

Лекція 2

Тема:

Зведення та групування даних

План

1. Суть біометричного зведення
2. Класифікації та групування
3. Принципи формування груп
4. Варіаційний ряд

Суть біометричного зведення

Перехід від одиничного до загального відбувається завдяки зведенню.

Суть зведення полягає в тому, що матеріали спостереження класифікують та агрегують. Елементи сукупності за певними ознаками об'єднують у групи, класи, типи, а інформацію про них агрегують як у межах груп, так і в цілому по сукупності.

Основне завдання зведення — виявити типові риси та закономірності масових явищ чи процесів. Зведення є основою подальшого аналізу інформації.

В результаті проведення спостережень над біологічними об'єктами отримують велику кількість фактичного матеріалу. Перш за все його потрібно впорядкувати з метою отримання необхідної інформації.

Спостереження – це збір первинних даних про об'єкт, який підлягає вивченню.

Кожне окреме спостереження називають **варіантою** або **датою**.

Групування – це процес систематизації результатів спостережень для отримання закладеної в них інформації, а також виявлення закономірностей, які властиві досліджуваним ознакам.

Види статистичного зведення

за складністю
побудови

за організацією
роботи

за технікою
зведення

періодичне
зведення

централізоване
зведення

автоматизоване
зведення

групове
зведення

деперсоналізоване
зведення

анонімізоване
зведення

Просте зведення – це лише простий підрахунок підсумків первинного статистичного матеріалу без будь-якої його систематизації. Так як просте зведення не готує матеріал для глибокого економічного аналізу, воно має обмежене застосування у статистичній практиці.

Групове (складне) зведення – це складна операція наукової обробки первинного матеріалу, яка включає декілька етапів роботи:

- 1) групування матеріалу за окремою або окремими ознаками;
- 2) розробка системи показників для характеристики типових груп і підгруп;
- 3) підбиття групових та загальних підсумків;
- 4) відображення результатів зведення у вигляді таблиць.

У разі централізованого зведення весь первинний матеріал спостереження зосереджується, систематизується та узагальнюється в центральному органі державної статистики – Державному Комітеті статистики України.

Децентралізоване зведення передбачає узагальнення матеріалу знизу до гори за ієрархічними сходами управління з відповідною обробкою на кожній із них, а Держкомстату України надсилаються уже зведені підсумки за певними адміністративними одиницями.

Автоматизоване зведення проводиться на автоматизованих робочих місцях за допомогою комп'ютерів.

Ручне зведення – це обробка первинних матеріалів ручним способом.

Класифікації та групування

Поділ сукупностей на групи, однорідні в тому чи іншому розумінні, пов'язаний з такими діями, як **систематизація, типологія, класифікація, групування**.

Традиційно зазначений поділ виконують за такою схемою: із множини ознак, які описують явище, добирають розмежувальні, а потім сукупність поділяють на групи та підгрупи відповідно до значень цих ознак.

**Головний принцип будь-якого поділу ґрунтується на
двох положеннях:**

- 1) в один клас, групу об'єднуються елементи певною мірою подібні між собою;
- 2) ступінь подібності між елементами, які належать до одного класу, значно вищий, ніж між елементами, що належать до різних класів.

**У кожному конкретному дослідженні вирішуються три
питання:**

- 1) що взяти за основу групування;
- 2) скільки груп, позицій необхідно відокремити;
- 3) як розмежувати групи.

Основою групування може бути будь-яка атрибутивна чи кількісна ознака, що має градації. Таку ознаку називають **групувальною**.

Залежно від складності масового явища (процесу) та мети дослідження групувальних ознак може бути одна, дві й більше.

У практиці часто вдаються до розбиття сукупностей за атрибутивними ознаками — **класифікації та номенклатури**.

Здебільшого йдеться про багатоступеневу класифікацію з докладною номенклатурою груп і підгруп, із чітко визначеними вимогами та умовами віднесення елементів сукупності до тієї чи іншої групи. Кожній класифікаційній позиції надається код (шифр), який замінює її назву і є постійним засобом ідентифікації під час передавання інформації по каналах зв'язку, комп'ютерної обробки тощо. Кожна класифікація є сталою, забезпечуючи порівнянність даних у просторі та часі.

Поряд з класифікацією для висвітлення певних аспектів конкретного дослідження використовують групування,

Статистичне групування — утворення однорідних груп одиниць сукупності за певною істотною ознакою, а також тих, що мають однакові або близькі значення групувальної ознаки, яку взято за основу утворення груп у процесі групування.

Статистичне угруповання виконує такі аналітичні функції:

- 1) вивчення структури та структурних зрушень;
- 2) визначення типів явищ, виокремлення однорідних груп і підгруп;
- 3) виявлення взаємозв'язків між ознаками.

Згідно з цими функціями розрізняють три види групувань: **структурне, типологічне, аналітичне.**

Структурне групування характеризує склад однорідної сукупності за певними ознаками.

Типологічне групування — це поділ якісно неоднорідної сукупності на класи, типи, однорідні групи. Основне завдання такого групування — ідентифікація типів.

Вибір групувальної ознаки та кількісних міжгрупових меж ґрунтується на всебічному теоретичному аналізі суті явища, його характерних рис та особливостей формування в конкретних умовах часу та простору.

Скориставшись групуванням, можна також виявити наявність та напрям зв'язку між ознаками, з яких одна розглядається як результат, інша — як фактор, що впливає на результат.

Висновок про наявність зв'язку можна зробити на підставі комбінаційного поділу за цими ознаками згідно з характером розміщення частот. Якщо результативна ознака кількісна, для кожної групи за факторною ознакою можна визначити середнє значення результативної ознаки.

За наявності зв'язку між ознаками групові середні результативної ознаки систематично змінюються від групи до групи (збільшуються чи зменшуються). Таке групування називається **аналітичним**.

Очевидно, аналітичне групування докладніше й виразніше, ніж комбінаційний поділ, описує зв'язок між ознаками. Зауважимо, що поділ групувань на три види певною мірою відносний. Адже часто групування універсальні: одночасно виділяються типи, визначається склад сукупності й виявляється взаємозв'язок між ознаками

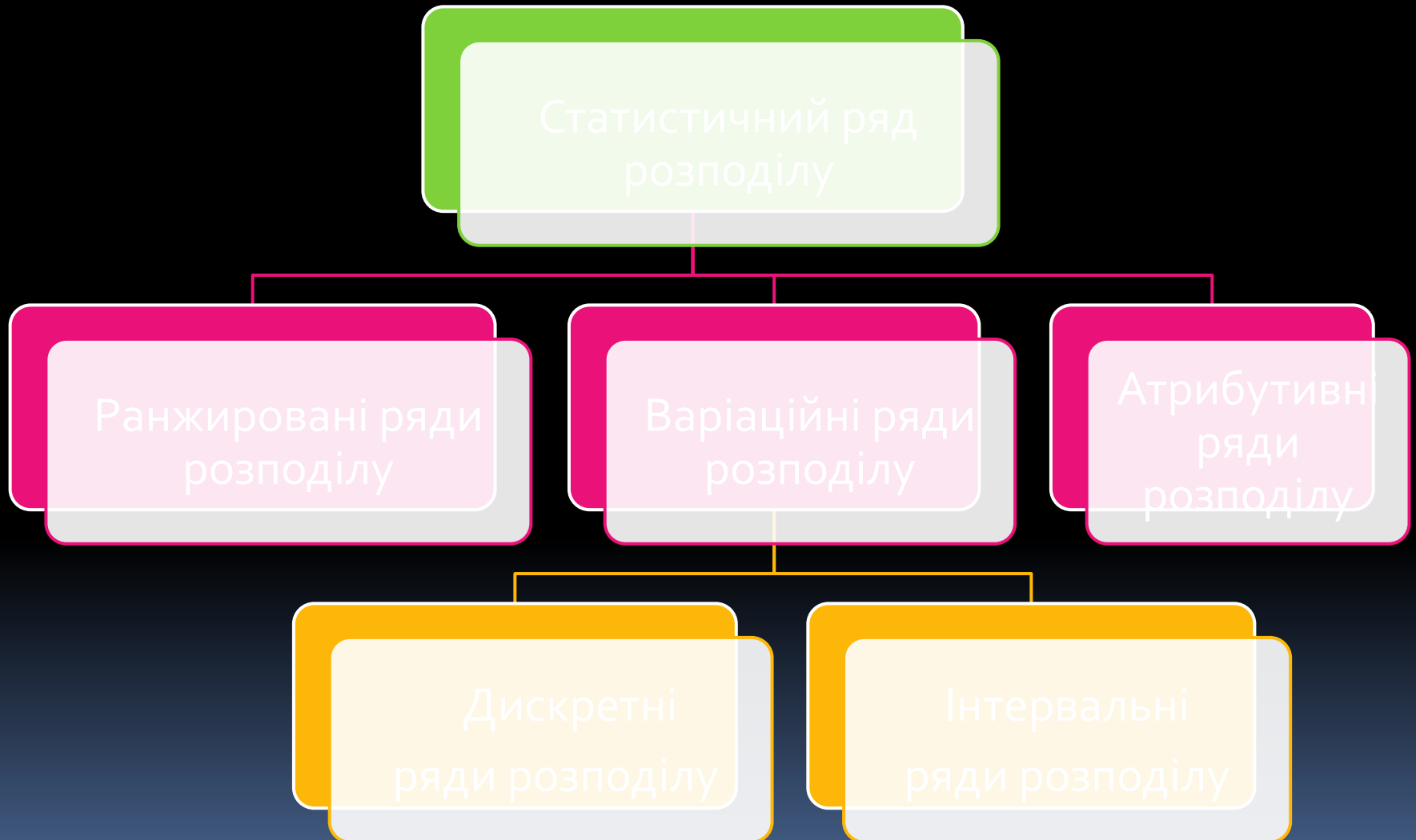
Принципи формування груп. Варіаційний ряд

Статистична сукупність формується під впливом причин та умов, з одного боку – типових, спільних для всіх елементів сукупності, а з іншого – випадкових, індивідуальних. Ці чинники взаємозв'язані, а їх спільна взаємодія визначає як індивідуальні значення ознак, так і розподіл останніх у межах сукупності. Характерні властивості структури статистичної сукупності відбиваються в рядах розподілу.

Англійський вчений **У. Дж. Рейхман** з приводу неупорядкованих сукупностей образно сказав, що *зіткнутися з масою неупорядкованих даних рівнозначно ситуації коли людину кидають у лісових хащах без компасу.*

Статистичний ряд розподілу

– це впорядковані статистичні сукупності.



Найпростішим видом статистичного ряду розподілу є **ранжирований ряд**, тобто ряд чисел, що знаходиться в порядку зростання чи спадання варіюючої ознаки. Вони будуються на основі ранжирування, тобто шляхом розміщення варіант в порядку зростання або спадання їх значення.

Наприклад, маємо 5 вимірів:

2,7,5,4,6.

Як бачимо, величина ознаки варіює від 2 до 7.

Розміщуємо цю сукупність в зростаючому порядку

2,4,5,6,7.

Це і є **ранжирований** ряд розподілу ознак.

Такий ряд не дозволяє судити про закономірності, закладені в розподілених даних: біля якої величини групується більшість показників; які є відхилення від цієї величини; яка загальна картина розподілу. З цією метою групують дані, показуючи, як часто зустрічаються окремі спостереження в загальному їх числі.

Розподіл одиниць сукупності за ознаками, що не мають кількісного виразу, називається **атрибутивним рядом** (наприклад, розподіл господарств за їх виробничим напрямом).

Ряди розподілу одиниць сукупності за ознаками, що мають кількісний вираз, називаються **варіаційними рядами**.

У таких рядах значення ознаки (варіанти) знаходяться в порядку зростання чи спадання. Кожне значення може декілька разів повторюватися і тоді вказаний ряд буде мати подвійний вигляд:

показники ознаки: 2,4,5,6,7

частота ознаки 3,1,2,4,2

Такий упорядкований ряд розподілу варіант, в якому вказана і їх повторюваність, називається і варіаційним рядом.

У варіаційному ряді розподілу розрізняють два елементи: **варіанта і частота.**

Варіанта – це окреме значення групувальної ознаки, **частота** – число, яке показує, скільки разів зустрічається кожна варіанта.

У біометрії обчислюється ще один елемент варіаційного ряду – **частість.**

Остання визначається, як відношення частоти випадків даного інтервалу до загальної суми частот.

Частість визначається в частках одиниці, відсотках (%) в проміле (‰).

Таким чином, **варіаційний ряд розподілу** – це подвійний ряд чисел, у якому варіанти розташовані в порядку зростання або спадання, вказані їх частоти або частоти.

Варіаційні ряди бувають **дискретні** (переривні) і **інтервальні** (непереривні).

Дискретні варіаційні ряди – це такі ряди розподілу, в яких варіанта як величина кількісної ознаки може приймати тільки певне значення. Варіанти різняться між собою на одну чи кілька одиниць.

При великій кількості варіант варіаційні ряди можуть набувати двох видів: безінтервальні та інтервальні.

Безінтервальний ряд – це ряд конкретних варіант з відповідною частотою і дискретністю показників без будь-якого інтервалу між ними.

Наприклад:

V – 6, 7, 8, 9, 10, 11, 12, 13, 14

f – 5, 10, 17, 20, 14, 9, 8, 6, 11

де V – варіанти, f – частота.

Однак, значна більшість ознак варіюють в широких межах і тому їх розподіляють в **інтервальні варіаційні ряди**.

Інтервальні (непереривні) варіаційні ряди – такі ряди розподілу, в яких значення варіанти дано у вигляді інтервалів, тобто значення ознак можуть відрізнятися одне від одного на скільки завгодно малу величину.

При побудові варіаційного ряду непереривної ознаки неможливо вказати кожне значення варіанти, тому сукупність розподіляється за інтервалами. Останні можуть бути рівні і нерівні. Для кожного з них вказуються частоти або частості.

Техніка складання таких рядів полягає в тому, що вся варіабельність (мінливість) ознаки від мінімуму до максимуму розподіляються на рівні інтервали, тобто проміжки, які називають **класами**.

кількість класів визначають за формулою:

$$k = 1 + 3,32lg \times n$$

n – величина сукупності.

Для визначення центральних значень всіх інших класів користуються формулою:

$$w_i = w_e + i$$

де w_i – центральне значення потрібного класу;

w_e – центральне значення попереднього класу.

Особливість центральних класів полягає в тому, що різниця між сусідніми класами точно дорівнює величині класового інтервалу.

Наступна операція полягає у встановленні частоти варіантів в кожному класі – f .

Сума частот варіаційного ряду повинна дорівнювати сумі варіант, що взяті для обрахунку.

Графічне зображення рядів розподілу (як і статистичних даних взагалі), крім досягнення наочності, переслідує й аналітичну мету. Графік дозволяє в найбільш простій і доступній формі піддати аналізу (візуально) статистичний ряд розподілу.

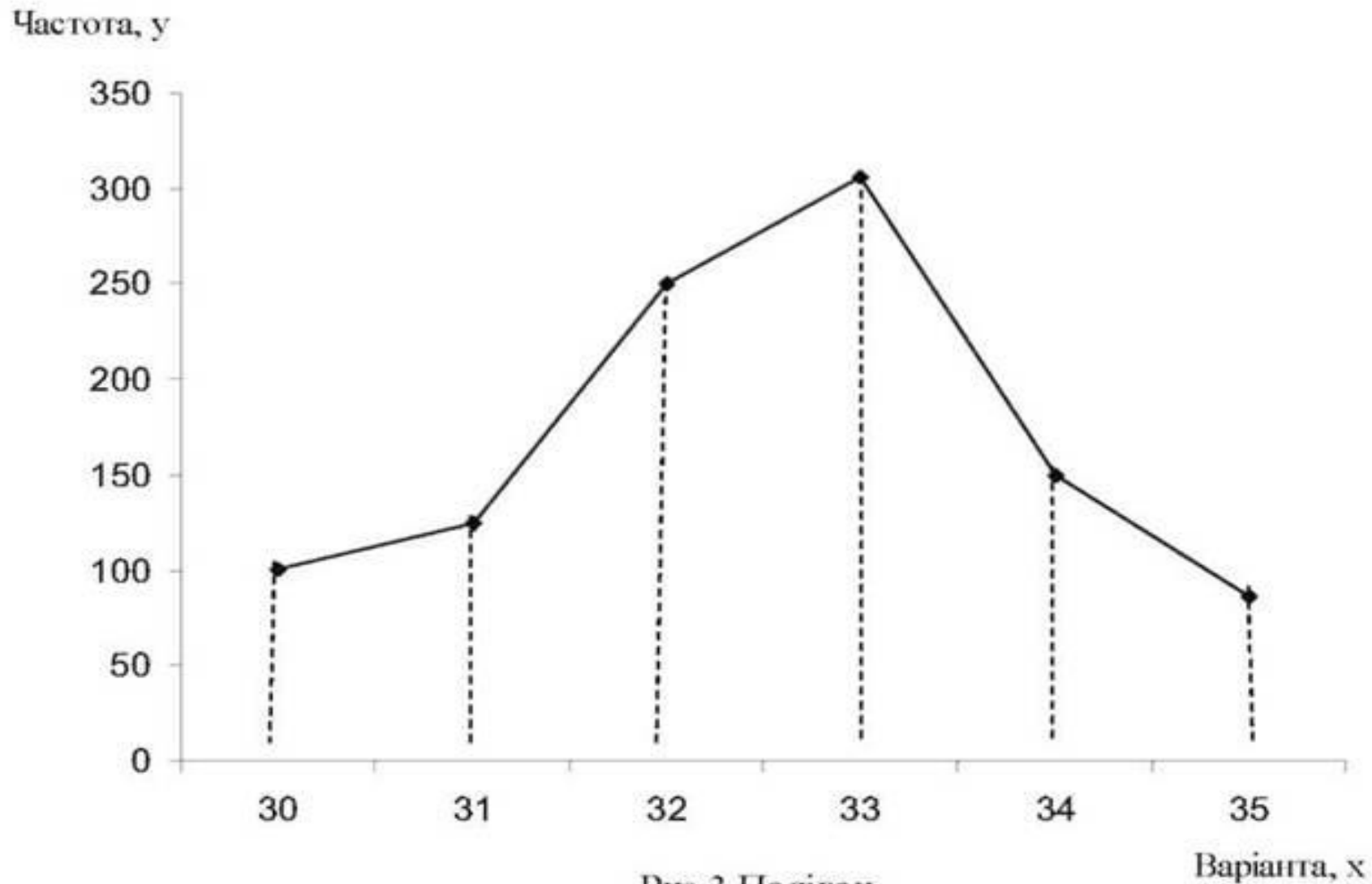
Варіаційні ряди залежно від виду і поставленої задачі їх аналізу графічно можуть бути зображені у виді **полігону, гістограми, кумуляти, огіви**.

Полігон розподілу будується в прямокутній системі координат, при цьому на осі абсцис відкладається варіанта, а на осі ординат – частота. За допомогою полігону розподілу, як правило, графічно зображуються дискретні варіаційні ряди. При побудові полігону для інтервальних рядів розподілу ордината, яка відповідає частоті (частоті) встановлюється перпендикулярно осі абсцис у точці, що відповідає центру інтервалу.

В такій побудові висота кожного класу буде пропорційна його частоті. А тому на перетині перпендикуляра, котрий помічає значення класу і ординати, яка відмічає частоту класу ставлять точку. Подібне роблять з другим класом, третім і т.д. Після цього точки послідовно поєднують плавною лінією, в результаті чого утворюється варіаційна крива або полігон розподілу. При цьому слід пам'ятати, що при побудові полігону необхідно завжди розпочинати і закінчувати проведення лінії нульовими класами, тобто як початок, так і кінець лінії повинні знаходитися на вісі абсцис.

Розміщення класів і їх частот на вісях абсцис і ординат повинно бути виконано в одному масштабі, тобто відстань між позначеннями класів, частот повинна бути однаковою для всіх, які будуть розміщені на графіку. Проте кожна вісь може мати свій масштаб.

Полігон



Такий спосіб зображення інтервального ряду називають способом **"навантажених ординат"**.

Він заснований на допущенні умови рівномірного розподілу частот у межах інтервалів. Якщо це так, то частоти можна віднести до конкретного значення варіанти, яке знаходиться в центрі інтервалу.

За таких умов середина (центр) інтервалу ніби **"навантажується"**.

Однак, графічне зображення варіаційного ряду за допомогою полігону не дає повного уявлення про число варіант в класі. Краще це відображає **гістограма**.

Гістограма являє собою ступінчасту фігуру у вигляді прямокутників, що примикають один до одного.

Порядок побудови цього виду графіків такий. На осі абсцис відкладають інтервали значень варіанти. Вони ж є основами прямокутників, висота яких (ордината) пропорційна частоті інтервалів.

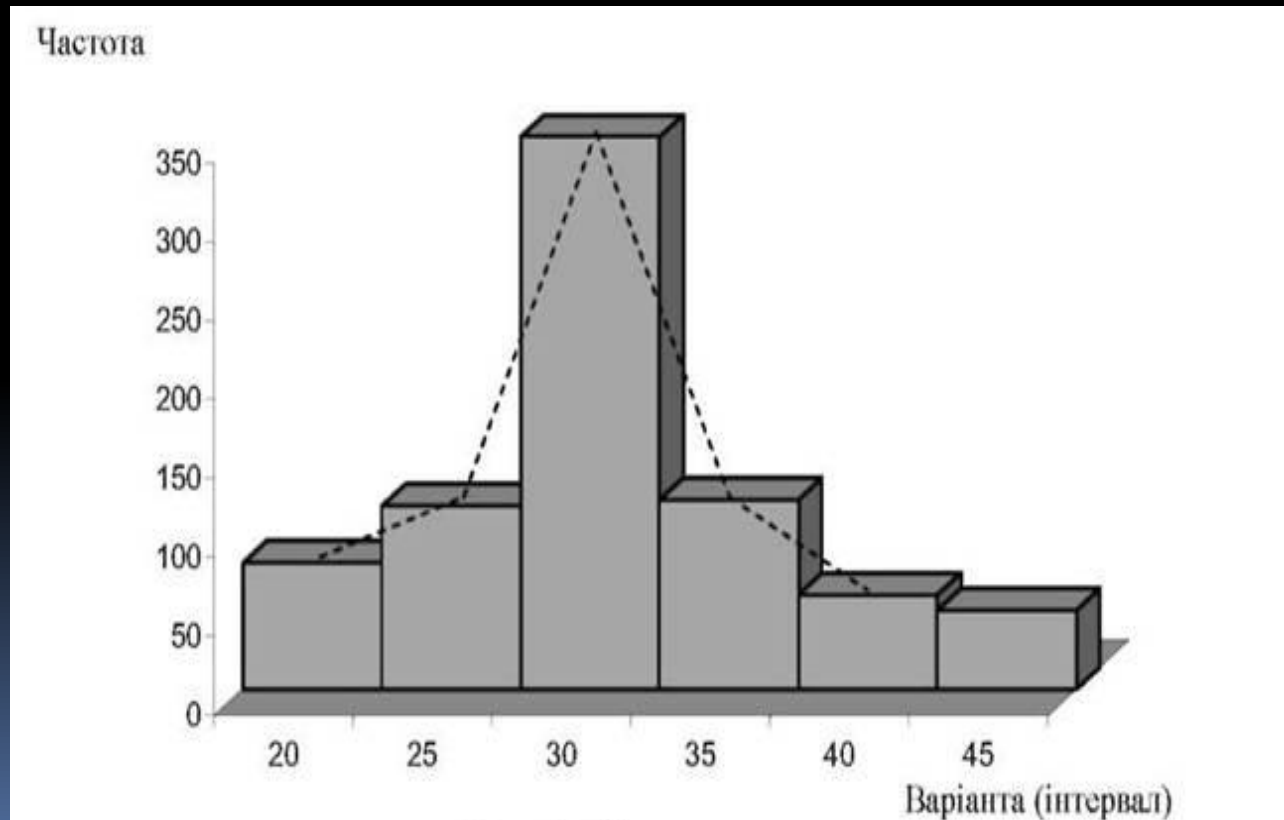


Рис. 4. Гістограма

Якщо необхідно зобразити на графіку інтервальний ряд розподілу з нерівними інтервалами, то гістограму будують не по частотах інтервалів, а по показниках щільностей розподілу.

При побудові гістограми по абсолютній щільності розподілу загальна площа її дорівнюватиме чисельності сукупності. При побудові графіка відносної щільності площа гістограми дорівнюватиме одиниці.

Інколи ідуть шляхом розчленування (подрібнення) інтервалів варіаційного ряду. У цьому випадку на графіку площа гістограми (дрібно ступінчастої) повинна відповідати попередній величині. Якщо процес подрібнення інтервалів продовжити, одержуючи все дрібніші їх параметри, то дрібноступінчаста гістограма в межі являтиме плавну криву.

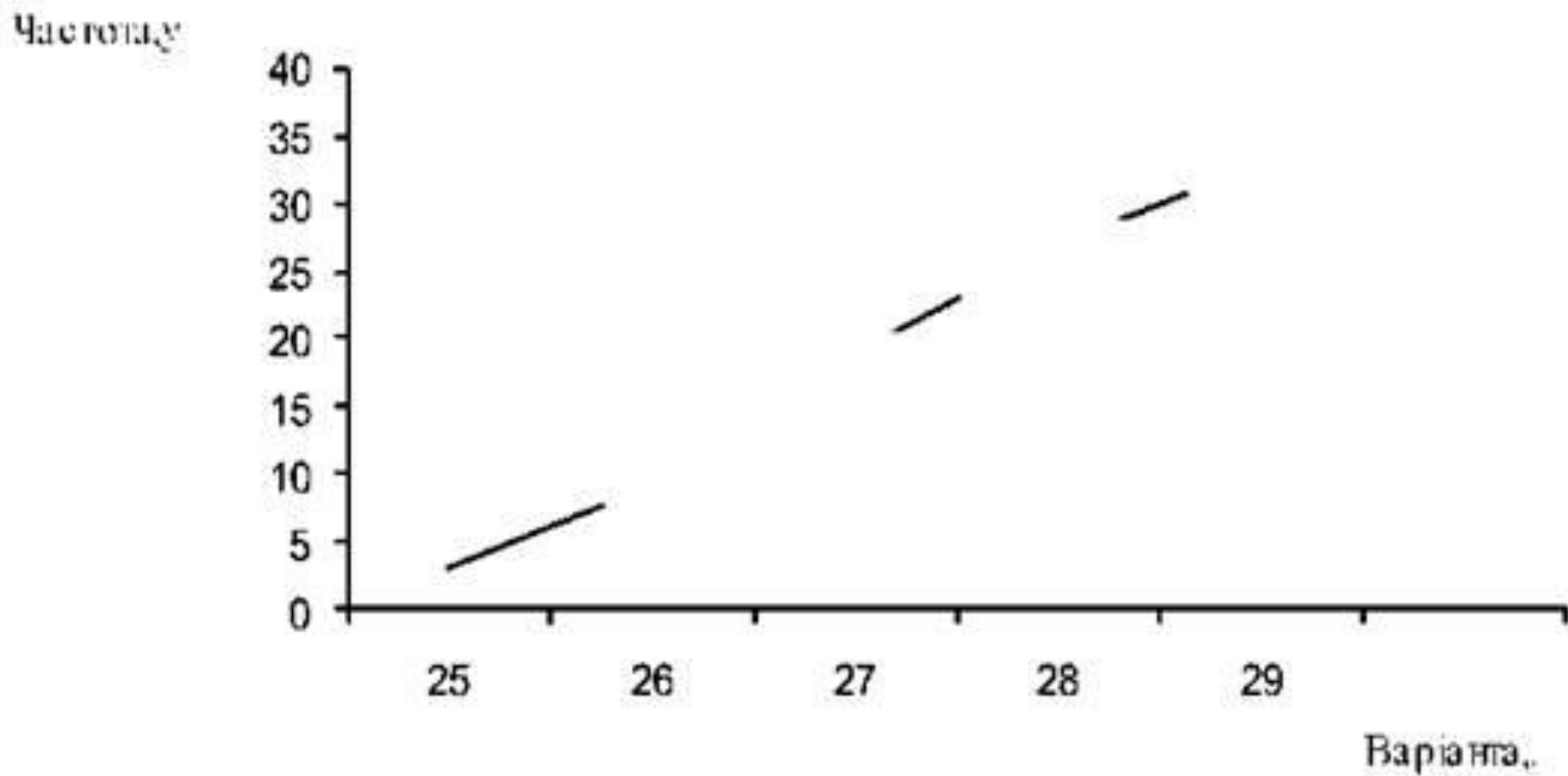
При зображенні варіаційного ряду з нагромадженими частотами у прямокутній системі координат одержується так звана крива сум – **кумулята**.

Якщо розподіл носить дискретний характер (переривний) на графіку на осі абсцис відкладають значення варіанти, на осі ординат – накопичені частоти.

При побудові кумуляти для інтервального ряду розподілу будують точки, абсциси яких – праві границі інтервалів, а ординати - відповідні їм нагромаджені частоти або частоти.

По інтервальному (неперервному) ряду розподілу будують точки, абсциси яких – праві границі інтервалів, а ординати – частоти, що їм відповідають.

Кумулята



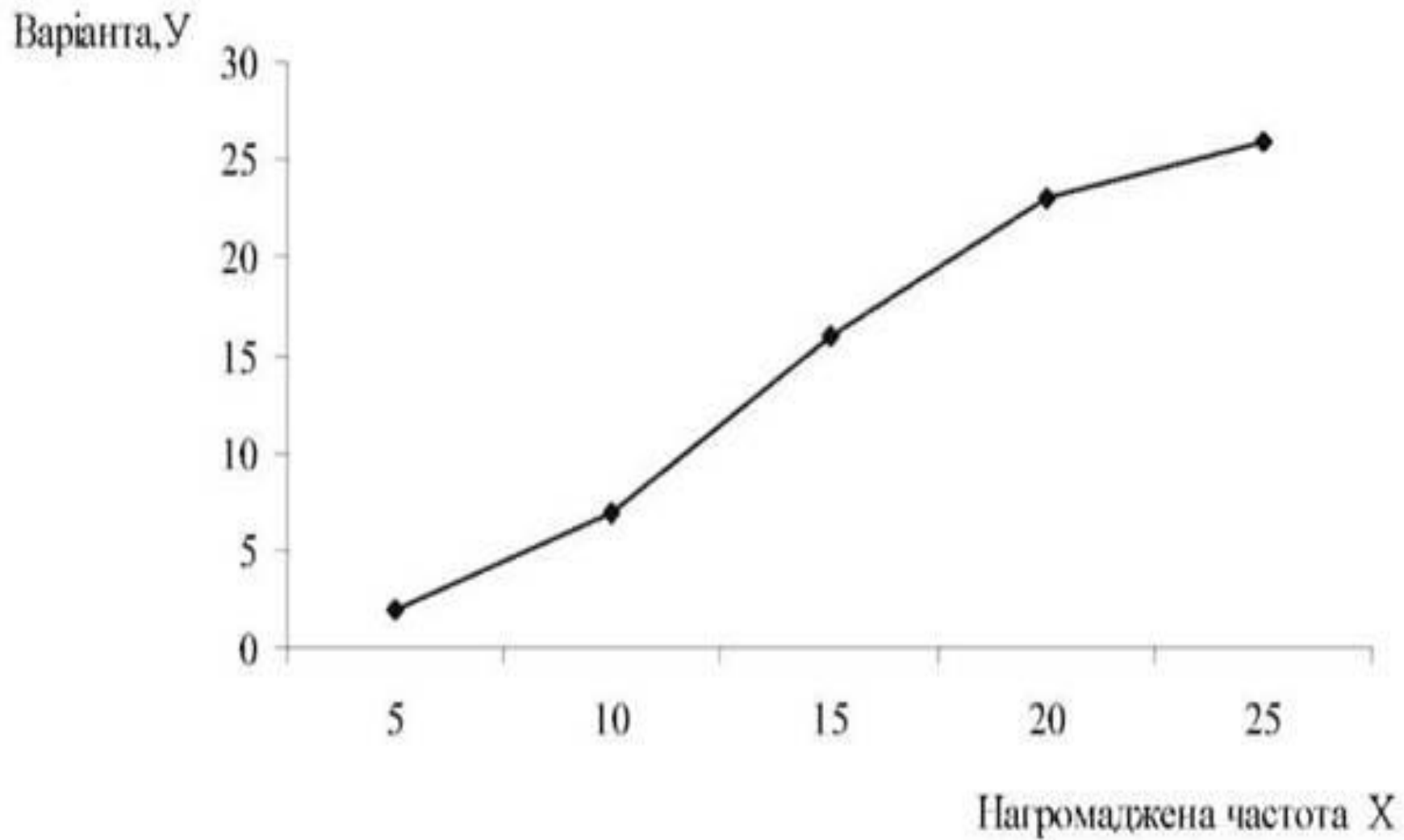
Аналогічно кумуляті в прямокутній системі координат будують **огіву**.

Різниця графіка лише в тому, що на осі абсцис наносять нагромаджені частоти, а на осі ординат – значення варіант. Щоб побачити форму огіву на графіку кумуляти, досить лист паперу з її зображенням повернути на 90° і подивитися на нього з протилежного боку.

Якщо збільшити до нескінченності число одиниць спостережень і зменшити величину інтервалу, графік ряду розподілу буде мати форму кривої.

Огіва застосовується зрідка, оскільки не дає повної характеристики варіаційного ряду, а має порівняльне значення, бо вказує на ступінь зростання частот.

Огіва



Форми статистичних рядів розподілів

Форми статистичних рядів розподілу



Схема форм статистичних розподілів

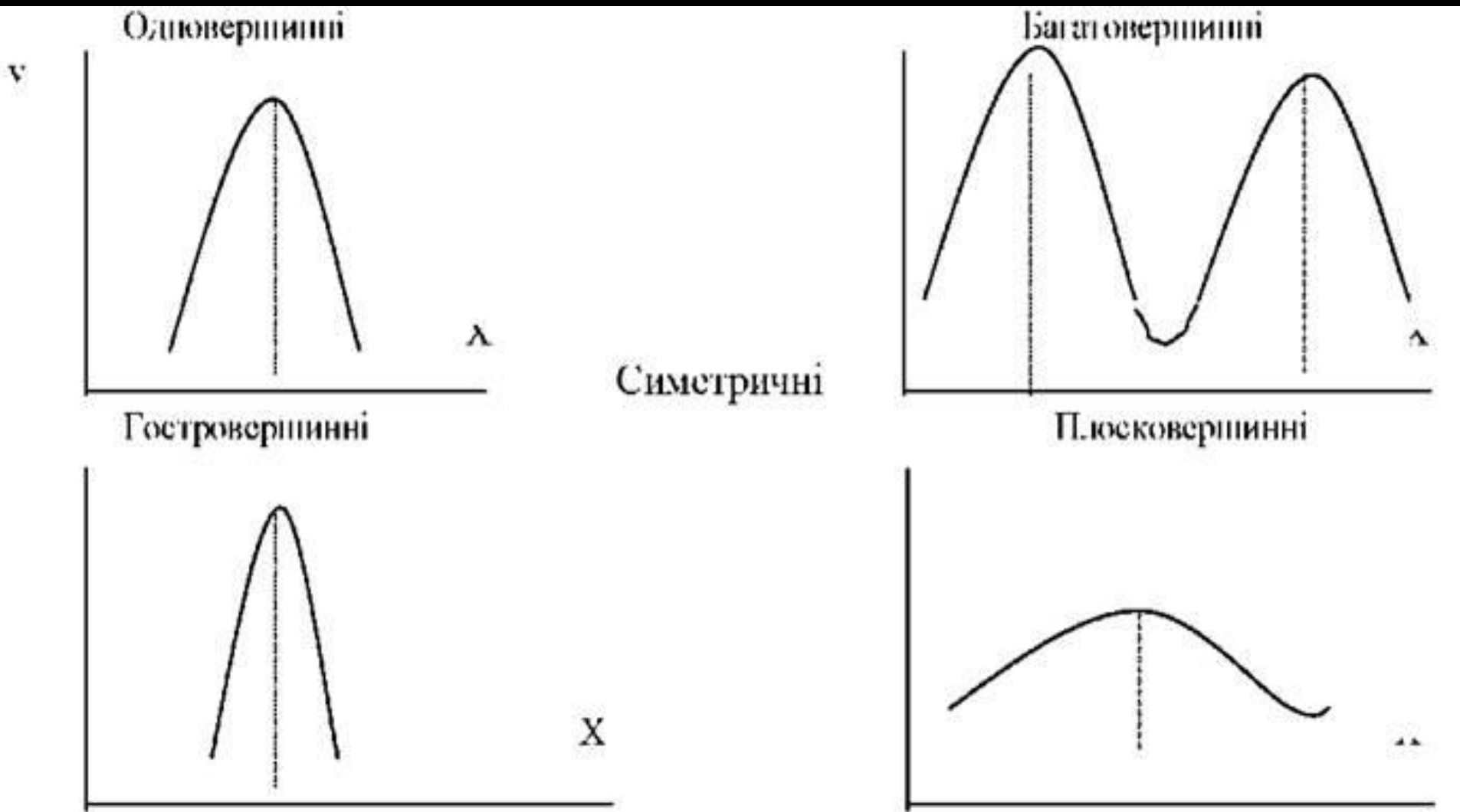
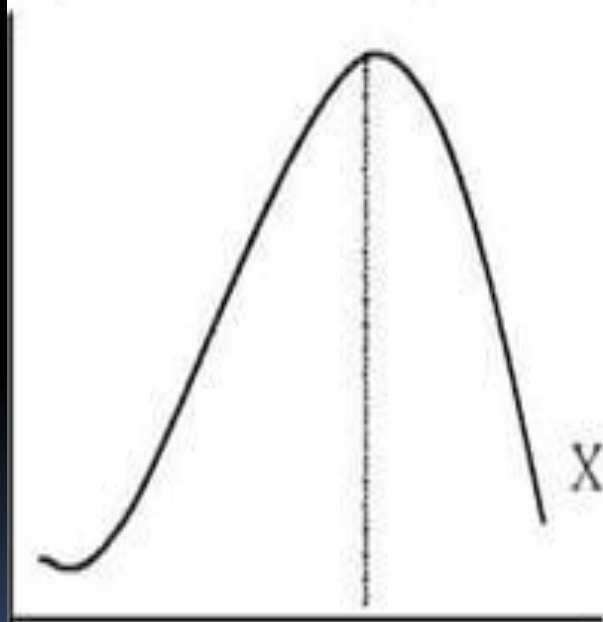


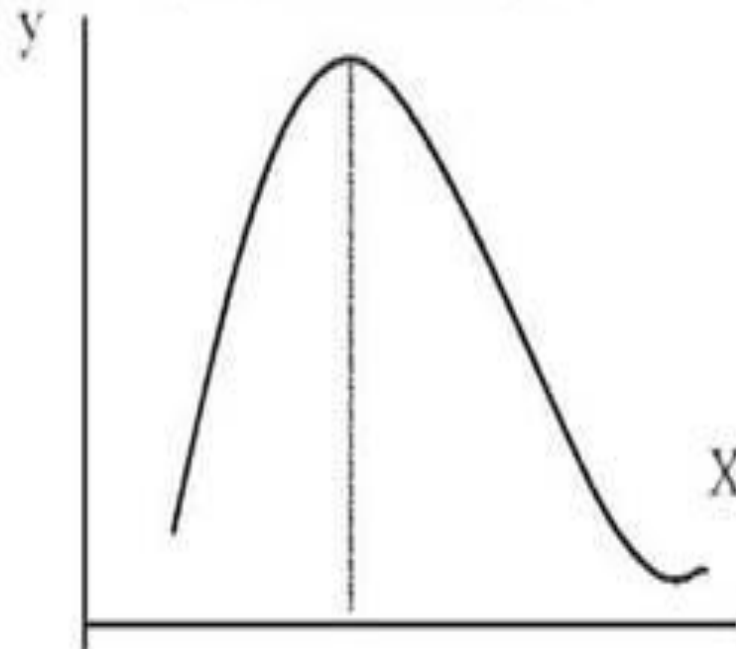
Схема форм статистичних розподілів

Помірно асиметричні

Лівостороння скошеність
(від'ємна асиметрія)



Правостороння скошеність
(додатня асиметрія)



Графіки форм статистичних розподілів

