процедури тут зводяться:

- до угруповання даних по їх значенням,
- побудови розподілу їх частот,
- виявлення центральних тенденцій розподілу (наприклад, середньої арифметичної, моди, медіани)
- до оцінки розкиду даних по відношенню до знайденої центральної тенденції.

2.1 Частотний розподіл

Як правило, в результаті емпіричного дослідження буває досить багато вихідних первинних даних, які підлягають статистичній обробці. Наприклад, колонка з 1000 тестових показників.

Оскільки їх багато – можна скласти таблицю частотного розподілу:

Ряд первинних даних виписують у перший стовпчик таблиці в порядку убування.

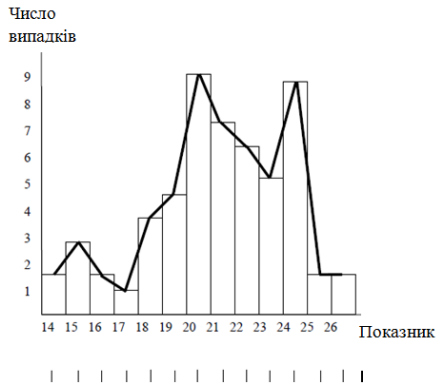
Коли показники розподілені по порядку, підраховують кількість випадків для кожного показника. Отримане таким способом число і є частота (кількість випадків) для відповідного показника. Сума всіх частот дорівнює загальному числу тестових показників (або обсягом вибірки n).

У другій стовпчик впишемо частоту зустрічаємості кожного первинного результату (див. табл. 1).

Таблиця 1 частотного розподілу даних по тесту

Первинний результат	Частота
26	2
25	2
24	8
23	5
22	6
21	7
20	9
19	5
18	4
17	1
16	2
15	3
14	2

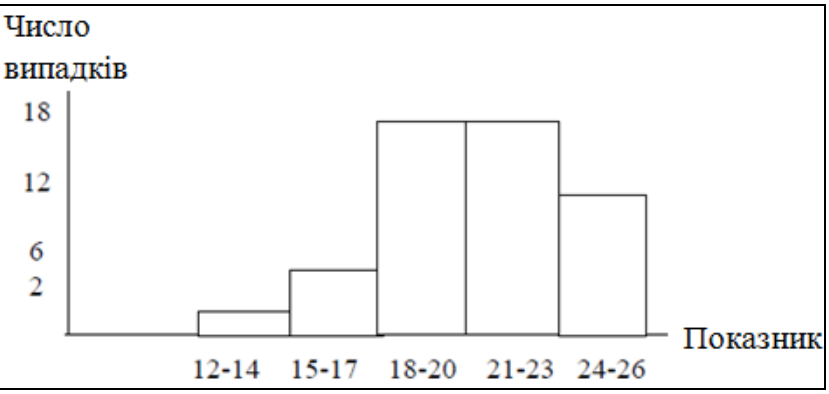
Інформація, що міститься в частотному розподілі, може бути також представлена графічно у вигляді кривої розподілу



Крива полігону частот і гістограма

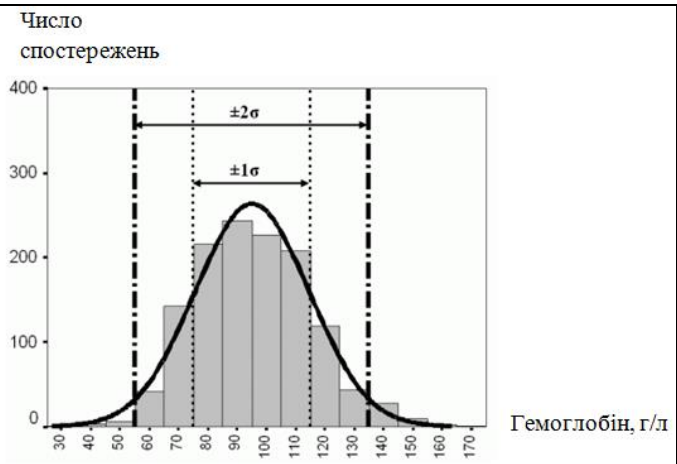
Для того, щоб зробити узагальнені дані про характер розподілення результатів по тесту і в разі, якщо отримано занадто велике число значень первинного результату, необхідно зробити угруповання даних і провести аналогічну процедуру побудови частотного розподілу. Згрупуємо дані, представлені в табл. 1, об'єднуючи їх по 3 одиниці в кожній групі (або з інтервалом 3 одиниці). Угруповання даних проводиться від мінімального значення до максимального. Для того, щоб інтервали значень були рівномірними, додамо ще два значення в нижній частині ряду (12 і 13), так як 26 балів є максимальним для даної методики. Частота цих первинних показників, що зустрічаються в нашій вибірці дорівнює 0. Частотний розподіл придбає такий вигляд:

Інтервал значень	Частота
24-26	12
21-23	18
18-20	18
15-17	6
12-14	2



Судячи з характеру розподілу, представленого на рис.2 воно не є нормальним і характеризується невеликою асиметрією зі зрушенням в бік високих значень.

Ідеальна нормальна крива зображена на даному рисунку



Цей тип кривої володіє важливими математичними властивостями і на ній засновані багато видів статистичного аналізу. По суті, ця крива означає, що число випадків максимально в середині розподілу і поступово спадає до її країв. Крива симетрична і має єдиний пік в центрі.

Більшість розподілів чисельних показників наближаються до нормальної кривої. Можна відзначити таку закономірність: чим більше група, тим ближче розподіл показників до нормальної кривої

2.2 Відсоткові показники

Відсоткові показники використовуються для того, щоб частотний розподіл за тією чи іншою змінною привести до основи 100 (аналогічно, пропорції використовуються для приведення даних до основи 1). У такому вигляді дані є кращими в інтуїтивному сенсі в порівнянні з «сирим» частотним розподілом.

Приклад (успішність здачі сесії):

Успішність	Частота	%	Пропорції
Були трійки	28	45,2	0,452
Без трійок, в основному на чотири	11	17,7	0,177
Без трійок, в основному на пять	13	21,0	0,210
На відмінно	10	16,1	0,161
Разом	62	100,0	1,000

2.3 Заходи центральної тенденції

Заходи центральної тенденції (мода, медіана і середнє арифметичне) дають інформацію про типове або центральне значення розподілу.

Мода говорить про значення, що найбільш часто зустрічається,

Медіана - про серединне значення,

Середнє арифметичне - про найбільш очікуване значення.

Найбільш часто використовуть середнє арифметичне. Його обчислюють, розділивши суму всіх значень даних на число цих даних.

$$\bar{x} = \frac{\sum x}{n}$$

2.4 Міри розкиду даних

Для більш повного опису результатів емпіричного дослідження використовуються міри розкиду (або варіативності) даних, що характеризують ступінь індивідуальних відхилень від центральної тенденції. Найбільш наочним і відомим способом подання розкиду є розмах розподілу, тобто різницю між найвищим і найнижчим результатом. Але ця міра вкрай неточна і нестійка, тому що вона визначається тільки двома показниками, і єдиний надзвичайно високий або низький результат може помітно вплинути на величину розмаху. Більш точний метод вимірювання розкиду даних заснований на обліку різниці між кожним індивідуальним результатом і середньоарифметичним значенням по групі. Такий мірою розкиду є **дисперсія або середній квадрат відхилення** (σ^2).

Дисперсія як статистична величина характеризує, наскільки

приватні значення відхиляються від середньої величини в даній вибірці. Чим більша дисперсія, тим більше відхилення або розкид даних.

Дуже часто замість дисперсії для виявлення розкиду приватних даних щодо середньої використовують похідну від дисперсії величину - стандартне (або вибіркове) відхилення. Воно дорівнює квадратному кореню, який витягується з дисперсії, і позначається тим же знаком, тільки без квадрата (σ). Ця величина в ряді випадків виявляється більш зручною характеристикою варіювання, ніж дисперсія, так як виражається в тих же одиницях, що і середня арифметична величина.

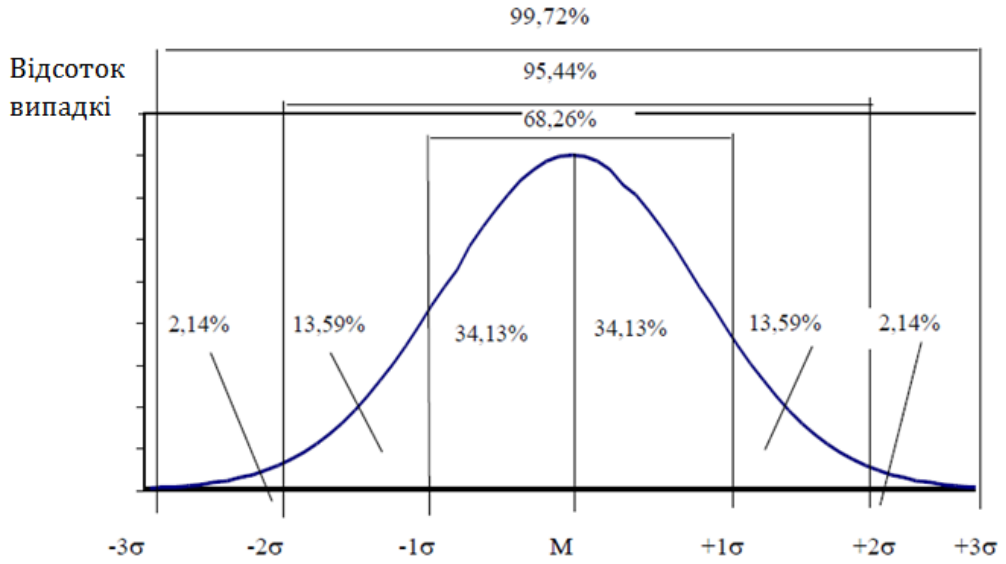
Визначають стандартне відхилення від середнього за формулою:

$$S_{\bar{x}} = \sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \cdots + \Delta x_n^2}{n}}$$

Для того, щоб більш точно оцінити стандартне відхилення для малих вибірок (з числом елементів менше 30), в знаменнику виразу під коренем треба використовувати не n, а n-1:

$$S_{\bar{x}} = \sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \cdots + \Delta x_n^2}{n-1}}$$

Особливо чіткою виявляється інтерпретація σ стосовно нормальної або до приблизно нормальної кривої розподілу, тому що тут існує пряма відповідність між σ і відносною кількістю випадків. На рис. по горизонтальній осі відкладені інтервали, які відповідні відхиленню в 1σ, 2σ і 3σ вправо і вліво від середнього значення (М). Відсоток випадків, що припадають на інтервал М+1σ в нормальному розподілі дорівнює 34,13. Оскільки крива симетрична, 34,13% випадків припадає також на інтервали від М до -1σ, так що діапазон від -1σ до + 1σ охоплює 68,26% випадків. Майже всі випадки (99,72%), тобто майже всі показники лежать від -3σ до + 3σ щодо середнього значення.



Більшість розподілів первинних результатів дослідження ближче до нормального розподілу, ніж до будь-якого іншого.

Властивості нормального розподілу

1. У нормальному розподілі всі міри центральної тенденції рівні між собою, тобто сходяться в одній точці на графіку (M = Me = Mo).

2. У нормальному розподілі приблизно 99% всіх значень досліджуваної змінної знаходиться в межі M±3σ. Відсотковий розподіл випадків на кривій нормального розподілу називають "законом трьох сігм" (див. рисунок).

3. Крива нормального розподілу має вигляд дзвона, вона симетрична (асиметрія відсутня, As = 0) і не має надто гострою або занадто плоскої вершини (ексцес відсутній, Ex = 0).

Нормальний розподіл → Можна застосовувати стандартні вторинні методи: t-критерій та дисперсійний аналіз

Нормальний розподіл (або мала вибірка) → Необхідно застосовувати стандартні непараметричні критерії

3. Методи вторинної статистичної обробки результатів

За допомогою вторинних методів статистичної обробки даних безпосередньо перевіряються, доводяться або спростовуються гіпотези, пов'язані з емпіричним дослідженням.

Найчастіше в дослідженнях застосовують такі методи вторинної статистичної обробки результатів:

1) Методи порівняння двох або кількох елементарних статистик (середніх, дисперсій, тощо), що відносяться до різних вибірок;

2) Методи встановлення статистичних зв'язків між змінними (наприклад, їх кореляції між собою);

3) Методи виявлення внутрішньої статистичної структури емпіричних даних (наприклад, факторний аналіз).

1) Методи порівняння елементарних статистик

Жодне біологічне дослідження не обходиться без порівнянь.

Порівнюють:

- результати, отримані двома різними групами досліджуваних;
- результати, отримані однією вибіркою випробовуваних, але в різний час або в різних умовах.

Про відмінності між ними судять зазвичай за різницею між середніми, дисперсіями і іншими вибірковими показниками.

У статистиці широке застосування отримала так звана ***нульова гіпотеза (Но)***.

Суть її зводиться до припущення, що різниця між генеральними параметрами порівнюваних груп дорівнює нулю і що відмінності, які спостерігаються між вибірковими характеристиками носять не систематичний, а виключно випадковий характер.

Для перевірки цієї гіпотези використовують спеціальні критерії достовірності (***параметричні критерії***): t - критерій Стьюдента і F - критерій Фішера (за умови нормального розподілу досліджуваної змінної). Перший використовують для порівняльної оцінки середніх величин, другий - для порівняльної оцінки дисперсій.

Якщо розподіл не нормальний, то використовують ***непараметричні критерії.***

ПАРАМЕТРИЧНІ СТАТИСТИЧНІ МЕТОДИ

t - критерій Стьюдента

Використовують для порівняння вибіркових середніх величин, що належить до двох совокупностей даних, і для вирішення питання про те, чи відрізняються середні значення статистично достовірно одне від іншого.

Метод Стьюдента різний для незалежних і залежних вибірок. **Незалежні вибірки** отримують при дослідженні двох різних груп випробовуваних. Для аналізу різниці середніх застосовують формулу:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{m_1^2 + m_2^2}}$$

Після того як за допомогою формули вирахувано показник t, за спеціальною таблицею для заданого числа ступенів свободи, що дорівнює $(n_1 + n_2)-2$ і обраної ймовірності припустимої помилки (таблицю), знаходять потрібне табличне значення t і порівнюють з ним розраховане значення t. Якщо обчислене значення t більше або дорівнює табличному, то роблять висновок про те, що порівнювані середні значення з двох вибірок дійсно статистично достовірно різняться з прийнятою ймовірністю допустимої помилки, тобто нульова гіпотеза не вірна.

До **залежних вибірок** відносяться, наприклад, результати однієї і тієї ж групи випробовуваних до і після впливу незалежної змінної.

Даний метод порівняння середніх величин застосовується тоді, коли необхідно, наприклад, встановити, вдався чи не вдався експеримент, надав або не чинив він вплив на рівень, наприклад, збільшення сполук чи клітин в організмі. Оцінюються залежні змінні на початку і в кінці експериментального дослідження.

Отримавши такі оцінки і обчисливши середнє по всій вивченій вибірці випробовуваних, ми можемо скористатися критерієм Стьюдента для точного встановлення наявності або відсутності статистично достовірних відмінностей між середніми до і після експерименту.

Якщо виявиться, що вони дійсно вірогідно розрізняються, то можна буде зробити певний висновок про те, що експеримент вдався. В іншому випадку немає переконливих підстав для такого висновку навіть в тому випадку, якщо самі середні величини на початку і в кінці експерименту за своїми абсолютними значеннями різні.

Для визначення достовірності різниці середніх у разі залежних вибірок застосовується наступна формула:

$$t = \frac{\sum d}{\sqrt{\frac{n\sum d^2 - (\sum d)^2}{n-1}}}$$

Дисперсійний аналіз (ANOVA)

На відміну від t-критерію **дисперсійний аналіз (ANOVA)** дозволяє порівнювати середні значення ***трьох і більше груп***.

В основі дисперсійного аналізу лежить припущення про те, що одні змінні можуть розглядатися як причини (фактори, незалежні змінні), а інші як слідства (залежні змінні).

Основна мета: дослідження значущості відмінності між середніми за допомогою порівняння дисперсій. Поділ загальної дисперсії на кілька джерел, дозволяє порівняти дисперсію, викликану відмінностями між групами, з дисперсією, викликаною внутрішньогруповою мінливістю.

Вибірки (3 і більше) можуть бути як рівними, так і нерівними за чисельністю, як пов'язаними, так і непов'язаними. За кількістю виявлених регульованих факторів дисперсійний аналіз може бути ***однофакторний*** (при цьому вивчається вплив одного фактора на результати експерименту), ***двофакторний*** (при вивченні впливу двох чинників) і ***багатофакторним*** (дозволяє оцінити не тільки вплив кожного з факторів окремо, але і їх взаємодію).

НЕПАРАМЕТРИЧНІ СТАТИСТИЧНІ МЕТОДИ

Непараметричні методи дозволяють обробляти дані "низької якості" з вибірок малого обсягу зі змінними, про розподіл яких мало що або взагалі нічого невідомо.

Непараметричні методи не ґрунтуються на оцінці параметрів (таких як середнє або стандартне відхилення) при описі вибіркового розподілу величини, що цікавить. Тому ці методи іноді також називаються вільними від параметрів або вільно розподіленими.

По суті, для кожного параметричного критерію є, принаймні, один непараметричний аналог. Ці критерії можна віднести до однієї з наступних груп:

А) критерії відмінності між незалежними вибірками

Б) критерії відмінності між залежними вибірками

В) критерії залежності між змінними

А) критерії відмінності між незалежними вибірками

Непараметричними альтернативами параметричного критерію для двох незалежних груп є:

1) U критерій Манна-Уїтні

2) Критерій серій Вальда-Вольфовиця
3) Двовибірковий критерій Колмогорова-Смірнова

Кілька незалежних груп: якщо ви маєте кілька груп, то можете використовувати Дисперсійний аналіз (ANOVA).

Його непараметричними аналогами є:

1) Рангові дисперсійний аналіз Краскела-Уолліса
2) медіанний тест

Б) критерії відмінності між залежними вибірками

Дві залежні вибірки: критерій Вілкоксона і ін.

Якщо ви хочете порівняти дві змінні, що відносяться до однієї і тієї ж вибірки (наприклад, математичні успіхи студентів на початку і в кінці семестру), то зазвичай використовується t-критерій для залежних вибірок.

Альтернативними непараметричними тестами є:

1) Критерій Вілкоксона парних порівнянь
2) Критерій знаків

Кілька залежних вибірок

Якщо розглядається більше двох змінних, що відносяться до однієї і тієї ж вибірки, то зазвичай використовується Дисперсійний аналіз (ANOVA) з повторними вимірами.

Альтернативним непараметричним методом є:

1) ранговий дисперсійний аналіз Фрідмана
2) Q критерій Кохрена

2) Кореляційний аналіз

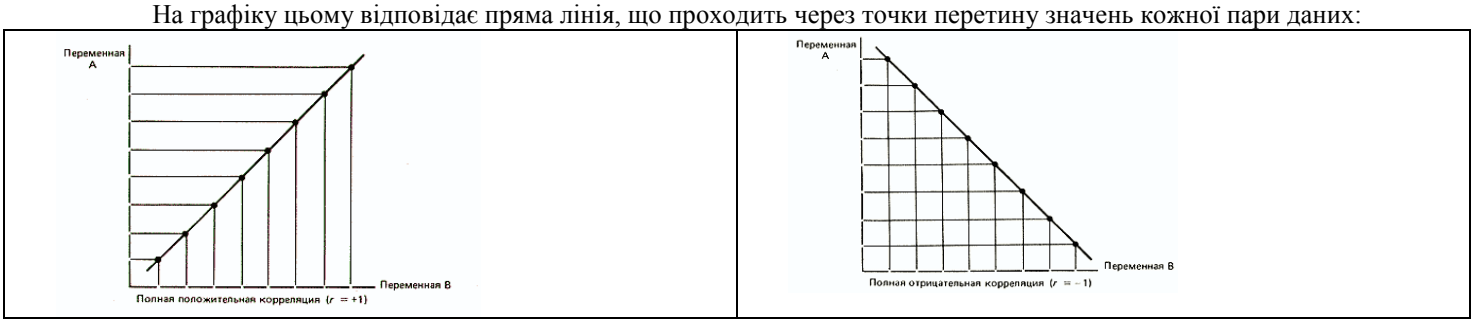
При вивченні кореляцій намагаються встановити, чи існує якийсь зв'язок між двома показниками в одній вибірці (наприклад, між зростанням і вагою дітей або між рівнем IQ і шкільною успішністю) або між двома різними вибірками (наприклад, при порівнянні пар близнюків), і якщо цей зв'язок існує, то чи супроводжується збільшення одного показника зростанням (позитивна кореляція) або зменшенням (негативна кореляція) іншого.

Можна використовувати два різні способи кореляційного аналізу: параметричний метод розрахунку коефіцієнта Браві-Пірсона (r) і обчислення коефіцієнта кореляції рангів Спірмена (r_s), який є непараметричним.

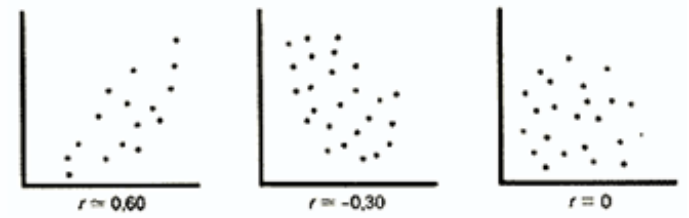
Коефіцієнт кореляції (позначається маленькою літерою **r**) і показує нам дві речі: 1) ступінь зв'язку двох змінних і 2) напрямок цього зв'язку (прямий або зворотний зв'язок).

Коефіцієнт кореляції - це величина, яка може варіювати в межах від +1 до -1. У разі повної позитивної кореляції цей коефіцієнт дорівнює плюс 1, а при повній негативній - мінус 1.

На графіку цьому відповідає пряма лінія, що проходить через точки перетину значень кожної пари даних:



У разі ж якщо ці точки не шикуються по прямій лінії, а утворюють «хмару», коефіцієнт кореляції за абсолютною величиною стає менше за одиницю і за мірою округлення цієї хмари наближається до нуля:



У разі якщо коефіцієнт кореляції дорівнює 0, обидві змінні повністю незалежні одна від одної.

При оцінці сили зв'язку коефіцієнтів кореляції використовується шкала Чеддока. При негативній кореляції значення сили зв'язку між змінними змінюють на протилежні. Таблиця аналізу сили зв'язку між змінними:

Значення	Інтерпретація
от 0 до 0,3	дуже слабкий
от 0,3 до 0,5	слабкий
от 0, 5 до 0,7	середній

от 0,7 до 0, 9	високий
от 0,9 до 1	дуже високий

Наприклад:

• якщо величина коефіцієнта кореляції між змінними дорівнює -0,36, то це слабка негативна кореляція, і швидше за все ми не будемо приймати її до уваги;

• якщо величина коефіцієнта кореляції дорівнює 0 - змінні не пов'язані між собою;

• якщо величина коефіцієнта кореляції між змінними дорівнює 0,25 то це дуже слабка кореляція і в більшості випадків ми не беремо її до уваги;

• якщо величина коефіцієнта кореляції між змінними дорівнює 0, 75 то це висока кореляція і в своїх інтерпретаціях нам варто звернути на неї увагу;

• якщо величина коефіцієнта кореляції дорівнює 1 – змінні повністю взаємопов'язані.

Однак для того, щоб можна було робити висновки про зв'язки між змінними, велике значення має обсяг вибірки: чим вибірка більше, тим вірогідніше величина отриманого коефіцієнта кореляції. Існують таблиці з критичними значеннями коефіцієнта кореляції Браве-Пірсона та Спірмена для різного числа ступенів свободи.

Коефіцієнт кореляції Браве-Пірсона (r) – це параметричний показник, для обчислення якого порівнюють середні і стандартні відхилення результатів двох вимірювань. При цьому використовують формулу:

$r = \frac{(\Sigma XY) - n\bar{X}\bar{Y}}{(n-1)s_X s_Y}$,	де ΣXY - сума добутків даних з кожної пари; n-число пар; X - середня для даних змінної X; Y - середня для даних змінної Y; Sx - стандартне відхилення для розподілу x; Sy - стандартне відхилення для розподілу y

Коефіцієнт кореляції рангів Спірмена (r_s) - це непараметричний показник, за допомогою якого намагаються виявити зв'язок між рангами відповідних величин в двох рядах вимірів. Цей коефіцієнт розраховувати простіше, проте результати виходять менш точними, ніж при використанні r.

4. Роль і значення графічного методу в статистиці

Графіком в статистиці називається умовне зображення статистичних даних у вигляді різних геометричних образів: точок, ліній, фігур, тощо. Головна перевага графіків - наочність.

Графіки в статистиці, як правило, використовуються для широкої популяризації даних і полегшення їх сприйняття неспеціалістами, тому в доповідях, промовах і повідомленнях використання статистичних даних часто здійснюється за допомогою графіків. Графіки широко використовуються для узагальнення і аналізу статистичних даних.

Загальні правила побудови графічного зображення

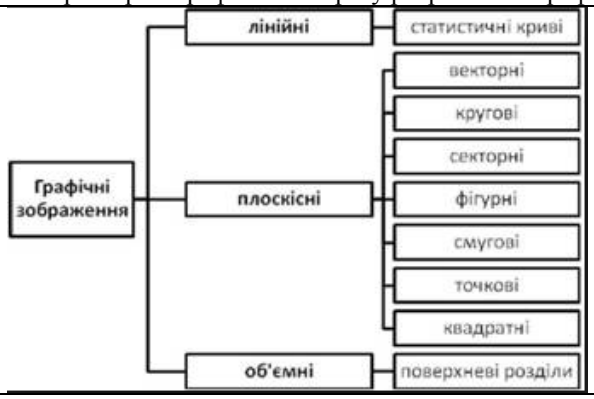
При побудові графіка важливо знайти такі способи зображення, які найкращим чином відповідають змісту і логічній природі зображуваних показників.

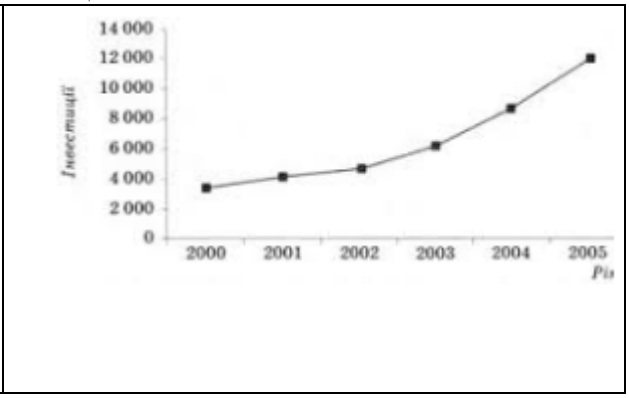
Кожен графік складається з **графічного образу** і **допоміжних елементів.**

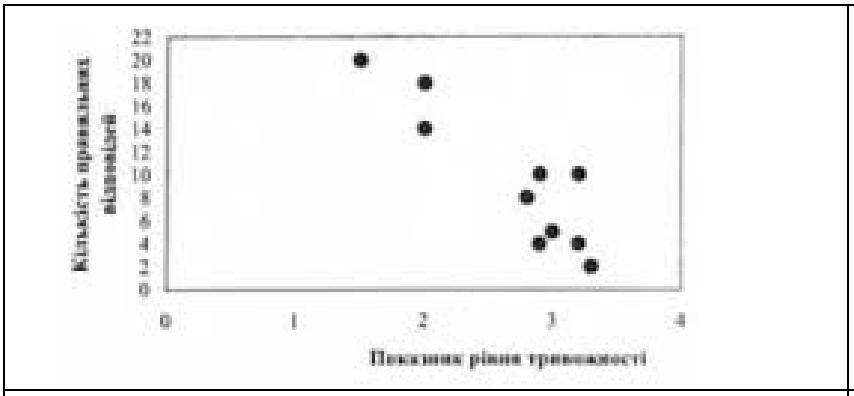
Графічний образ (основа графіка) - це геометричні знаки, тобто сукупність точок, ліній, фігур, за допомогою яких зображуються статистичні показники. Важливо правильно вибрати графічний образ, який повинен відповідати меті графіка і сприяти найбільшій виразності зображуваних статистичних даних. Графічний образ може являти собою ряд стовпчиків або квадратів і т.п.

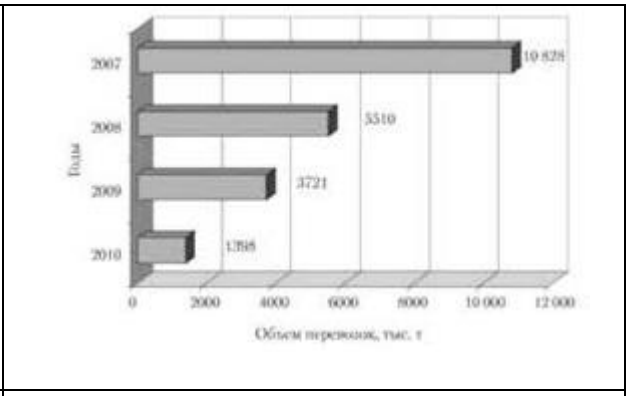
Допоміжні елементи уможливлюють читання графіка, його розуміння і використання. До них відносяться: заголовок графіка, підписи, пояснення.

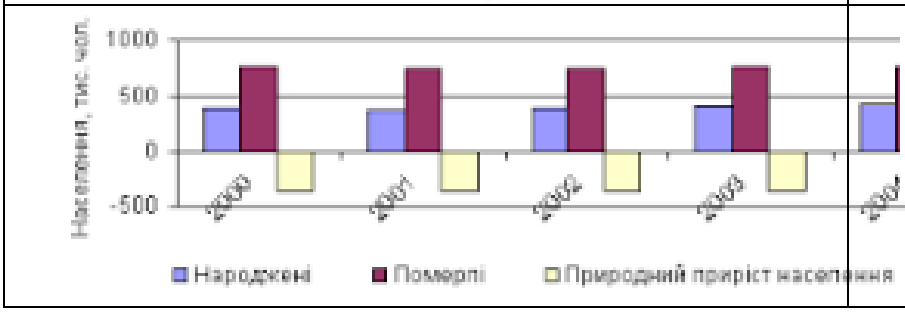
За характером графічного образу розрізняють графіки об'ємні, лінійні і площинні

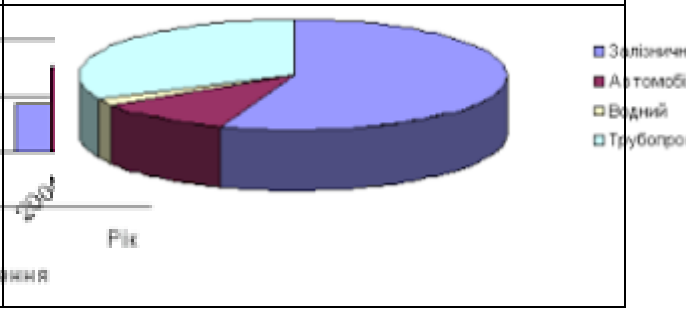





Кількість правильних відповідей
Показник рівня тривожності


2007
2008
2009
2010
10 525
5510
3721
1705
0
2000
4000
6000
8000
10 000
12 000


Населення, тис. чол.
1000
500
0
-500
Народжені
Померлі
Природний приріст населення


Рік