Лекція 4.

ВАЖЛИВІ ЗАКОНИ РОЗПОДІЛУ

План

1. Формування і види рядів статистичних даних. Закономірність розподілу.
2. Аналіз рядів розподілу, характеристики: центра, розміру та ступеня варіації, форми розподілу.
3. Нормальний розподіл. Графічне зображення рядів розподілу.

1. Складовою частиною зведеної обробки даних статистичного спостереження є побудова і формування **рядів розподілу** з метою виявлення основних властивостей та закономірностей досліджуваної сукупності.

Ряд розподілу – це групування одиниць сукупності за однією певною ознакою. Характеризується двома величинами – варіантом і частотою:

- **варіант** – окреме значення групувальної ознаки;
- **частота** – число, що показує як часто варіант зустрічається в ряді.

Сума усіх частот визначає чисельність сукупності та її обсяг. Замість частот іноді зручніше вживати **частку** – відносну частоту, що виражена в долях одиниці (коефіцієнтом) чи відсотком до підсумку. Відповідно сума часток дорівнює 1-ці чи 100%. Абсолютна чисельність групи (частота) та відносна частота (частка) – це **частотні характеристики** утвореного ряду розподілу.

У співвідношенні варіант і частот проявляється **закономірність розподілу**. А саме: в деяких рядах із збільшенням значення варіюючої ознаки частоти спочатку збільшуються, а після досягнення max величини в середині ряду зменшуються. Це вказує на те, що частоти змінюються закономірно у зв′язку зі зміною варіюючої ознаки. Такі закономірності змін частот і називають закономірностями розподілу.

Закономірності розподілу найбільш яскраво проявляються тільки у масових спостереженнях.

Ряди розподілу поділяються на атрибутивні і варіаційні.

Атрибутивний ряд утворюють за **якісною ознакою**. Варіанти розташовують у їх логічній послідовності. Прикладом такого ряду є розподіл населення за статтю, характером праці, освітою, національністю.

Варіаційний ряд утворюється групуванням за **кількісною ознакою**. Варіанти такого ряду упорядковують за зростанням (або спаданням) їх значень – ранжують. Це дозволяє досить легко визначити min та max значення ознаки, відстань між ними, варіанти, що найчастіше повторюються, розділити дані по групах.

За характером варіації кількісні ознаки поділяються на:

- **дискретні ознаки** – це дані, які виражені тільки цілими числами без проміжних значень, наприклад, число дітей в сім’ї, кількість тварин у групі;
- **неперервні ознаки** – можуть приймати у визначених межах будь-яке значення, не тільки ціле, а і дробове, наприклад, маса тіла у кілограмах і грамах;

Варіаційні ряди, побудовані за цими ознаками, називаються відповідно **дискретними** чи **інтервальними рядами**.

У тих випадках, коли число варіантів дискретної ознаки досить велике, для неї будують інтервальний ряд.

Аналіз закономірностей розподілу передбачає:

 визначення типового рівня ознаки, який є центром тяжіння розподілу;

 вимірювання варіації ознаки – ступеня згрупованості індивідуальних значень ознаки навколо центра розподілу;

 оцінку особливостей варіації, ступеня її відхилення від симетрії.

Базою аналізу закономірностей розподілу слугує варіаційний ряд – дискретний або інтервальний з рівними інтервалами.

2. **Ряд розподілу** може бути охарактеризований системою характеристик (статистичних оцінок, показників), серед яких розрізняють:

 характеристики центра групування,

 характеристики варіації,

 характеристики форми розподілу.

Центром розподілу називається значення ознаки, навколо якого групуються інші варіанти. До **характеристик центра** розподілу належать **середня, мода** і **медіана**.

Основною характеристикою центра розподілу вважається середня, яка спирається на усю інформацію про досліджувану сукупність одиниць. Однак у деяких випадках середня повинна бути доповнена чи навіть замінена модальним значенням або медіаною. Наприклад, під час контролю якості зручніше користуватися медіаною, оскільки визначення медіани для ранжируваного ряду даних не потребує спеціального розрахунку. Окрім того, на відміну від середньої, вона не чутлива до крайніх значень взятої контрольної проби. В рядах з відкритими інтервалами (до..., від...) також доцільніше використовувати моду та медіану.

Середня, мода і медіана для якісно однорідної сукупності незначно відрізняються одна від одної. У симетричних розподілах $\mathbf{x_{сер} = Mo = Me}$.

До **характеристик варіації**, що показують як дані розподілені навкруг середньої, належать:

- **розмах варіації,**
- **середні арифметичне та квадратичне (стандартне) відхилення,**
- **дисперсія** – середній квадрат відхилень,
- **коефіцієнти варіації**.

Дисперсію та стандартне відхилення найчастіше застосовують у статистичній практиці, оскільки вони входять до більшості теорем теорії ймовірностей, що слугують фундаментом математичної статистики. Крім того, дисперсію можна розкласти на складові, які дозволяють оцінити вплив різних факторів, що зумовлюють варіацію ознаки.

Між абсолютними показниками варіації існують такі співвідношення:

$\mathbf{\sigma > d}$ (правило мажорантності), $\sigma = \mathbf{1{,}25d}$, $\mathbf{R = 6}\sigma$.

В умовах нормального розподілу, якщо обсяг сукупності досить великий, встановлена залежність між стандартним відхиленням та кількістю спостережень:

 в межах $\mathbf{x_{сер} \pm 1}\sigma$ розташовано **68,3%** кількості спостережень (членів ряду);

 в межах $\mathbf{x_{сер} \pm 2}\sigma$ – **95,4%** спостережень;

 в межах $\mathbf{x_{сер} \pm 3}\sigma$ – **99,7%** спостережень.

На практиці майже не зустрічаються відхилення, що перевищують $\pm\mathbf{3}\sigma$. Відхилення $\mathbf{3}\sigma$ вважається максимально можливим. Таке положення називають **“правилом 3-х сигм”**.

Відносні показники варіації застосовують для визначення однорідності (при $V \leq 33\%$) досліджуваної сукупності.

До **характеристик форми** розподілу, які ілюструють ступінь скошеності і рівень гостро- чи плосковершинності розподілу належать **коефіцієнти асиметрії** та **ексцесу**.

Статистичні ряди представляють у формі деякої кривої, до якої наближається графік ряду при збільшенні обсягу сукупності і зменшенні довжини інтервалів групування.

Крива розподілу – це графічне зображення у вигляді неперервної лінії змін частот у варіаційному ряді, функціонально пов′язаних зі зміною варіант (співвідношення варіант і частот). Крива може характеризувати емпіричний (за даними спостережень) або теоретичний (утворюється при дії лише основних, істотних причин) розподіл.

Аналіз варіаційних рядів зводиться до співставлення емпіричного і теоретичного розподілів і визначенню ступеня розбіжностей між ними.

За формою ряди розподілу поділяють на одно-, дво- та багатовершинні.

Багатовершинність свідчить про неоднорідність досліджуваної сукупності і вимагає перегрупування даних для виділення більш однорідних груп. **Одновершинні** переважно характеризують якісно однорідні сукупності і можуть бути симетричні та асиметричні (скошені), гостро- і плосковершинні.

Симетричним є розподіл, в якому частоти будь-яких двох варіантів, рівновіддалених в обидва боки від центра, рівні між собою. Для нього виконується рівність $x_{сер} = Me = Mo$.

Асиметричним є розподіл, у якому частоти по обидва боки від центра змінюються неоднаково, тобто вершина розподілу зміщена. Існує правостороння ($x_{сер} > Me > Mo$) і лівостороння ($x_{сер} < Me < Mo$) асиметрії. Асиметрія виникає як результат обмеженої варіації ознак в одному напрямку чи впливу домінуючої причини розвитку явища, яка призводить до зміщення центра розподілу.

Найпростішими показниками асиметрії є:

 абсолютне відхилення – різниця між середніми ($x_{сер} – Mo$) чи ($x_{сер} – Me$); чим більша різниця, тим більша асиметрія ряду;

 відносне відхилення – **коефіцієнт асиметрії (As)** – $As = (x_{сер} – Mo) / \sigma$, чи $As = (x_{сер} – Me) / \sigma$, при $As = 0$ – маємо симетричний розподіл, $As > 0$ – правостороння асиметрія, $As < 0$ – лівостороння асиметрія.

Для симетричного розподілу розраховують **коефіцієнт ексцесу (Е)**. Ексцес, це випад вершини емпіричного розподілу уверх чи униз від вершини кривої нормального розподілу. Якщо $E = 0$ – це нормальний розподіл (симетричний), $E > 0$ – гостровершинний, $E < 0$ – плосковершинний.

Необхідно відмітити, що хоча показники асиметрії та ексцесу характеризують безпосередньо лише форму розподілу ознаки в межах досліджуваної сукупності, однак їх визначення має не тільки описове значення. Часто асиметрія та ексцес надають конкретні вказівки щодо подальшого дослідження вивчаємих явищ. Наприклад, поява значного від′ємного ексцесу вказує на якісну неоднорідність досліджуваної сукупності. Крім того, ці показники дозволяють зробити висновки щодо можливості застосування даного емпіричного розподілу до типу кривих нормального розподілу.

3. Серед найпоширеніших теоретичних розподілів найчастіше використовується **нормальний розподіл (розподіл Лапласа – Гауса)**, який відображає нормальна крива (симетрична відносно max ординати). Такий розподіл є результатом впливу на значення ознаки необмеженої кількості незалежних один від одного факторів, як це буває в природі.

Нормальний розподіл повністю визначений двома параметрами – середньою арифметичною ($x_{сер}$) та стандартним (середнім квадратичним) відхиленням (σ). Значення ознаки при нормальному розподілі переважно зосереджуються біля центра розподілу – $x_{сер}$. Значення ознаки, які істотно відхиляються від $x_{сер}$, зустрічаються рідко. Підпорядкованість цьому закону буде більш точнішою у разі одночасної дії великої кількості випадкових величин. Якщо жодна з випадково діючих причин за своєю дією не буде переважаючою над іншими, то закон розподілу дуже близько підходить до нормального. Наприклад, за нормальним законом розподілені вага та зріст людини, відхилення у виробничому процесі

при нормальному рівні організації та технології, в розподілі населення означеного віку за розміром взуття та багато інших явищ, у яких проявляється велика кількість незалежних значень спостережуваних ознак, серед яких немає суттєво відмінних від решти значень ознаки статистичної сукупності.

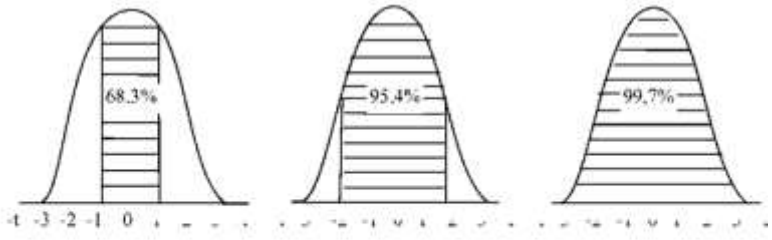
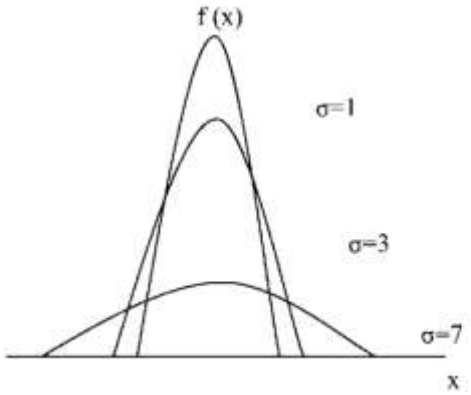


Рис.13. Нормальний розподіл з одно-, дво- та трьосигмовими границями



Поняття нормального розподілу покладено в основу багатьох методів статистики. Його використовують як стандарт для порівняння інших розподілів, застосовують у вибірковому, кореляційно-регресивному, факторному методах статистичного дослідження. Нормальний розподіл подібний до одновершинних симетричних розподілів, тому його розглядають як перше наближення в разі статистичного моделювання. При розв′язуванні багатьох задач необхідно встановити, за яким законом розподілена ознака статистичної сукупності. Щоб перевірити нормальність розподілу, тобто відповідність досліджуваного розподілу нормальному, слід частоти фактичного розподілу порівняти з теоретичними частотами, характерними для нормального розподілу. Для цього за фактичними даними, підставляючи їх у відповідну формулу, обчислюють теоретичні частоти кривої нормального розподілу. Ступінь відповідності між фактичним і теоретичним розподілами оцінюється за допомогою показників (критеріїв) узгодженості.

Критерій згоди – це статистичний показник, що використовується для з'ясування розбіжностей між прийнятою статистичною моделлю і спостережуваними даними ознаки, які має описати ця модель. Одним з найпоширеніших є **критерій Пірсона (хі-квадрат) –** χ^2:

$$\chi^2 = \sum (f_{теор.} - f_{емп.})^2 / f_{теор.}$$

Чим більшою буде різниця між емпіричними та теоретичними частотами, тим більше буде значення критерію Пірсона.

Обчисливши фактичне значення критерію, порівнюють його з табличним (критичним) значенням:

якщо $\chi^2_{ф.} > \chi^2_{табл.}$, тобто χ^2 попадає у критичну область – розбіжність між емпіричними та теоретичними частотами є істотною і її не можна пояснити випадковими коливаннями даних, а емпіричний розподіл є принципово відмінним від теоретичного;

якщо $\chi^2_{ф.} < \chi^2_{табл.}$, відхилення фактичних частот від теоретичних вважається випадковим, неістотним; емпіричний розподіл відповідає теоретичному.

Існують ще такі критерії перевірки відповідності нормальному розподілу як критерій Колмогорова-Смирнова, Романовського, Ястремського, Фішера, Вілконсона. Необхідною умовою використання цих критеріїв є достатньо велике число спостережень – не менше 100 (20-30 за критерієм Пірсона).

Для зображення статистичних рядів розподілу використовують такі графіки:

- полігон,
- гістограма,
- кумулята і огіва,
- крива концентрації (Лоренца).

Полігон – зображення варіаційного (переважно дискретного) ряду у вигляді ламаної лінії, що з'єднує сукупність точок в прямокутній системі координат. Значення ознаки відкладається на осі абсцис (X), а частоти (частки, щільність) – на осі ординат (Y).

Гістограма – сходинковий графік для інтервального варіаційного ряду. Утворені прямокутники пропорційні за висотою частотам варіантів для кожного інтервалу. У випадку нерівних інтервалів висота прямокутників пропорційна щільності розподілу ознаки у конкретному інтервалі. Гістограму можна перетворити на полігон, з'єднавши середини вершин стовпчиків лінією.

За гістограмою зручно визначати модальне значення ознаки – праву верхню вершину модального (з максимальною ординатою) прямокутника з'єднуємо з правою вершиною попереднього, а ліву вершину модального прямокутника – з лівою вершиною післямодального прямокутника. Абсциса точки перетину з'єднувальних прямих буде модою розподілу.

Кумулята і **огіва** – криві нагромаджених (кумулятивних) підсумків частот або часток. Використовують для зображення як дискретних так і інтервальних рядів. Будуючи кумуляту, на абсцисі відкладають варіанти, на ординаті – нагромаджені частоти. У разі побудови огіви, яка є дзеркальним відображенням кумуляти, навпаки. За цими кривими визначають, скільки одиниць сукупності, або яка їх частка не перевищує певного значення групувальної ознаки (для дискретного ряду) чи верхньої межі відповідного інтервалу (для інтервального ряду).

Кумулятивні криві надають можливість графічного визначення медіани – остання ордината кумуляти ділиться навпіл і через середину проводиться пряма паралельно осі абсцис до перетину з кривою. Значення абсциси цього перетину є медіаною.

Крива Лоренца – різновид кумулятивної кривої, яка відображає рівномірність розподілу частот варіаційного ряду та концентрацію обсягу досліджуваного явища.