Тема 9. ОСНОВНІ ПОНЯТТЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ: ВИБІРКОВІ СПОСТЕРЕЖЕННЯ ТА ВИБІРКОВІ ОЦІНКИ

План

1.Предмет і завдання математичної статистики. Основні положення вибіркового методу. Генеральна сукупність та вибірка

Математична статистика – розділ математики, що вивчає закономірності, які мають місце в масових явищах і статистичних сукупностях.

Завдання математичної статистики:

1. Створення методів збирання та обробки експериментальних значень випадкової величини.
2. Визначення невідомих параметрів розподілу випадкової величини і законів розподілу за статистичними даними.
3. Перевірка правдоподібності прийнятої гіпотези про закон розподілу.

Генеральна та вибіркова сукупності

Генеральною сукупністю називається загальна кількість однорідних об'єктів, що мають певну ознаку.

Наприклад, всі студенти вищого навчального закладу є генеральною сукупністю, а спільна ознака – середній бал за здану сесію.

Розподіл ознаки як випадкової величини для всієї генеральної сукупності пов'язаний з рядом труднощів. Тому з усієї генеральної сукупності виділяють деяку випадковим чином та у відповідний спосіб відібрану частину. Це частина називається ***вибірковою сукупністю*** (вибіркою).

Об’ємом сукупності (вибіркової – $\boldsymbol{n}$ або генеральної – $\boldsymbol{N}$) називається кількість об’єктів цієї сукупності.

Відношення об'єму вибірки до об'єму генеральної сукупності називається ***відносним показником вибірки.***

Метод, який базується на тому, що за даними дослідження вибірки, яка виділена з даної генеральної сукупності, робиться висновок про всю генеральну сукупність, називається ***вибірковим методом***.

Вибірка називається ***репрезентативною***, якщо кожен об’єкт генеральної сукупності має однакову можливість попасти у вибірку.

Тоді, згідно закону великих чисел, результати її дослідження будуть близькі до тих, які б ми отримали, вивчаючи всю генеральну сукупність.

Способи відбору статистичних даних:

1. ***Повторний*** – якщо вибраний елемент повертається до всіх елементів і може бути випадково вибраний повторно.

2. ***Безповторний*** – якщо вибраний елемент не повертається до всіх елементів.

Якщо генеральна сукупність ділиться на окремі групи, то зроблена вибірка може бути ***типовою, механічною, серійною.***

Якщо вибірку роблять з типової частини генеральної сукупності, то маємо типову вибірку. *К*оли об’єкти обираються через певний інтервал механічно, то маємо механічну вибірку. Якщо з генеральної сукупності вибирають не по одному об'єкту, а серіями, які і досліджується, то маємо серійну вибірку. В економічних дослідженнях часто використовують комбінований відбір.

Повернутись до плану.

2. Статистичний розподіл. Первинна обробка і графічне подання вибіркових даних

Дослідження вибірки починають зі складання статистичного розподілу.

У математичній статистиці замість слова «дані» вживають термін «варіанти».

Нехай з генеральної сукупності зроблено вибірку об’єму $\boldsymbol{n}$. Значення $\boldsymbol{x_i}$ вибірки – ***варіанти***. Числову характеристику варіанти називають ***ознакою***.

Варіанти, розташовані в ряд у порядку зростання, називаються ***варіаційним або статистичним рядом***. Якщо при цьому $\boldsymbol{x_i}$ повторюється n_i раз ($\boldsymbol{i}$ = 1,2,…k;) а $\sum_{i=1}^{k} n_i = n$, то величина $\boldsymbol{n_i}$

називається ***частотою*** варіанти x_i, а величина $\frac{n_i}{n}$ – ***відносною частотою*** варіанти x_i.

Варіанта, що ділить варіаційний ряд на дві рівні частини, називається ***медіаною***. Це варіанта середини ряду.

Варіанта, що має найбільшу відносну частоту появи, називається ***модою***.

Різниця між меншим та найбільшим значеннями варіанти називається ***розмахом вибірки***.

Відношення кількості значень варіант до загального об'єму вибірки називається ***відносною частотою появи варіанти***.

Статистичним або емпіричним розподілом вибірки називають таблицю значень ознаки, що розташовані у порядку зростання та відповідних їм частот (відносних частот).

x_i	x_1	...	x_k
Частоти, n_i	n_1	...	n_k

Розрізняють:

- дискретні розподіли;
- інтервальні розподіли або гістограма (послідовність інтервалів та відповідних їм частот).

Для побудови гістограми потрібно визначити кількість інтервалів розподілу та крок розподілу. Вважаємо, що досліджуємо вибірку розташовану у порядку зростання значень.

Тобто відомі: обсяг вибірки – n, максимальне – $x_{\max}$ та мінімальне значення – $x_{\min}$.

Тоді крок визначається (формула Стерджеса):

$$h = (x_{\max} - x_{\min})/(1 + 3{,}322 \cdot \log_{10}(n)).$$

Якщо це число не ціле, то береться ціле значення та додається одиниця.

Початок першого інтервалу визначається:

$a_0 = x_{\min} - h/2$.

Тоді кінець останнього інтервалу групування визначають як: $a_m = x_{m-1} + h/2$

Кількість інтервалів $\boldsymbol{m}$ визначається з умови, що $x_{\max}$ попадає в останній інтервал. Якщо n_j кількість влучень в j інтервал (частота) то виконується умова:

$$\sum_{j=1}^{m} n_j = n.$$

Відносна частота визначається:

$$w_j = n_j / n \ (\sum_{j=1}^{m} w_j = 1).$$

Емпірична функція розподілу (кумулята) визначається:

$$F(x) = \sum_{j=1}^{x=a_j} w_j$$

Таблиця, що використовується для побудови гістограми має наступний вигляд

Номер інтервалу	**1**	**2**	**……..**	**m**
Межі інтервалу	$(a_0;a_1]$	$(a_1;a_2]$	……..	$(a_{m-1};a_m]$
Середина інтервалу	$\bar{a}_1 = \frac{a_0 + a_1}{2}$	$\bar{a}_2 = \frac{a_1 + a_2}{2}$	………	$\bar{a}_m = \frac{a_{m-1} + a_m}{2}$
Частота	n_1	n_2	………	n_m
Відносна частота	$w_1 = n_1 / n$	$w_2 = n_2 / n$	………	$w_m = n_m / n$
Кумулята	w_1	$w_1 + w_2$	$w_1 + w_2 +..$	$w_1 + w_2 + ... + w_m$

Статистичні характеристики по згрупованій вибірці:

$$\overline{X}_в = \sum_{j=1}^{m} w_j \cdot \bar{a}_j$$

$$\sigma_в^2 = \sum_{j=1}^{m} w_j \cdot \bar{a}_j^2 - \overline{X}_в^2$$

$$A = \sum_{j=1}^{m} w_j \cdot \left(\bar{a}_j - \bar{X}_в\right)^3 / \sigma_в^3$$

$$E = \sum_{j=1}^{m} w_j \cdot \left(\bar{a}_j - \bar{X}_в\right)^4 / \sigma_в^4 - 3$$

Де $\bar{a}_j$ - центри інтервалів групування.

Приклад 9.1. (дискретний розподіл).

За тиждень продано 50 пар взуття, що мають наступні розміри (табл.). Побудувати ряд розподілу та емпіричну функцію розподілу.

Розмір взуття	37	38	40	42	43
Частота	10	12	15	10	3
Відносна частота	0,2	0,24	0,3	0,2	0,06
Кумулята	0,2	0,44	0,74	0,94	1,0

Приклад 9.2. (інтервальний розподіл).

Отримано наступний варіаційний ряд довжин колосків обсягом 30 спостережень (см) : 8;8;9;9;9;9;10;10;10;10;10;11;11;11;11;11;11;12;12;12;12;13;13;13;14;14; 15;15;16;16.

Побудувати таблицю інтервального розподілу та зробити оцінки математичного сподівання, дисперсії, коефіцієнтів асиметрії та ексцесу.

Розв'язання:

За формулою Стерджеса знайдемо крок гістограми

$(\, x_{min} = 8см; \quad x_{maa} = 16см; \quad n = 30\,)$:

$h = (16 - 8)/(1 + 3{,}322\lg 30) \approx 1{,}35$.

Ціла частина від цього числа дорівнює 1. Тобто крок дорівнює 2см. Початкова точка розбивки дорівнює 7см. Знайдемо інтервали та відповідні частоти (табл.)

Номер інтервалу	1	2	3	4	5
Межі інтервалу	(7;9]	(9;11]	(11;13]	(13;15]	(15:17]
Середина інтервалу	8	10	12	14	16
Частота	6	11	7	4	2
Відносна частота	0,2	0,37	0,23	0,13	0,07
Кумулята	0,2	0,57	0,8	0,93	1,0

За згрупованими даними зробимо оцінку статистичних характеристик:

$$\overline{X}_в = 0{,}2\cdot 8 + 0{,}37\cdot 10 + 0{,}23\cdot 12 + 0{,}13\cdot 14 + 0{,}07\cdot 16 = 11 см$$

$$\sigma_в^2 = 0{,}2\cdot 8^2 + 0{,}37\cdot 10^2 + 0{,}23\cdot 12^2 + 0{,}13\cdot 14^2 + 0{,}07\cdot 16^2 - 11^2 = 5{,}32; \quad \Rightarrow \quad \sigma_в = 2{,}31$$

$$\gamma = (0{,}2(8-11)^3 + 0{,}37(10-11)^3 + 0{,}23(12-11)^3 + 0{,}13(14-11)^3 + 0{,}07(16-11)^3)/2{,}31^3 \approx 0{,}55$$

$$\mu = (0{,}2(8-11)^4 + 0{,}37(10-11)^4 + 0{,}23(12-11)^4 + 0{,}13(14-11)^4 + 0{,}07(16-11)^4)/2{,}31^4 - 3 \approx -0{,}50$$

Емпіричну функцію розподілу можна використовувати як оцінку (наближене значення) функції розподілу $F(x)$ випадкової величини X (генеральної сукупності).

Функція $F^*(x)$ має всі властивості функції розподілу $F(x)$:

1. $F^*(x) \in [0;1]$;
2. $F^*(x)$ – не спадна функція;
3. $F^*(x) = 0$ при $x \le x_1$, де x_1 – найменша варіанта; $F^*(x) = 1$ при $x \ge x_k$, де x_k – найбільша варіанта.

Повернутись до плану.

3. Статистична обробка дискретних величин

Статистичні дані вибірки розташовують у порядку зростання значення варіанти. Для кожної варіанти визначають частоту і відносну частоту появи.

Будують графік залежності відносної частоти від значень варіанти. Цей графік називають ***полігоном*** частот або многокутником розподілу (рис. 9.1.).

Знаходять нагромаджуючі частоти n_i^* варіант. Нагромаджуючою частотою n_i^* варіанти x_i, називається відношення кількості членів сукупності, які мають значення ознаки, що не перевищує величину x_i до загальної кількості членів вибірки.

Будують графік залежності нагромаджуючої частоти від значень варіанти, який називається ***кумулятою.***

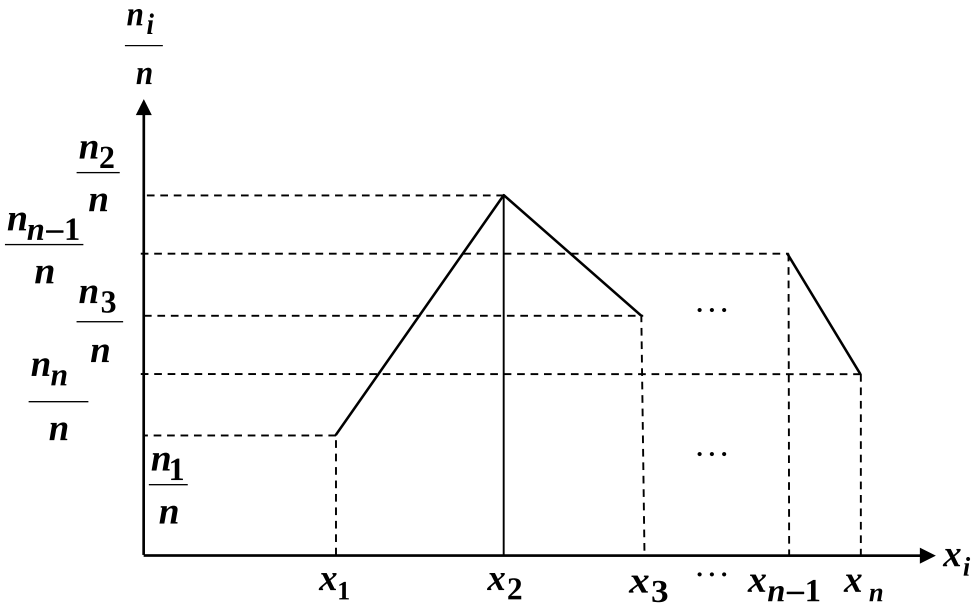

Рис. 9.1. Полігон

Нагромаджуюча частота є ***емпіричною функцією*** розподілу.

З експерименту маємо набір значень випадкової неперервної величини. Імовірність появи конкретного значення завжди є нуль. Це означає, що зібрані значення не є точні, а визначені наближено.

У зв'язку з цим після побудови варіаційного ряду встановлюють гарантований діапазон зміни випадкової величини і розбивають його на інтервали. Визначають частоту попадання значень в кожен інтервал і відносну частоту попадання.

Повернутись до плану.

4. Числові характеристики вибіркової сукупності

Щоб характеризувати найбільш суттєві властивості статистичного розподілу, як і в теорії ймовірностей, використовують середні показники або, як їх називають – ***вибіркові числові характеристики***.

1. ***Вибіркова середня –*** $\overline{X}_в$**.**

Для не згрупованих даних вибіркову середню розраховують за формулою:

$$\overline{X}_в = \frac{\sum_{i=1}^{n} x_i}{n} \; ;$$

для згрупованих даних :

$$\overline{X}_в = \frac{\sum_{j=1}^{m} x_j n_j}{n}.$$

1. ***Вибіркова дисперсія*** $-D_в$ **:**

Для не згрупованих даних вибіркову дисперсію розраховують за формулою:

$$D_в = \frac{\sum_{i=1}^{n}(x_i - \overline{X}_в)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\overline{X}_в\right)^2$$

для згрупованих даних:

$$D_в = \frac{\sum_{j=1}^{m}\left(\overline{x}_j - \overline{X}_в\right)^2 n_j}{n} = \frac{\sum_{j=1}^{m} x_j^2 \cdot n_j}{n} - \left(\overline{X}_в\right)^2 \; ,$$

де m – кількість груп.

Незміщена оцінка дисперсії визначається:

$S^2 = \frac{n}{n-1} D_в$, або:

для не згрупованих даних

$$S^2 = \frac{\sum_{j=1}^{m}\left(\overline{x}_j - \overline{X}_в\right)^2 n_j}{n-1}$$

і для згрупованих даних:

$$S^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{X}_в)^2 n_i}{n-1}$$

2. ***Вибіркове середньоквадратичне відхилення*** – величина $S = \sigma_в = \sqrt{D_в}$.

3. ***Коефіцієнт варіації*** використовують для порівняння міри розсіювання значень ознак навколо вибіркової середньої у різних вибірках :

$$V = \frac{\sigma_в}{\overline{X}_в} \cdot 100\%.$$

Повернутись до плану.