Лекція 3. СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

3.1. Числові характеристики дискретного статистичного розподілу вибірки

Вибіркова середня величина $\left(\overline{x}_B\right)$.

$$\overline{x}_B = \frac{\sum_{i=1}^{k} x_i n_i}{n}, \qquad (3.1)$$

У випадку, коли всі варіанти з'являються у вибірці лише по одному разу, тобто $n_i = 1$, то вибіркова середня величина:

$$\overline{x}_B = \frac{\sum_{i=1}^{k} x_i}{n}$$

Формулу (3.1) можна отримати при $p = 1$ із узагальненої формули для степеневих середніх

$$\overline{x}_p = \left(\frac{\sum_{i=1}^{k} x_i^p n_i}{n} \right)^{\frac{1}{p}}, \qquad (3.2)$$

тоді $\overline{x}_1 = \overline{X}$.

При $p = -1$ (середнє гармонічне): $\overline{x}_{-1} = \overline{x}_{гарм} = \dfrac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$.

При $p = 0$, яка після розкриття невизначеності при обчисленні границі, коли $p \to 0$, набуває вигляду $\overline{x}_0 = \overline{x}_{геом} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$ (середнє геометричне).

При $p=2$: $\overline{x}_2=\sqrt{\sum_{i=1}^{k}\frac{x_i^2 n_i}{n}}$.

Модою $\left(Mo^*\right)$ дискретного статистичного розподілу вибірки називають варіанту, що має найбільшу частоту появи: $Mo^*=x_i$, якщо $n_i=\max n_i$.

Якщо всі значення в групі зустрічаються з однаковою частотою, то вважається, що моди немає. Такий розподіл називають *антимодальними*.

Якщо дві варіанти ряду мають однакову частоту і є суміжними між собою (сусідні) та їхні частоти більше ніж частоти будь-якої іншої варіанти цього ряду, то значення моди є середнє цих двох значень.

Якщо те ж саме відноситься до двох несуміжних варіант, то існує дві моди, а розподіл називають *бімодальним*.

Бімодальний розподіл указує на якісну неоднорідність сукупності за досліджуваною ознакою.

Медіаною $\left(Me^*\right)$ дискретного статистичного розподілу вибірки називають варіанту, яка поділяє варіаційний ряд на дві частини, рівні за кількістю варіант:

$$Me^*=\begin{cases} x_i, & n=2\cdot i-1, \quad (\textit{непарна кількість варіант}) \\ \dfrac{x_i+x_{i+1}}{2}, & n=2\cdot i, \quad (\textit{парна кількість варіант}) \end{cases},$$

де x_i - середина варіаційного ряду.

Відхилення варіант.

Різницю $\left(x_i-\overline{x}_B\right)\cdot n_i$ називають *відхиленням варіант*.

Розмах вибірки (варіацій) (R):

$$R = x_{\max} - x_{\min} \tag{3.3}$$

Дисперсія вибірки — це середнє арифметичне квадратів відхилень варіант відносно $\overline{x}_B$, яке обчислюється за формулою:

$$D_B = \frac{\sum_{i=1}^{k} \left(x_i - \overline{x}_B\right)^2 \cdot n_i}{n} \tag{3.4}$$

або

$$D_B = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{n} - \left(\overline{x}_B\right)^2 \tag{3.5}$$

Середнє квадратичне відхилення вибірки σ_B:

$$\sigma_B = \sqrt{D_B}\,, \tag{3.6}$$

Коефіцієнт варіації (V):

$$V = \frac{\sigma_B}{\overline{x}_B} \cdot 100\% \tag{3.7}$$

Коефіцієнт варіації є оцінкою надійності середньої. При величині *V*=5% варіація вважається слабкою, *V*=6-10 % -помірною, *V*=16-20 % - значною *V*=21-50 % - великою; *V*> 50 % -дуже великою.

Для малих вибірок величина коефіцієнта варіації повинна бути не більше 33 %. Якщо $\overline{x}_B$=1; *V*=1.

Назва	Значення
Вибіркове середнє $\overline{x}_B$	характеризує центральну тенденцію ряду даних за умови нормальності розподілу вибірки
Дисперсія D_B	характеризує розкид даних відносно середнього арифметичного; застосовується в статистичних тестах для порівняння розкидів різних наборів нормально розподілених даних; має квадратичну розмірність аналізованої величини (наприклад, см2)
Середньоквадратичне відхилення σ_B	характеризує розкид даних відносно середнього арифметичного; застосовується для опису нормально розподілених виборок; має таку ж розмірність, що і аналізована величина (наприклад, см)
мінімум; максимум min; max	певною мірою характеризує розкид даних не залежно від характеру розподілу; дуже чутливі до «викидів» показники. Викид – деякий компонент вибірки, що сильно відрізняється від усіх інших компонентів цієї вибірки.
Розмах $R = x_{\max} - x_{\min}$	чим менший розмах варіації, тим стійкіше досліджуваний процес, тобто його можна охарактеризувати як більш передбачуваний.
Коефіцієнт варіації (V)	виражає стандартне відхилення відносно середнього значення. Тобто він намагається пояснити, наскільки великим є значення середньоквадратичного відхилення відносно середнього значення.

3.2. Числові характеристики інтегрального статистичного розподілу вибірки

Медіана

Для визначення медіани інтегрального статистичного розподілу вибірки необхідно визначити медіанний частковий інтервал. Якщо, наприклад, на i-му інтервалі $\left[x_{i-1} - x_i\right]$ $F^*\left(x_{i-1}\right) < 0,5$ і $F^*\left(x_i\right) > 0,5$, то, беручи до уваги, що досліджувана ознака X є неперервною і при цьому $F^*\left(x\right)$ є неспадною функцією, всередині інтервалу $\left[x_{i-1} - x_i\right]$ неодмінно існує таке значення $X = Me$, де $F^*\left(Me\right) = 0,5$.

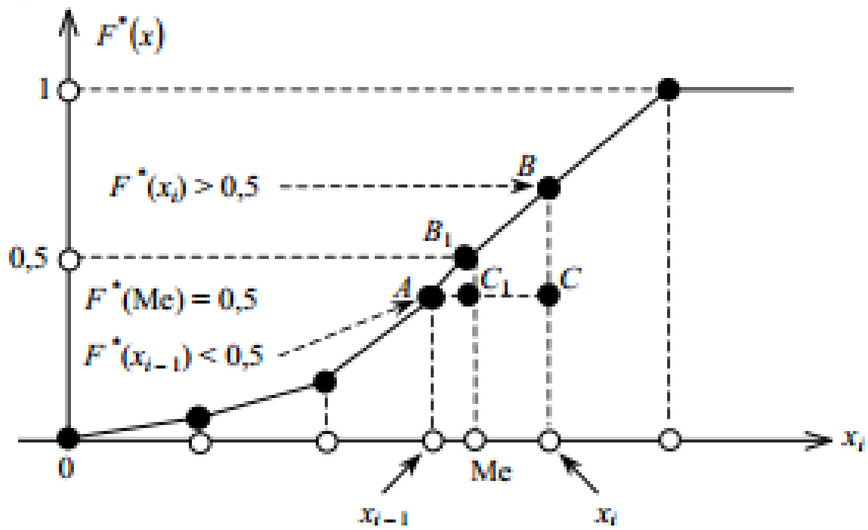

Рис. 2.2. Графік емпіричної функції розподілу за інтегральним статистичним розподілом вибірки (комулята)

З подібності трикутників ΔABC і ΔAB_1C_1 маємо:

$$\frac{x_i - x_{i-1}}{Me^* - x_{i-1}} = \frac{F^*(x_i) - F^*(x_{i-1})}{0{,}5 - F^*(x_{i-1})} \rightarrow Me^* = x_{i-1} + \frac{0{,}5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} \cdot h \qquad (3.8)$$

де $h = x_i - x_{i-1}$ називаються *кроком*.

Мода

Для визначення моди інтервального статистичного розподілу необхідно знайти модальний інтервал, тобто такий частинний інтервал, що має найбільшу частоту появи.

Використовуючи лінійну інтерполяцію, моду обчислимо за формулою

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} \cdot h \qquad (3.9)$$

де x_{i-1} - початок модального інтервалу;

h - довжина, або крок, часткового інтегралу;

n_{Mo} - частота модального інтервалу;

n_{Mo-1} - частота домодального інтервалу;

n_{Mo+1} - частота післямодального інтегралу.

Визначення вибіркової середньої величини:

де $x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}$ - середина часткових інтервалів, і має вигляд:

$x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}$	x_1^*	x_2^*	x_3^*	$\dots$	x_k^*
n_i	n_1	n_2	n_3	$\dots$	n_k

Тоді *вибіркова середня величина* $(\overline{x}_B)$: $\overline{x}_B = \frac{\sum_{i=1}^{k} x_i^* n_i}{n}$ (3.10)

Дисперсія вибірки (D_B): $D_B = \frac{\sum_{i=1}^{k} \left(x_i^*\right)^2 \cdot n_i}{n} - \left(\overline{x}_B\right)^2$ (3.11)

Середнє квадратичне відхилення (σ_B): $\sigma_B = \sqrt{D_B}$ (3.12)

3.3. Емпіричні моменти

Початкові емпіричні моменти

Середнє зважене значення варіант у степені k $(k = 1, 2, 3, \dots)$ називають початковим емпіричним моментом $k-$го порядку $\left(v_k^*\right)$, який обчислюється за формулою:

$$v_k^* = \frac{\sum_{i=1}^{m} x_i^k n_i}{n} \quad (3.13)$$

При $k = 1$ дістанемо початковий момент першого порядку:

$$v_1^* = \frac{\sum_{i=1}^{m} x_i n_i}{n} = \overline{x}_B \quad (3.14)$$

При $k = 2$ дістанемо початковий момент другого порядку:

$$v_2^* = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} \quad (3.15)$$

m - кількість варіантів ряду.

Отже, дисперсію вибірки можна подати через початкові моменти першого та другого порядків, а саме:

$$D_B = \nu_2^* - \left(\nu_1^*\right)^2 \qquad (3.16)$$

Центральний емпіричний момент $k-$ го порядку.

Середнє зважене відхилення варіант у степені k $\left(k = 1, 2, 3, \ldots\right)$ називають центральним емпіричним моментом $k-$ го порядку

$$\mu_k^* = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)^k n_i}{n} \qquad (3.17)$$

При $k = 1$ дістанемо:

$$\mu_1^* = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)\cdot n_i}{n} = \frac{\sum_{i=1}^{m} x_i n_i}{n} - \overline{x}_B \cdot \frac{\sum_{i=1}^{m} n_i}{n} = \overline{x}_B - \overline{x}_B = 0$$

При $k = 2$ дістанемо:

$$\mu_2^* = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)^2 n_i}{n} = D_B,$$

де m - кількість варіантів ряду.

На практиці найчастіше застосовують центральні емпіричні моменти третього та четвертого порядків, що обчислюються за формулами:

$$\mu_3^* = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)^3 n_i}{n} \qquad (3.18)$$

$$\mu_4^* = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}_B\right)^4 n_i}{n} \qquad (3.19)$$

Підносячи до третього та четвертого степені відхилення варіант, подаємо μ_3^* та μ_4^* через відповідні початкові моменти:

$$\mu_3^* = \nu_3^* - 3\nu_2^* \cdot \nu_1^* + 2\left(\nu_1^*\right)^2 \qquad (3.20)$$

$$\mu_4^* = \nu_4^* - 4\nu_3^* \cdot \nu_1^* + 6\nu_2^* \cdot \left(\nu_1^*\right)^2 - 3\left(\nu_1^*\right)^4 \qquad (3.21)$$

Коефіцієнт асиметрії $\left(A_s^*\right)$

$$A_s^* = \frac{\mu_3^*}{\sigma_B^3}. \qquad (3.22)$$

Якщо коефіцієнт скошеності дорівнює нулю, то розподіл симетричний, якщо не дорівнює нулю –асиметричний. У випадках, коли $A_s^* > 0$, розподіл має лівосторонню асиметрію, коли $A_s^* < 0$ – правосторонню асиметрію, асиметрія відсутня коли $A_s^* = 0$.

У статистичній практиці прийнято вважати, що при значенні коефіцієнта $A_s^* < \pm 0,25$ асиметрія є незначною, при значенні $A_s^* > \pm 0,25$ – емпіричний розподіл відрізняється від нормального значним зміщенням.

Похибка коефіцієнта асиметрії можна знайти за формулою:

$$\delta_A = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}},$$ де n - обсяг вибірки.

Ексцес $\left(E_s^*\right)$

$$E_s^* = \frac{\mu_4^*}{\sigma_B^4} - 3. \qquad (3.23)$$

При нормальному розподілі ексцес $E_s^* = 0$, при гостровершинному (вершина фактичного розподілу виступає над вершиною нормального розподілу) $E_s^* > 0$, при туповершинному (плосковершинному) (вершина фактичного розподілу знаходиться нижче вершини нормального розподілу) $E_s^* < 0$.

У тих випадках, коли величина коефіцієнта ексцесу не перебільшує ±0,4, крива фактичного розподілу вважається слабоексцесивною.

Максимальне значення від'ємного ексцесу становить –2. У цьому випадку вершина кривої фактичного розподілу опускається до осі абсцис і крива розподілу ділиться на дві самостійні одновершинні криві.

Похибка коефіцієнта ексцесу можна знайти за формулою:

$$\delta_E = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}$$, де n - обсяг вибірки.