Практична робота N 1

Тема: „Інтервальний розподіл та його числові характеристики."

Зразок розв'язання типового варіанту

Задача. *В результаті вимірювання зростання ВВП підприємств регіону (у % до попереднього року) отримані наступні результати:*

Таблиця 1

2,35	1	1,5	9,78	1,41	1,59	2,55	2,31	9,07	6,03
2,94	2,49	2,88	9,54	2,55	8,16	8,17	2,52	4,77	7,62
7,44	7,95	2,84	5,19	3,94	13	1,56	5,48	12,27	5,22

Завдання:

1) Побудувати відповідний інтервальний розподіл (для визначення кількості інтервалів використати формулу Стерджеса).

2) Побудувати гістограму частот; емпіричну функцію розподілу $F^*(x)$ і її графік.

3) Знайти числові характеристики отриманого інтервального розподілу:

3.1) показники центру розподілу $\overline{x}_B$, Mo^*, Me^*;

3.2) показники варіації (D_B, σ_B, R, V; S^2, S) та оцінити однорідність сукупності;

3.3) показники форми розподілу (A_s^* та E_s^*).

Інтерпретувати отримані результати.

Розв'язання. **1)** Для представлення дискретного розподілу у вигляді інтервального слід увесь діапазон вхідних даних розділити на K інтервалів однакової ширини h і згрупувати дані (підрахувати частоту входження варіант вибірки у кожен з отриманих інтервалів). Зазвичай

$$5 \le K \le 20.$$

Число K рахують одним з трьох способів

$$K \approx \sqrt{n}, \qquad (1)$$

$$K < 5\lg n, \qquad (2)$$

$$K \approx 1 + 3{,}322\lg n, \qquad (3)$$

де n - обсяг (об'єм) вибірки. Формулу (3) називають ***формулою Стерджеса*** і використовують найчастіше. Зрозуміло, що K має бути цілим, тому знак «$\approx$» слід сприймати як вказівку на заокруглення отриманого у формулі числа до цілого (за правилами заокруглення).

При цьому довжину інтервалу обчисляють за формулою

$$h = \frac{x_{\max} - x_{\min}}{K} = \frac{R}{K}, \qquad (4)$$

де R - розмах вибірки. Усі інтервали беруться напіввідкритими

$$I_i = [x_{i-1}, x_i) \quad (1 \le i \le K-1), \qquad (5)$$

крім останнього

$$I_K = [x_{K-1}, x_K], \qquad (6)$$

який береться замкненим, щоб інтервалів розбиття було рівно K.

Кінці інтервалів поділу знаходяться так:

$$x_0 = x_{\min}, \quad x_i = x_{i-1} + h \quad (1 \le i \le K) \qquad (7)$$

Знайдемо K та h для нашого прикладу. Обсяг вибірки $n = 30$. Тоді за формулою Стерджеса (3) маємо

$$K \approx 1 + 3{,}322 \lg 30 \approx 1 + 3{,}322 \cdot 1{,}4771 = 1 + 4{,}7991 = 5{,}7991 \approx 6,$$

отже, вибірку слід розбити на 6 рівних інтервалів. Враховуючи, що найменший елемент вибірки $x_{\min} = 1$, а найбільший $x_{\max} = 13$, визначимо ширину кожного інтервалу за формулою (4)

$$h = \frac{x_{\max} - x_{\min}}{K} = \frac{13-1}{6} = 2$$

(у випадку нецілого h отримане за формулою (4) число заокруглюється, але тоді правий кінець останнього інтервалу поділу за рахунок накопичення похибки заокруглення буде більшим за $x_{\max}$ і, отже, в такому разі останній інтервал також можна брати напіввідкритим, як і всі попередні). Кінці інтервалів поділу:

$$x_0 = x_{\min} = 1,\ x_1 = x_0 + h = 1 + 2 = 3,\ x_2 = x_1 + h = 3 + 2 = 5,\ x_3 = x_2 + h = 5 + 2 = 7,$$
$$x_4 = x_3 + h = 7 + 2 = 9,\ x_5 = x_4 + h = 9 + 2 = 11,\ x_6 = x_5 + h = 11 + 2 = 13 = x_{\max}.$$

Оскільки $x_K = x_{\max}$, то останній інтервал поділу беремо замкненим. Маємо:

$I_1 = [x_0, x_1) = [1,3)$; в цей інтервал попадають варіанти 1; 1,41; 1,5; 1,56; 1,59; 2,31; 2,35; 2,49; 2,52; 2,55; 2,55; 2,84; 2,88; 2,94 – усього 14 варіант, отже, $n_1 = 14$;

$I_2 = [x_1, x_2) = [3,5)$; в цей інтервал попадають варіанти 3,94; 4,77 – усього 2 варіанти, отже, $n_2 = 2$;

$I_3 = [x_2, x_3) = [5,7)$; в цей інтервал попадають варіанти 5,19; 5,22; 5,48; 6,03 – усього 4 варіанти, отже, $n_3 = 4$;

$I_4 = [x_3, x_4) = [7,9)$; в цей інтервал попадають варіанти 7,44; 7,62; 7,95; 8,16; 8,17 – усього 5 варіант, отже, $n_4 = 5$;

$I_5 = [x_4, x_5) = [9,11)$; в цей інтервал попадають варіанти 9,07; 9,54; 9,78 – усього 3 варіанти, отже, $n_5 = 3$;

$I_6 = [x_5, x_6] = [11,13]$ - останній, 6-й інтервал беремо замкненим, щоб не збільшилась кількість інтервалів поділу; в цей інтервал попадають варіанти 12,27; 13 – усього 2 варіанти, отже, $n_6 = 2$. Перевірка: $n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 14 + 2 + 4 + 5 + 3 + 2 = 30 = n$. Заносимо результати в таблицю:

Таблиця 2

$I_i, h=2$	[1,3)	[3,5)	[5,7)	[7,9)	[9,11)	[11,13]
n_i	14	2	4	5	3	2

Отримали інтервальний розподіл, що відповідає вихідному дискретному розподілу при поділі на 6 інтервалів.

2) За даними таблиці 2 побудуємо гістограму частот, емпіричну функцію розподілу $F^*(x)$ та її графік.

Для побудови гістограми частот слід у прямокутній декартовій системі координат будувати прямокутники з основами $[x_{i-1}, x_i]$ (на вісі Ox) і висотами $\frac{n_i}{h}$ (на вісі Oy), $1 \le i \le K$. Утворена стовпчаста фігура носить назву ***гістограма частот***. При цьому сума площ усіх прямокутників, які утворюють гістограму частот, дорівнює обсягу вибірки n. Справді,

$$S_{фіг} = \sum_{i=1}^{K} h \cdot \frac{n_i}{h} = \sum_{i=1}^{K} n_i = n .$$

Маємо:

$\frac{n_1}{h} = \frac{14}{2} = 7$, отже, на [1,3] висота прямокутника 7 од;

$\frac{n_2}{h} = \frac{2}{2} = 1$, отже, на [3,5] висота прямокутника 1 од;

$\frac{n_3}{h} = \frac{4}{2} = 2$, отже, на [5,7] висота прямокутника 2 од;

$\frac{n_4}{h} = \frac{5}{2}$, отже, на [7,9] висота прямокутника 2,5 од;

$\frac{n_5}{h} = \frac{3}{2}$, отже, на [9,11] висота прямокутника 1,5 од;

$\frac{n_6}{h} = \frac{2}{2} = 1$, отже, на [11,13] висота прямокутника 1 од.

Будуємо:

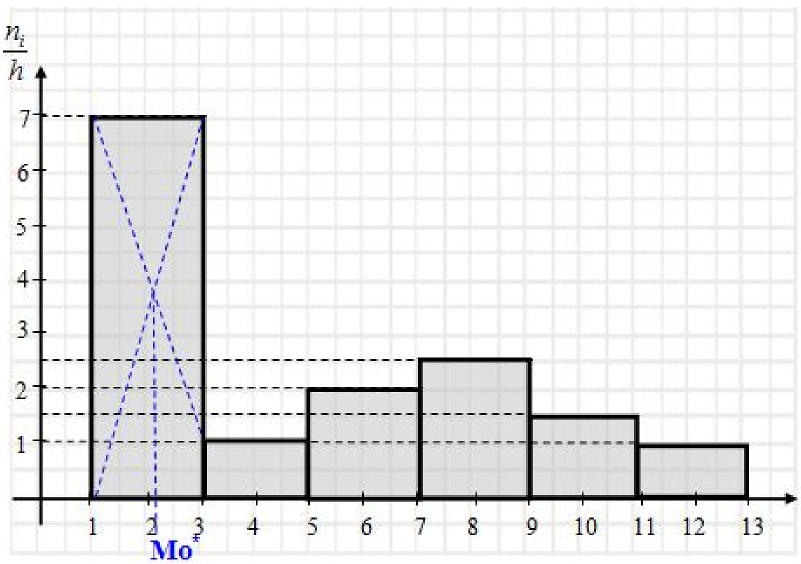


Рис. 1. *Гістограма частот*

Зауваження 1: *гістограма використовується для уявлення про вигляд графіка функції щільності розподілу випадкової величини X. Для цього слід з'єднати відрізками середини верхніх сторін прямокутників гістограми; в результаті утвориться ламана (полігон), яку можна розглядати як графік функції, котра є оцінкою справжньої щільності ймовірностей (див. рис. 1.1).*

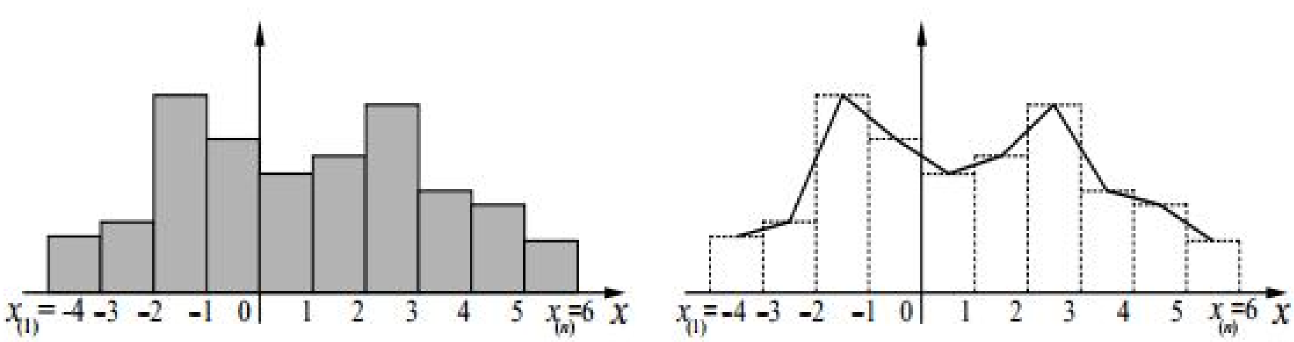


Рис. 1.1. *Побудова за гістограмою функції (полігону), яка є оцінкою щільності ймовірностей*

Для побудови емпіричної функції розподілу зручно таблицю 2 доповнити рядком відносних частот і нагромаджених відносних частот (табл. 3)

Тоді алгоритм побудови $F^*(x)$ дуже простий: лівіше лівого кінця першого інтервалу емпірична функція розподілу дорівнює нулеві, а на кожному з інтервалів розподілу її значення співпадає з нагромадженою відносною частотою цього інтервалу (4-й рядок табл. 3).

Таблиця 3

$I_i, h=2$	[1,3)	[3,5)	[5,7)	[7,9)	[9,11)	[11,13]
n_i	14	2	4	5	3	2
$W_i = \frac{n_i}{n}$	$\frac{14}{30}$	$\frac{2}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{3}{30}$	$\frac{2}{30}$
нагромаджені W_i	$\frac{14}{30}$	$\frac{16}{30}$	$\frac{20}{30}$	$\frac{25}{30}$	$\frac{28}{30}$	1

Маємо

$$F^*(x) = \begin{cases} 0, & x \le 1, \\ \frac{14}{30}, & 1 < x \le 3, \\ \frac{16}{30}, & 3 < x \le 5, \\ \frac{20}{30}, & 5 < x \le 7, \\ \frac{25}{30}, & 7 < x \le 9, \\ \frac{28}{30}, & 9 < x \le 11, \\ 1, & 11 < x \le 13, \end{cases} \qquad (8)$$

Графік емпіричної функції $F^*(x)$ у випадку інтервального розподілу є неперервною кривою (на відміну від східчастого вигляду для дискретного розподілу). Для його побудови слід на координатну площину нанести точки $(x_i, F^*(x_i))$, де x_i - праві кінці відповідних інтервалів ($1 \le i \le K$),

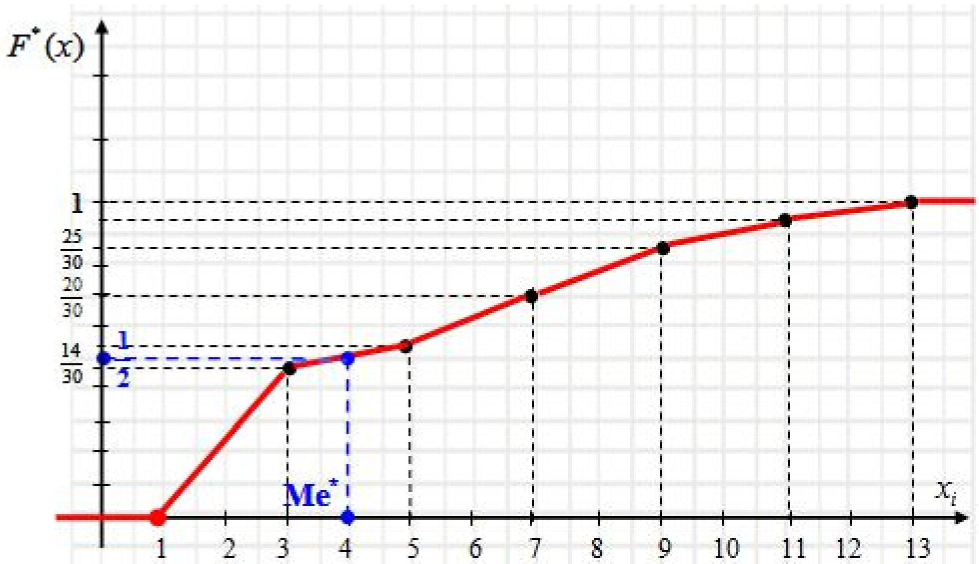


Рис. 2. *Графік емпіричної функції розподілу*

після чого з'єднати отримані точки плавною лінією. В нашому прикладі це будуть точки $(1,0)$, $\left(3, \frac{14}{30}\right)$, $\left(5, \frac{16}{30}\right)$, $\left(7, \frac{20}{30}\right)$, $\left(9, \frac{25}{30}\right)$, $\left(11, \frac{28}{30}\right)$, $(13,1)$. У прямокутній декартовій системі координат зображаємо вищевказані точки і будуємо графік $F^*(x)$ (рис. 2).

3.1) а) Для обчислення $\overline{x}_B$ (а також D_B, σ_B, V, A_s^*, E_s^*) треба перейти від інтервального розподілу до дискретного, замінюючи інтервали їхніми серединами

$$x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}, \qquad (9)$$

а частоти n_i залишаючи незмінними (тобто середині x_i^* інтервалу $[x_{i-1}, x_i)$ відповідатиме та сама частота n_i). Результати подамо у вигляді наступної таблиці

Таблиця 3.1

Інтервали $I_i, h=2$	*Середини інтервалів* $x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}$	*Частоти* n_i	*Нагромаджені частоти* S_i
[1,3) - модальний	2	14	14
[3,5) - медіанний	4	2	16 ($>30/2=15$)
[5,7)	6	4	20
[7,9)	8	5	25
[9,11)	10	3	28
[11,13]	12	2	30

Середнє вибіркове $\overline{x}_B$ обчислимо за формулою

$$\overline{x}_B = \frac{\sum_{i=1}^{K} x_i^* n_i}{n} = \frac{x_1^* n_1 + x_2^* n_2 + \ldots + x_K^* n_K}{n} \qquad (10)$$

Маємо:

$$\overline{x}_B = \frac{2\cdot 14 + 4\cdot 2 + 6\cdot 4 + 8\cdot 5 + 10\cdot 3 + 12\cdot 2}{30} = \frac{154}{30} \approx 5{,}13$$

Інтерпретація результату: *отримане середнє вибіркове значення $\overline{x}_B = 5{,}13$ вказує на те, що приріст ВВП підприємств регіону у досліджуваній сукупності порівняно з попереднім роком в середньому склав 5,13%.*

б) Знайдемо моду Mo^* розподілу (варіанту з найбільшою частотою).

Графічно її можна знайти, маючи побудовану гістограму частот або відносних частот, в наступний спосіб: знаходимо найвищий (модальний) прямокутник, з'єднуємо його ліву верхню вершину з лівою верхньою вершиною наступного за ним прямокутника, а праву верхню вершину найвищого прямокутника – з правою верхньою вершиною попереднього прямокутника гістограми. Точку перетину проведених ліній проектуємо на вісь Ox і отримуємо Mo^*. Якщо на гістограмі декілька модальних прямокутників однакової максимальної (порівняно з іншими) висоти, то мод у розподілі буде декілька і для їх знаходження слід проробити вищеописану процедуру з кожним модальним прямокутником.

Якщо найвищий (модальний) прямокутник – самий перший на гістограмі, як у нашому випадку (див. рис. 1), то природно, попереднього прямокутника немає, а отже, праву верхню вершину модального прямокутника у цьому випадку з'єднуємо з його ж нижньою лівою вершиною. Аналогічно, в ситуації, коли найвищий (модальний) прямокутник на гістограмі останній, то його ліва верхня вершина з'єднується з його ж правою нижньою вершиною (бо відсутній наступний прямокутник). З рис. 1 видно, що для нашого прикладу мода дещо більше 2.

Для явного знаходження моди використовується формула

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot h = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} \cdot h, \qquad (11)$$

де

x_{i-1} – початок модального інтервалу (інтервалу з найбільшою частотою n_i);

n_{Mo-1} – частота домодального інтервалу (інтервалу, що передує модальному);

n_{Mo} – частота модального інтервалу;

n_{Mo+1} – частота післямодального інтервалу (інтервалу, що йде одразу після модального);

h – ширина інтервалу розподілу.

Знаходимо модальний інтервал, користуючись таблицею 3.1: найбільша частота (у третьому стовпці) $n_1 = 14$, отже, модальним є перший інтервал [1,3). Оскільки він перший, то домодального інтервалу немає, а отже, його частота $n_{Mo-1} = 0$. Маємо:

$$x_{i-1} = 1;\ n_{Mo-1} = 0;\ n_{Mo} = 14;\ n_{Mo+1} = 2;\ h = 2.$$

Тоді з (11) знаходимо:

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot h = 1 + \frac{14-0}{(14-0)+(14-2)} \cdot 2 = 1 + \frac{14}{26} \cdot 2 = \frac{27}{13} \approx 2,08.$$

Отримане аналітично значення Mo^* справді дещо більше за 2, що попередньо ми встановили графічно (за рис. 1).

Інтерпретація результату: *значення моди $Mo^* = 2,08$ означає, що у вибірці найбільша частка підприємств збільшила свій ВВП на $2,08$ % порівняно з минулим роком.*

в) Знайдемо медіану Me^* розподілу (варіанту, що ділить варіаційний ряд на 2 рівні половини).

Графічно її можна знайти, маючи графік емпіричної функції розподілу $F^*(x)$. Для цього на вісі Oy треба знайти значення 0,5, провести через цю точку пряму, паралельну Ox, до перетину з графіком $F^*(x)$ і точку перетину спроектувати на вісь Ox. Це й буде значення Me^*. З рис. 2 видно, що $Me^* \approx 4$.

Для явного знаходження медіани використовується одна з двох наступних формул

$$Me^* = x_{i-1} + \frac{\frac{1}{2}\sum_{i=1}^{K} n_i - S_{Me-1}}{n_{Me}} \cdot h = x_{i-1} + \frac{\frac{n}{2} - S_{Me-1}}{n_{Me}} \cdot h, \qquad (12)$$

де x_{i-1} – початок медіанного інтервалу (першого інтервалу, на якому нагромаджені частоти більше або дорівнюють половині обсягу вибірки, тобто $S_i \geq n/2$);

S_{Me-1} – нагромаджена частота домедіанного інтервалу (інтервалу, що передує медіанному);

n_{Me} – частота медіанного інтервалу;

n – обсяг вибірки;

h – ширина інтервалу розподілу

або

$$Me^* = x_{i-1} + \frac{0,5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} \cdot h, \qquad (13)$$

де x_{i-1} – початок медіанного інтервалу;

x_i – кінець медіанного інтервалу;

$F^*(x)$ – емпірична функція розподілу;

h – ширина інтервалу розподілу.

Для обчислення медіани за формулою (12) знаходимо спочатку медіанний інтервал, користуючись таблицею 3.1: для цього проглядаємо останній стовпчик, який містить значення нагромаджених частот S_i і шукаємо там перше значення, що дорівнює $n/2 = 30/2 = 15$ чи переважає його. В нашому випадку – це $S_2 = 16 > 15 (= 30/2 = n/2)$, отже, медіанним є другий інтервал [3,5) і

$$x_{i-1} = 3;\ S_{Me-1} = 14;\ n_{Me} = 2;\ n = 30;\ h = 2.$$

Підставивши знайдені значення у (12), знаходимо

$$Me^* = x_{i-1} + \frac{\frac{n}{2} - S_{Me-1}}{n_{Me}} \cdot h = 3 + \frac{\frac{30}{2} - 14}{2} \cdot 2 = 4.$$

Якщо зручніше скористатись формулою (13), то попередньо знаходимо, користуючись аналітичним виглядом (8) емпіричної функції розподілу $F^*(x)$

$$F^*(x_{i-1}) = F^*(3) = \frac{14}{30},\ F^*(x_i) = F^*(5) = \frac{16}{30}.$$

Тоді згідно з (13) маємо:

$$Me^* = 3 + \frac{0,5 - \frac{14}{30}}{\frac{16}{30} - \frac{14}{30}} \cdot 2 = 4.$$

Аналітичні розрахунки підтвердили наш результат, попередньо встановлений графічним способом.

Інтерпретація результату: *значення медіани $Me^* = 4$ вказує на те, що половина підприємств із досліджуваної вибірки має приріст ВВП, менший від 4% (порівняно з минулим роком), а приріст ВВП іншої половини – більше 4%.*

Зауваження 2: величини $\overline{x}_B$, Mo^*, Me^* є показниками центру розподілу. Кожну з цих величин використовують для опису властивостей ряду розподілу різного типу, адже не завжди $\overline{x}_B$ адекватно відображає реальну тенденцію («ситий і голодний – в середньому поїли»). Для симетричних рядів розподілу

$$\overline{x}_B = Mo^* = Me^*,$$

а в помірно асиметричних рядах

$$3\left(\overline{x}_B - Me^*\right) \approx \overline{x}_B - Mo^*.$$

Для характеризації асиметричних рядів використовують не $\overline{x}_B$, а Me^*.

У нашому випадку $\overline{x}_B \neq Mo^* \neq Me^*$, що свідчить про несиметричність ряду розподілу. Маємо: $\overline{x}_B = \frac{154}{30} \approx 5{,}13;\ Mo^* = \frac{27}{13} \approx 2{,}08;\ Me^* = 4$. Тоді

$$3\left(\overline{x}_B - Me^*\right) \approx 3(5{,}13 - 4) = 3 \cdot 1{,}13 = 3{,}39,$$

$$\overline{x}_B - Mo^* \approx 5{,}13 - 2{,}08 = 3{,}05,$$

отже, у розподілі є виражена асиметрія (він не є ні симетричним, ні помірно асиметричним).

На характер асиметрії вказуватиме показник коефіцієнту асиметрії A_s^*, який ми знайдемо у п. 3.3.

3.2) Знайдемо показники варіації (D_B, σ_B, R, V ; S^2, S) та оцінимо однорідність сукупності.

Дисперсія D_B може бути обчислена за однією з двох формул

$$D_B = \frac{\sum_{i=1}^{K}(x_i - \overline{x}_B)^2 n_i}{n} \qquad (14)$$

або

$$D_B = \frac{\sum_{i=1}^{K} x_i^2 n_i}{n} - \left(\overline{x}_B\right). \qquad (15)$$

Модифікуємо таблицю 3.1 (видалимо останній стовпець (нагромаджені частоти) і додамо стовпець квадратів x_i^2 дискретних варіант) для зручного використання формули (15):

Таблиця 3.2

Інтервали $I_i, h=2$	Середини інтервалів $x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}$	x_i^2	Частоти n_i
[1,3)	2	4	14
[3,5)	4	16	2
[5,7)	6	36	4
[7,9)	8	64	5
[9,11)	10	100	3
[11,13]	12	144	2

Тоді

$$D_B = \frac{\sum_{i=1}^{K} x_i^2 n_i}{n} - \left(\overline{x}_B\right) = \frac{4\cdot14 + 16\cdot2 + 36\cdot4 + 64\cdot5 + 100\cdot3 + 144\cdot2}{30} - \left(\frac{154}{30}\right)^2 =$$

$$= \frac{1140}{30} - \frac{23716}{900} = \frac{10484}{900} \approx 11{,}65.$$

Середньоквадратичне відхилення σ_B знайдемо за формулою

$$\sigma_B = \sqrt{D_B}. \qquad (16)$$

Маємо:

$$\sigma_B = \sqrt{D_B} = \sqrt{\frac{10484}{900}} \approx 3{,}413.$$

Інтерпретація результату: *значення $\sigma_B = 3{,}413$ показує, що в середньому приріст ВВП по розглядуваних підприємствах відхиляється від вибіркового середнього $\overline{x}_B$ на 3,413%.*

Зауваження 3: *при малих обсягах вибірки вибіркова дисперсія є зміщеною оцінкою, тому при $n \le 30$ для більш точного оцінювання використовують виправлену дисперсію S^2 та виправлене середньоквадратичне відхилення S, які обчисляються за формулами*

$$S^2 = \frac{n}{n-1} D_B, \qquad (17)$$

де n - обсяг вибірки, а D_B - вибіркова дисперсія, обчислена за формулою (14) чи (15).

$$S = \sqrt{S^2}. \qquad (18)$$

Зрозуміло, що

$$S^2 > D_B,\ \textit{бо}\ \frac{n}{n-1} > 1,\ \textit{а значить, і}\ S > \sigma_B.$$

У нашому прикладі

$$S^2 = \frac{n}{n-1} D_B = \frac{30}{30-1} \cdot \frac{10484}{900} = \frac{10484}{870} \approx 12,05$$ - виправлена дисперсія;

$$S = \sqrt{S^2} = \sqrt{\frac{10484}{870}} \approx 3,471$$ - виправлене середньоквадратичне відхилення.

Розмах вибірки R знайдемо за формулою

$$R = x_{\max} - x_{\min}. \qquad (19)$$

У п. 1 ми його вже знаходили, коли шукали ширину інтервалу поділу ($x_{\min} = 1$, $x_{\max} = 13$)

$$R = x_{\max} - x_{\min} = 13 - 1 = 12.$$

Інтерпретація результату: *значення* $R = 12$ *вказує на те, що приріст ВВП підприємств досліджуваної вибірки коливається в межах 12%.*

Останній з показників варіації, за яким можна робити висновок про однорідність вибірки або порівнювати вибірки з різними значеннями $\overline{x}_B$ - це коефіцієнт варіації. Його обчисляють за формулою

$$V = \frac{\sigma_B}{\overline{x}_B} \cdot 100\% \qquad (20)$$

та інтерпретують отримане значення за таблицею 4. Фактично коефіцієнт оцінює інтенсивність коливань варіант відносно їх середньої величини

У нашому прикладі

$$V = \frac{\sigma_B}{\overline{x}_B} \cdot 100\% = \frac{3,413}{5,13} \cdot 100\% \approx 66,53\%$$

Таблиця 4

№	*Межі коефіцієнта варіації, %*	*Висновок про однорідність вибірки*
1	$0 \le V \le 33$	вибірка однорідна
2	$33 < V < 66$	вибірка помірно неоднорідна
3	$V \ge 66$	вибірка істотно неоднорідна

Зауваження 4: *В деяких джерелах зауважують, що ранжування, подане у табл. 4, коректне для вибірок, розподіл яких близький до нормального, а для інших типів вибірок коефіцієнт варіації слугує критерієм надійності вибіркової середньої: при* $0 \le V \le 40$ *вибіркова середня* $\overline{x}_B$ *надійна, тобто вибірка однорідна (немає ефекту «ситий і голодний – в середньому поїли»), а при* $V > 40$ *- вибіркова середня* $\overline{x}_B$ *не надійна, а вибірка неоднорідна.*

Інтерпретація результату: *значення* $V = 66,53\% > 66\%$, *отже, вибірка істотно неоднорідна.*

3.3) **Центральні емпіричні моменти *k*-го порядку**. Середнє зважене відхилення варіант у степені *k* (*k*=1,2,3,…) називають ***центральним емпіричним моментом* k*-го порядку***

$$\mu_k^* = \frac{\sum_{i=1}^{K} (x_i - \overline{x}_B)^k n_i}{n} \qquad (21)$$

При k=1 з (21) маємо:

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

отже, ***перший центральний момент для будь-якого розподілу дорівнює нулеві***:

$$\mu_1^* = 0.$$

При k=2 з (21) маємо:

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

отже, ***другий центральний момент для будь-якого розподілу дорівнює дисперсії вибірки***.

Важливими для характеристики вибірки є третій μ_3^* та четвертий μ_4^* центральні моменти, з допомогою яких обчислюються коефіцієнт асиметрії A_s^* та ексцес E_s^*.

$$A_s^* = \frac{\mu_3^*}{\sigma_B^3}, \qquad (22)$$

де

$$\mu_3^* = \frac{\sum_{i=1}^{K} (x_i - \overline{x}_B)^3 n_i}{n} \qquad (23)$$

Якщо варіанти статистичного розподілу вибірки симетрично розміщені відносно $\overline{x}_B$, то в цьому разі $A_s^* = 0$, оскільки $\mu_3^* = 0$. Якщо вибірка має нормальний закон розподілу, то для неї $A_s^* = 0$. Показники асиметрії оцінюють зміщення ряду розподілу вліво або вправо по відношенню до осі симетрії нормального розподілу. Загальний вигляд гістограм при симетричному та обох видах асиметричного розподілу наведено на рис. 3.

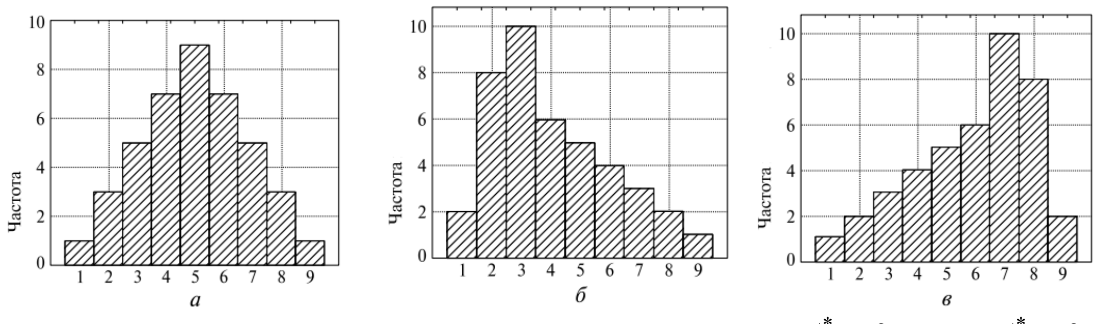


Рис. 3. *Гістограми симетричного (а) розподілу та розподілів з* $A_s^* > 0$ *(б) та* $A_s^* < 0$ *(в)*

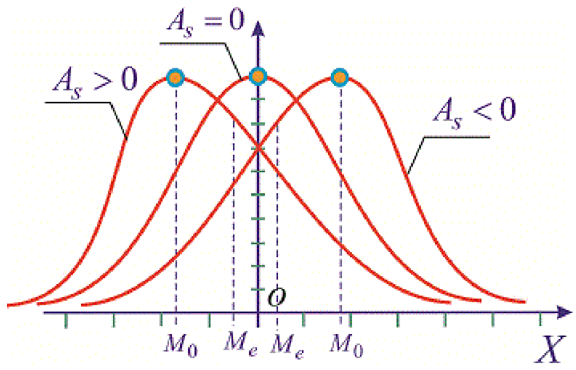


Рис. 3.1. *Види асиметрії*

У разі асиметричного розподілу вершина кривої, що наближає графік щільності розподілу, знаходиться не в середині, а зміщена або вліво, або вправо (рис. 3.1). Якщо вершина зміщена вліво, то права частина кривої виявляється довше лівої, тобто має місце ***правостороння асиметрія*** (***$A_s^* > 0$***), що характеризується нерівністю $\overline{x}_B > Me^* > Mo^*$.

Якщо ж вершина кривої зміщена вправо, то ліва частина кривої виявляється довша за праву, тобто має місце ***лівостороння асиметрія*** (***$A_s^* < 0$***), для якої має місце нерівність $\overline{x}_B < Me^* < Mo^*$.

Таблиця 5

№	*Знак коефіцієнта асиметрії*	*Висновок про симетричність вибірки*	*Характеристика середніх вибірки*
1	$A_s^* = 0$	вибірка симетрична відносно $\overline{x}_B$	$\overline{x}_B = Me^* = Mo^*$
2	$A_s^* < 0$	від'ємна (лівостороння) асиметрія	$\overline{x}_B < Me^* < Mo^*$
3	$A_s^* > 0$	додатна (правостороння) асиметрія	$\overline{x}_B > Me^* > Mo^*$

Чим більше за модулем значення коефіцієнта асиметрії, тим більш асиметричним є розподіл. У деяких літературних джерелах встановлюють таке ранжування:

Таблиця 5.1

№	*Межі коефіцієнта асиметрії*	*Висновок про асиметрію розподілу*
1	$0 < \left\| A_s^* \right\| \le 0,25$	асиметрія незначна
2	$0,25 \le \left\| A_s^* \right\| \le 0,5$	асиметрія помірна
3	$\left\| A_s^* \right\| > 0,5$	асиметрія суттєва

Ексцес E_s^* є показником «крутості» (гостровершинності) варіаційного ряду в порівнянні з нормальним розподілом (табл. 6). Якщо ексцес додатний, то полігон варіаційного ряду має більш крутий шпиль. Це говорить про скупчення значень ознаки в центральній зоні ряду розподілу, тобто про переважну появу у вибірці значень, близьких до середньої величини. Якщо ексцес від'ємний - то полігон має більш пологу вершину в порівнянні з нормальною кривою. Це означає, що значення варіант не концентруються в центральній частині ряду, а розсіяні по всьому діапазону від мінімального до максимального значення. Чим більше абсолютна величина ексцесу, тим істотніше розподіл відрізняється від нормального. Для нормального закону розподілу, як відомо, $E_s^* = 0$. Ексцес обчислюється за формулою

$$E_s^* = \frac{\mu_4^*}{\sigma_B^4} - 3, \qquad (24)$$

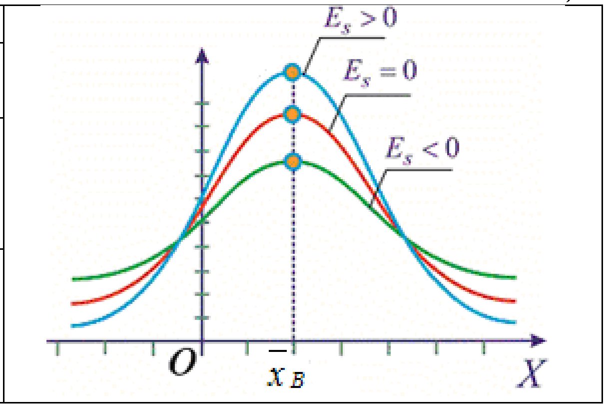
де

$$\mu_4^* = \frac{\sum_{i=1}^{K} (x_i - \overline{x}_B)^4 n_i}{n}. \qquad (25)$$

Зауваження 5: *при обчисленні коефіцієнта асиметрії та ексцесу для дискретного розподілу використовують формули (23) та (25), а у випадку інтервального розподілу варіанти x_i замінюють на x_i^* - середини інтервалів розподілу.*

Таблиця 6

№	*Знак ексцесу*	*Висновок*
1	$E_s^* = 0$	нормальний розподіл
2	$E_s^* < 0$	туповершинний розподіл
3	$E_s^* > 0$	гостровершинний розподіл



За значеннями показників асиметрії і ексцесу розподілу можна судити про близькість розподілу до нормального, що буває важливо для оцінки результатів кореляційного і регресійного аналізу, можливостей імовірнісної оцінки прогнозів.

Для зручного обчислення A_s^* та E_s^* видозмінимо таблицю 3.2, вилучивши стовпець x_i^2 та доповнивши стовпцями $(x_i - \overline{x}_B)^3$, $(x_i - \overline{x}_B)^4$, $(x_i - \overline{x}_B)^3 \cdot n_i$, $(x_i - \overline{x}_B)^4 \cdot n_i$ та рядком «сума».:

Таблиця 7

Інтервали $I_i, h=2$	*Середини інтервалів* x_i^*	*Частоти* n_i	$(x_i - \overline{x}_B)^3$	$(x_i - \overline{x}_B)^3 \cdot n_i$	$(x_i - \overline{x}_B)^4$	$(x_i - \overline{x}_B)^4 \cdot n_i$
[1,3)	2	14	-30,7623704	-430,673185	96,38876049	1349,443
[3,5)	4	2	-1,4557037	-2,91140741	1,649797531	3,299595
[5,7)	6	4	0,650962963	2,603851852	0,564167901	2,256672
[7,9)	8	5	23,55762963	117,7881481	67,5318716	337,6594
[9,11)	10	3	115,2642963	345,7928889	560,9529086	1682,859
[11,13]	12	2	323,770963	647,5419259	2223,227279	4446,455
		Сума	---	**680,1422222**	---	**7821,97**

Це дозволить нам у 5-му та 7-му стовпцях рядка «сума» отримати чисельники формул (23) та (25) відповідно. Обчислення зручно проводити, використовуючи табличний процесор MS EXCEL (або власно створений програмний продукт).

Маємо:

$$\mu_3^* = \frac{\sum_{i=1}^{K}(x_i - \overline{x}_B)^3 n_i}{n} = \frac{680,14}{30} \approx 22,671.$$

Тоді за формулою (22) знаходимо коефіцієнт асиметрії

$$A_s^* = \frac{\mu_3^*}{\sigma_B^3} = \frac{22,671}{(3,413)^3} \approx \frac{22,671}{39,757} \approx 0,57$$

Інтерпретація результату: *значення* $A_s^* = 0,57 > 0$; *відповідно до таблиці 5.1 маємо суттєву додатну асиметрію, отже, вершина щільності розподілу зміщена вліво. До цього ж висновку приводить виконання нерівності* $\overline{x}_B > Me^* > Mo^*$: $5,13 > 4 > 2,08$

Знайдемо четвертий центральний момент та ексцес вибірки. Підставляючи число, знайдене у 7-му стовпці рядка «Сума» таблиці 7, у формулу (25), знайдемо

$$\mu_4^* = \frac{\sum_{i=1}^{K}(x_i - \overline{x}_B)^4 n_i}{n} = \frac{7821,97}{30} \approx 260,73\,.$$

Далі за формулою (24) обчисляємо ексцес

$$E_s^* = \frac{\mu_4^*}{\sigma_B^4} - 3 = \frac{260,73}{(3,413)^4} - 3 = \frac{260,73}{135,689} - 3 \approx -1,078\,.$$

Інтерпретація результату: *значення* $E_s^* = -1,078 < 0$, *отже, розподіл туповершинний (табл. 6).*

Завдання для самостійного розв'язання

Задача. *За вхідними даними Вашого варіанту (номер варіанта відповідає номеру студента в журналі академгрупи):*

1) Побудувати відповідний інтервальний розподіл (для визначення кількості інтервалів використати формулу Стерджеса).
2) Побудувати гістограму частот; емпіричну функцію розподілу $F^*(x)$ і її графік.
3) Знайти числові характеристики отриманого інтервального розподілу:
   3.1) показники центру розподілу $\overline{x}_B$, Mo^*, Me^*;
   3.2) показники варіації (D_B, σ_B, R, V; S^2, S) та оцінити однорідність сукупності;
   3.3) показники форми розподілу (A_s^* та E_s^*).

Інтерпретувати отримані результати.

Варіант 1 (Байрамов Алі)

Наявні дані про число тонн вантажів, які перевозяться щотижня поромом деякого морського порту в період навігації:
398, 412, 560, 474, 544, 690, 587, 600, 613, 457, 504, 477, 530, 641, 359, 566, 452, 633, 474, 499, 580, 606, 344, 455, 505, 396, 347, 441, 390, 632, 400, 582.

Варіант 2 (Беленчук Олексій)

Наявні дані про денну виручку супермаркету (у тисячах гривень):
221 233 180 215 235 260 201 234 211 237 200 254 245 207 243 251 210 245 250 223 223 265 255 239 195 250 245 227 231 256

Варіант 3 (Березний Ігор)

Число пасажирів компанії «МОТОР-СІЧ» одного з рейсів протягом місяця між Запоріжжям і Києвом склало:
128, 121, 134, 118, 123, 109, 120, 116, 125, 128, 121, 129, 130, 131, 127, 119, 114, 124, 110, 126, 134, 125, 128, 123, 128, 133, 132, 136, 134, 129

Варіант 4 (Бужак Андрій)

Наявний розподіл абітурієнтів за числом балів, отриманих ними на тестуванні з математики:
200, 197, 122, 124, 175, 183, 135 176, 200, 161,152, 123, 151, 144, 189, 182, 167, 183, 191, 200, 124, 190, 200, 185, 174, 156, 200, 165, 163, 185

Варіант 5 (Бурле Павло)

З метою визначення продуктивності праці робітників цеху пораховано кількість деталей, виготовлених кожним робочим за зміну:
42 43 38 40 43 38 40 41 39 42 42 43 44 42 44 43 40 42 41 32 43 42 39 38 45 39 42 40 41 39

Варіант 6 (Василевич Павло)

Обстеження підприємств області щодо зростання виробітку на одного робітника (у % до попереднього року) дали результати:
80 100 91 102 103 115 122 118 119 120 104 82 107 111 102 112 115 116 114 115 106 92 109 117 99 98 118 108 107 112.

Варіант 7 (Волощук Назарій)

Фіксувались кращі результати стрибків у довжину легкоатлетів (см):
795 795 796 797 803 789 804 790 783 804 804 795 796 797 803 799 804 799 800 804 833 830 832 837 834 832 835 836 835 833.

Варіант 8 (Георгіян Євген)

Наведено вагу пацієнтів (в кг), які бажають пройти курс лікування, щоб знизити вагу:
103 90 95 101 106 79 98 91 87 120 93 88 111 82 84 86 81 95 79 83 105 77 61 74 90 82 93 88 67 95

Варіант 9 (Гончаров Олександр)

Наявний розподіл розмірів пар чоловічого взуття, що продані магазином протягом дня:
39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44

Варіант 10 (Григорчук В'ячеслав)

Вимірювання продуктивності праці бригади шахтарів (в тонах вугілля за зміну) дали такі результати :
331 356 391 299 364 317 386 360 402 390 365 411 383 374 357 306 427 382 381 373 303 338 315 401 334 345 367 324 385 326

Варіант 11 (Денис Денис)

Вимірювання витрат бензину легкового автомобіля на 1000 км пробігу дали такі результати (л):
90 90 89 89 91 91 90 89 91 90 86 89 88 86 85 86 87 88 90 77 79 81 78 81 79 80 79 78 87 92.

Варіант 12 (Дісар Іван)

Вимірювання довжини заготовок після їх первісної обробки дали такі результати (в міліметрах):
1151 1158 1152 1155 1160 1151 1154 1156 1160 1151 1153 1155 1154 1158 1151 1158 1151 1154 1153 1157 1154 1154 1152 1154 1155 1152 1153 1158 1157 1155.

Варіант 13 (Дручук Роман)

Дослідження тривалості годин горіння електролампочок дали такі результати:
750 751 756 769 757 760 743 745 759 750 739 752 758 750 758 753 747 749 751 762 748 754 752 739 755 744 764 760 739 748.

Варіант 14 (Дубець Василь)

Маємо дані про місячній зарплаті робітників (в гривнях) одного з цехів:
3015 2870 2965 2935 2989 2750 2972 3190 3140 3244 3007 3019 2825 3027 2925 3070 2839 3085 3040 3133 3090 3190 3112 3015 3279 2970 2890 3032 2948 3045

Варіант 15 (Дуплава Олександр)

Через кожну годину вимірювалася напруга струму в електромережі. При цьому були отримані наступні значення (в вольтах):
227, 229, 215, 230, 232, 223, 220, 222, 218, 219, 222, 221, 227, 226, 226, 209, 211, 215, 218, 220, 216, 220, 220, 221, 225, 224, 212, 217, 219, 220.

Варіант 16 (Жупник Евеліна)

Маємо дані про денну виручку продовольчого магазину (грн.):
5023 4970 4985 4974 5048 4996 5016 4978 4918 5028 4676 4970 4985 4638 4720 4816 4935 4917 4918 5028 3718 3812 3716 3875 3788 3852 3794 3790 3802 3755