Розділ 1. ОСНОВНІ ПОНЯТТЯ МАТЕМАТИЧНОЇ СТАТИСТИКИ

1.1. Поняття вибіркового методу в статистиці

Математична статистика – це розділ прикладної математики, предметом якого є розробка раціональних прийомів і методів отримання, опису та обробки експериментальних даних з метою вивчення закономірностей масових випадкових явищ.

Основними завданнями математичної статистики є:

– визначення за статистичними даними законів розподілу випадкових величин;

– визначення за статистичними даними параметрів розподілу випадкових величин;

– визначення за статистичними даними виду зв'язку між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкта);

– визначення сили (тісноти зв'язку) між різними явищами (об'єктами) або властивостями одного і того ж явища (об'єкта);

– перевірка вірогідності статистичних гіпотез;

– розробка рекомендацій щодо проведення експерименту та обробки його результатів.

У прикладних дослідженнях, зазвичай, необхідно вивчити сукупність однорідних об'єктів або спостережень за якою-небудь кількісною або якісною ознакою.

Сукупність об'єктів або спостережень, всі елементи якої підлягають вивченню при статистичному аналізі, називається **генеральною сукупністю.**

Генеральна сукупність може бути скінченою або нескінченною. Так, при вивченні розподілу населення за родом занять розглядається велика, але скінчена генеральна сукупність об'єктів. При вивченні впливу яскравості освітлення робочого місця на продуктивність праці працівника генеральна сукупність спостережень теоретично нескінченна, оскільки яскравість освітлення може змінюватися неперервно у межах певного інтервалу.

Кількість об'єктів (спостережень) генеральної сукупності називається **об'ємом генеральної сукупності** і позначається $\boldsymbol{N}$.

На практиці рідко є можливість досліджувати кожний елемент генеральної сукупності, оскільки це пов'язано з великими витратами засобів, коштів і часу, а іноді з псуванням або знищенням досліджуваних об'єктів. У деяких випадках дослідити всі об’єкти генеральної сукупності взагалі неможливо. Тому при статистичному аналізі, як правило, вивчається не вся генеральна сукупність, а деяка її частина.

Частина об'єктів генеральної сукупності, використовувана в ході дослідження, називається **вибіркою**. У соціологічному дослідженні об’єкти вивчення називаються **респондентами**.

Кількість об'єктів (спостережень) вибірки називається її **об'ємом** і позначається $\boldsymbol{n}$.

Наприклад, продукція у кількості N одиниць, вироблена підприємством

на протязі року, є генеральною сукупністю. Для дослідження якості продукції на практиці розглядається вибірка, що складається з *n* одиниць продукції. Ознакою якості в даному дослідженні служить відповідність вибраної одиниці товару сертифікатним вимогам.

Ціль вибіркового методу в статистиці полягає в тому, що висновки, зроблені на основі вивчення вибірки, розповсюджуються на всю генеральну сукупність.

Слід зазначити, що незалежно від способу організації вибірки вона повинна правильно відображати кількісні та якісні співвідношення генеральної сукупності, тобто бути **репрезентативною**. Крім того, всі елементи генеральної сукупності повинні мати однакову ймовірність бути відібраними у вибірку, тобто вибірка має бути **випадковою**.

Існує кілька типів ймовірнісної вибірки, які відрізняються між собою характером використаних дослідником прийомів:

– проста ймовірнісна вибірка, яка проводиться шляхом випадкового відбору об’єктів у вибірку;

– стратифікована вибірка, що використовується тоді, коли цілі та завдання дослідження вимагають відбору об’єктів для вивчення за певними груповими критеріями;

– багатоступінчаста вибірка, для якої характерно декілька послідовних змін одиниць відбору.

Для результатів, що отримані при вибірковому дослідженні, необхідна перевірка на точність і статистичну значущість; спосіб формування вибірки та її об’єм повинні відповідати певному методу обробки даних.

1.2. Шкали вимірювань

Статистичній обробці підлягають тільки ті ознаки об’єктів або фактори, які можна виміряти за деякою шкалою. **Шкала** − числова система, що відображає досліджувані властивості та ознаки об’єкта. Наприклад, на етапі розробки інструментарію дослідник-соціолог, чітко визначившись з: а) метою, б) завданнями, в) об’єктом і г) предметом дослідження, будує систему оціночних суджень (запитань) з розгорнутою системою підказок (альтернатив відповіді). Це і є шкала, в якій має бути відображений весь діапазон думок досліджуваної ознаки (запитання).

Існують такі шкали вимірювань: шкала найменувань (номінальна); шкала порядку; шкала інтервалів; шкала відношень.

Шкала найменувань (класифікації, номінальна)

Якщо дані вимірюються за шкалою найменувань, то над ними можливі тільки операції порівняння, такі як „рівні” або „нерівні”. Дані номінальної шкали необхідні для ідентифікації певного об’єкта – місце розташування деякої організації, номер філіалу деякої фірми, номер методики навчання і т. ін.

Шкала порядку

Якщо дані вимірюються за шкалою порядку, їх можна порівняти за величиною „більше”, „менше” або „рівні”. За такою шкалою вимірюються, наприклад, експертні оцінки, вік респондентів і т. ін.

Шкала інтервалів

Якщо дані вимірюються за шкалою інтервалів, до них можна застосувати операції: порівняння − „більше”, „менше”, „рівні”; додавання і віднімання. Прикладом даних, які належать до цієї шкали, є результати вимірювання температури, тиску.

Шкала відношень

Якщо дані вимірюються за шкалою відношень, їх можна порівняти за величиною та виконати всі арифметичні операції: додавання, віднімання, множення і ділення. Такою шкалою кодується вага, маса, ріст, довжина, дохід, обсяг виробництва і т. ін.

В емпіричних дослідженнях найбільш використовуваною вважається шкала порядкового типу.

Від шкали вимірювання даних залежить можливість знаходження числових характеристик (табл. 1.1) і застосовування певного статистичного методу.

Таблиця 1.1 – Шкали вимірювань і числові характеристики

Назва шкали	**Числові характеристики**
Найменувань	Частоти
Порядку	Частоти, середнє, мода, медіана
Інтервалів	Частоти, середнє, мода, медіана, дисперсія
Відношень	Всі відомі

Для об’єктивного визначення типу шкали користуються способом її створення на основі шкали Лайкерта. Ця методика має ще назву «метод сумарних оцінок». На першому етапі група дослідників розробляє набір (навіть до сотні одиниць) роздумів відносно предмета вивчення. Наступним етапом є пілотажне дослідження на невеликій вибірці. Після отримання його результатів проводиться якісний аналіз набору даних і відсіюються малоефективні значення ознак.

1.3. Статистичні ряди та їх графічна інтерпретація

Припустимо, що необхідно вивчити деяку ознаку генеральної сукупності X, для чого було проведено n вимірювань цієї ознаки і складено вибірку її значень $\{x_1, x_2, ..., x_n\}$ об'єму n.

Різні елементи вибірки називаються **варіантами**. Число n_i, що показує, скільки разів варіанта x_i зустрічається у вибірці, називається **частотою варіанти**. Число w_i, що дорівнює відношенню частоти варіанти n_i до об'єму вибірки n, називається **відносною частотою варіанти** x_i:

$$w_i = \frac{n_i}{n}. \qquad (1.1)$$

Ряд варіант, розташованих в порядку зростання їх значень, називається **варіаційним рядом**. Послідовність, що складається із варіант і відповідних їм частот (відносних частот), називається **статистичним рядом** або **рядом розподілу**. Групування кількісних результатів вимірювань у вигляді статистичних рядів є необхідним для застосування статистичних методів

аналізу даних і побудови статистичних моделей.

Ознака X є випадковою величиною, а статистичний ряд – емпіричним (тобто отриманим у результаті експерименту або спостережень) законом її розподілу.

Статистичний ряд називається **дискретним**, якщо він є законом розподілу дискретної випадкової величини, та **інтервальним**, якщо він є законом розподілу неперервної випадкової величини.

Дискретний статистичний ряд у загальному вигляді можна зобразити таблицею (табл. 1.2):

Таблиця 1.2

Варіанти x_i	x_1	x_2	…	x_k
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

де k – кількість варіант.

Інтервальний статистичний ряд у загальному вигляді можна представити таблицею (див. табл. 1.3):

Таблиця 1.3

Інтервали $[a_i; a_{i+1})$	$[a_1; a_2)$	$[a_2; a_3)$	…	$[a_{к-1}; a_к)$
Частоти n_i (відносні частоти w_i)	n_1 (w_1)	n_2 (w_2)	…	n_k (w_k)

де k – кількість інтервалів.

Для статистичних рядів виконуються рівності: $\sum_{i=1}^{k} n_i = n$, $\sum_{i=1}^{k} w_i = 1$.

Для побудови інтервального статистичного ряду множину значень варіант розбивають на інтервали $[a_i; a_{i+1})$, тобто проводять їх групування. Кількість інтервалів k рекомендується розраховувати за формулою Стерджерса:

$$k = 1 + 1,4 \ln n . \qquad (1.2)$$

Довжина кожного із інтервалів Δ розраховується за формулою

$$\Delta = \frac{x_{\max} - x_{\min}}{k}, \qquad (1.3)$$

де $x_{\max}$, $x_{\min}$ – максимальне і мінімальне значення у варіаційному ряді.

Підраховуючи кількість значень варіант, що потрапили в інтервал $[a_i; a_{i+1})$, отримують частоти n_i для $i = \overline{1,k}$.

Для наочності використовують графічне зображення статистичних рядів у вигляді полігону частот (відносних частот) та, виключно у випадку інтервального ряду, гістограми.

Полігоном частот (відносних частот) називається ламана лінія, що сполучає точки з координатами: (x_i; n_i) або (x_i; w_i) для $i = \overline{1,k}$ у разі дискретного статистичного ряду; (c_i; n_i) або (c_i; w_i) у разі інтервального ряду, де c_i – середина

i-того інтервалу, $c_i = \dfrac{a_i + a_{i+1}}{2}$.

Гістограмою називається ступінчаста фігура, яка складається з прямокутників з основами, що дорівнюють довжинам інтервалів Δ, та висотами, які пропорційні частотам n_i (відносним частотам w_i) і обчислюються як відношення частот n_i (відносних частот w_i) до довжин Δ відповідних інтервалів. Площа гістограми частот дорівнює об'єму вибірки n, а площа гістограми відносних частот дорівнює одиниці.

За статистичним рядом можна встановити емпіричну функцію розподілу та емпіричну щільність розподілу випадкової величини X.

Емпіричною функцією розподілу називається функція

$$F_n(x) = \frac{1}{n}\sum_{x_i<x} n_i = \sum_{x_i<x} w_i \,. \qquad (1.4)$$

Відмітимо, що для інтервального ряду вказуються не конкретні значення варіант, а тільки їх частоти на інтервалах. Тому емпірична функція розподілу визначена тільки на кінцях інтервалів.

Кумулятою називається крива, що проходить через точки з координатами: $(a_i;\ F_n(a_i))$, де $i = \overline{1,k}$.

Для зображення результатів соціологічних досліджень використовують **огіву Гальтона** – криву, яка сполучає точки з координатами: $(F_n(a_i),\ a_i)$, де $i = \overline{1,k}$.

Емпіричною щільністю розподілу для інтервального ряду називається функція

$$f_n(x) = \begin{cases} \dfrac{n_i}{n\Delta} = \dfrac{w_i}{\Delta}, \text{ якщо } a_i \le x \le a_{i+1},\ i = \overline{1,k} \\ 0, \quad \text{якщо } x < a_1 \text{ або } x > a_{k+1},\ i = \overline{1,k} \end{cases}. \qquad (1.5)$$

Приклад 1.1. У результаті тестування службовців деякої компанії були отримані такі результати (у балах): 39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 38, 42, 41, 43, 39, 37, 43, 41, 38, 42, 40, 41, 42, 40, 41. Побудувати дискретний статистичний ряд для випадкової величини X – оцінки службовців, полігон частот, емпіричну функцію розподілу та її графік.

Розв'язок. Для побудови дискретного статистичного ряду записуємо у порядку зростання різні значення випадкової величини X і відповідні частоти (табл. 1.4). Останній стовпчик таблиці використовується для перевірки правильності побудови статистичного ряду (усього у тестуванні приймали участь 25 осіб, тому сума частот повинна дорівнювати 25).

Таблиця 1.4

x_i	37	38	39	40	41	42	43	44	Сума
n_i	1	2	2	4	6	5	4	1	$\sum_{i=1}^{k} n_i = 25$

Полігон частот даного розподілу зображено на рис. 1.1.

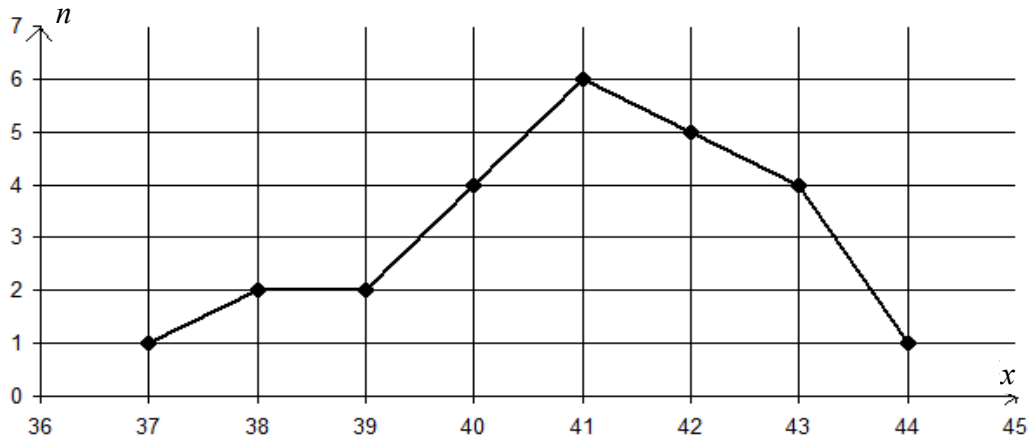

Рисунок 1.1. Полігон частот

Для побудови емпіричної функції розподілу доповнимо таблицю двома рядками (табл. 1.5). В першому рядку обчислимо суму частот варіант, що менше x_i (тобто $\sum_{x_i<x} n_i$). Отримаємо:

якщо $x < x_1 = 37$, то $\sum_{x_i<37} n_i = 0$, оскільки таких значень X немає;

якщо $x < x_2 = 38$, то $\sum_{x_i<38} n_i = n_1 = 1$;

якщо $x < x_3 = 39$, то $\sum_{x_i<39} n_i = n_1 + n_2 = 1 + 2 = 3$;

якщо $x < x_4 = 40$, то $\sum_{x_i<40} n_i = n_1 + n_2 + n_3 = 1 + 2 + 2 = 5$;

якщо $x < x_5 = 41$, то $\sum_{x_i<41} n_i = n_1 + n_2 + n_3 + n_4 = 1 + 2 + 2 + 4 = 9$;

якщо $x < x_6 = 42$, то $\sum_{x_i<42} n_i = n_1 + n_2 + n_3 + n_4 + n_5 = 1 + 2 + 2 + 4 + 6 = 15$;

якщо $x < x_7 = 43$, то $\sum_{x_i<43} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 1 + 2 + 2 + 4 + 6 + 5 = 20$;

якщо $x < x_8 = 44$, то $\sum_{x_i<44} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 + n_7 = 20 + 4 = 24$;

якщо $x > x_8 = 44$, то $\sum_{x_i<x} n_i = 25$ – це означає, що всі значення X менші числа, більшого за 44.

В другому рядку запишемо значення функції, обчислені за формулою (1.4). Графік отриманої емпіричної функції розподілу зображено на рис. 1.2.

Таблиця 1.5

x_i	37	38	39	40	41	42	43	44	
n_i	1	2	2	4	6	5	4	1	
$\sum_{x_i<x} n_i$	0	1	3	5	9	15	20	24	25
F_i	0	0,04	0,12	0,2	0,36	0,6	0,8	0,96	1

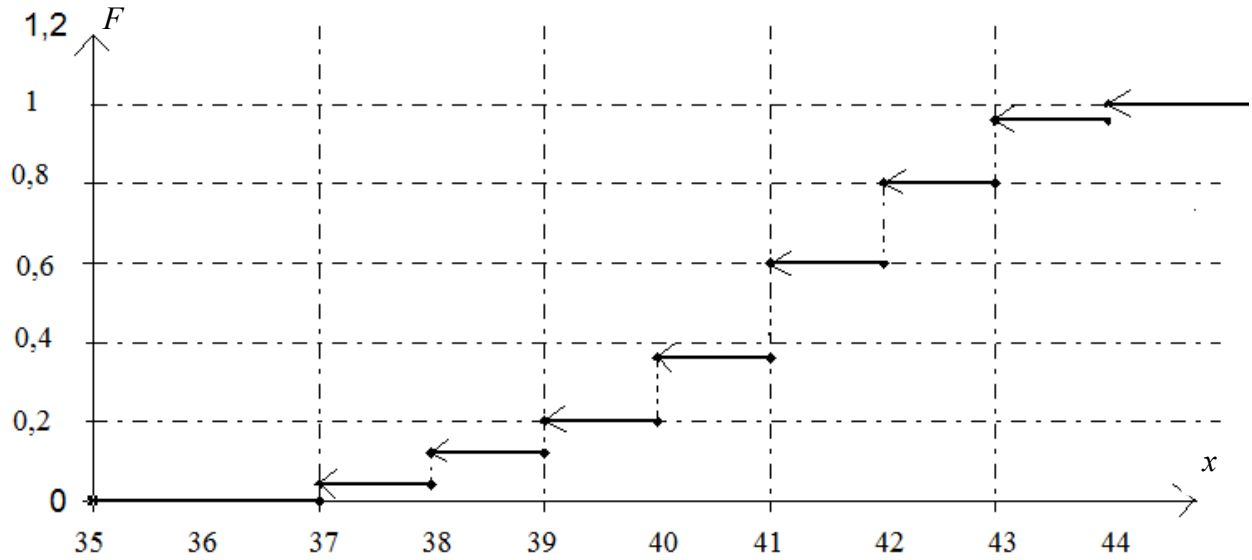

Рисунок 1.2. Графік емпіричної функції розподілу

Приклад 1.2. За даними вибіркового дослідження було отримано розподіл родин за доходом на одного їх члена в умовних одиницях (табл. 1.6). Побудувати інтервальний статистичний ряд, полігон частот, гістограму, полігон відносних частот, емпіричні функцію і щільність розподілу та їх графіки.

Таблиця 1.6

28,92	27,54	22,36	29,09	32,19	26,04	17,06	26,83	24,55	33,22
17,53	30,07	36,27	24,24	26,03	31,05	13,94	14,56	21,40	23,04
13,09	38,84	25,57	22,87	6,11	27,79	25,68	16,30	17,93	24,37
28,92	27,54	22,36	29,06	32,19	26,04	17,06	26,83	24,55	33,22
17,53	30,07	36,27	24,24	26,03	31,05	13,94	14,56	21,40	23,04

Розв'язок. Табл. 1.6 містить 50 даних, тобто $n = 50$. Для побудови інтервального статистичного ряду знаходимо: кількість інтервалів за формулою (1.2): $k = 1 + 1,4 \ln 50 \approx 6,477 \approx 7$; $x_{\max} = 38,84$, $x_{\min} = 6,11$; довжина кожного інтервалу за формулою (1.3): $\Delta = \dfrac{x_{\max} - x_{\min}}{k} = \dfrac{38,84 - 6,11}{7} \approx 4,68$. Отже, за початок першого інтервалу обираємо $a_1 = x_{\min} = 6,11$. Тоді $a_2 = a_1 + \Delta = 6,11 + 4,68 = 10,79$. Аналогічно, $a_3 = 15,47$; $a_4 = 20,15$; $a_5 = 24,83$; $a_6 = 29,51$; $a_7 = 34,19$; $a_8 = 38,87$.

Підраховуючи кількість варіант, що попали в кожний інтервал, отримаємо інтервальний статистичний ряд (табл. 1.7).

Таблиця 1.7

$[a_i; a_{i+1})$	[6,11; 10,79)	[10,79; 15,47)	[15,47; 20,15)	[20,15; 24,83)	[24,83; 29,51)	[29,51; 34,19)	[34,19; 38,87)
n_i	1	5	6	12	15	8	3
n_i/Δ	0,214	1,068	1,282	2,564	3,205	1,709	0,641

За даними табл. 1.7 будуємо гістограму (рис. 1.3).

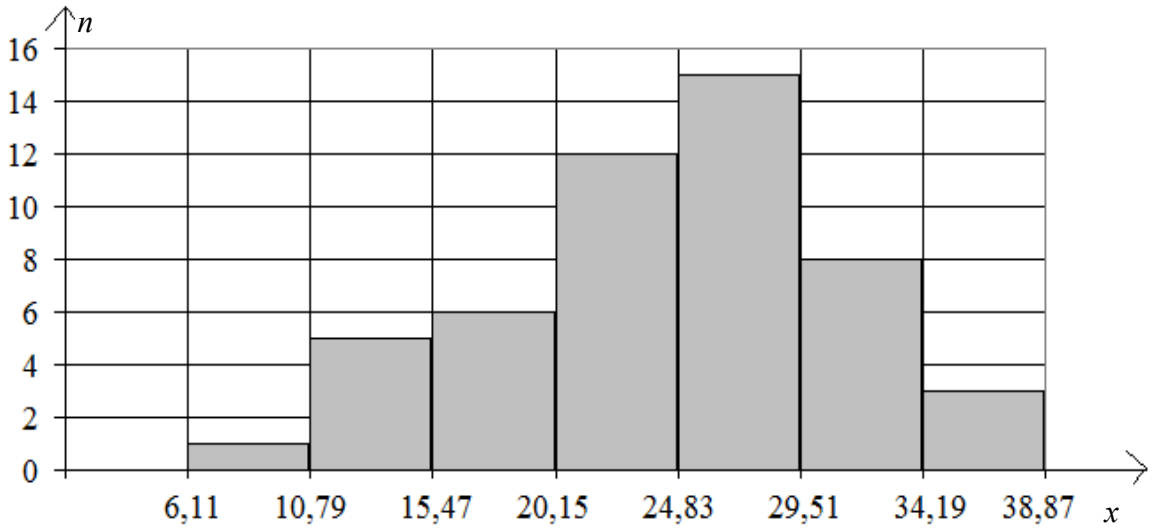

Рисунок 1.3. Гістограма

Для побудови полігона частот і полігона відносних частот обчислимо середини інтервалів за формулою: $c_i = \frac{a_i + a_{i+1}}{2}$. Отримаємо: $c_1 = \frac{6,11 + 10,79}{2} = 8,45$; $c_2 = 13,13$; $c_3 = 17,81$; $c_4 = 22,49$; $c_5 = 27,17$; $c_6 = 31,85$; $c_7 = 36,51$.

Розрахуємо відносні частоти за формулою (1.1): $w_1 = \frac{n_1}{n} = \frac{1}{50} = 0,02$; $w_2 = 0,1$; $w_3 = 0,12$; $w_4 = 0,24$; $w_5 = 0,3$; $w_6 = 0,16$; $w_7 = 0,06$.

Результати оформимо у вигляді таблиці (табл. 1.8), останній стовпчик якої будемо використовувати для перевірки правильності розрахунків.

Таблиця 1.8

c_i	8,45	13,13	17,81	22,49	27,17	31,85	36,51	Перевірка
n_i	1	5	6	12	15	8	3	$\sum_{i=1}^{k} n_i = 50$
w_i	0,02	0,1	0,12	0,24	0,3	0,16	0,06	$\sum_{i=1}^{k} w_i = 1$

За даними табл. 1.8 будуємо полігон частот (рис. 1.4) і полігон відносних частот (рис. 1.5).

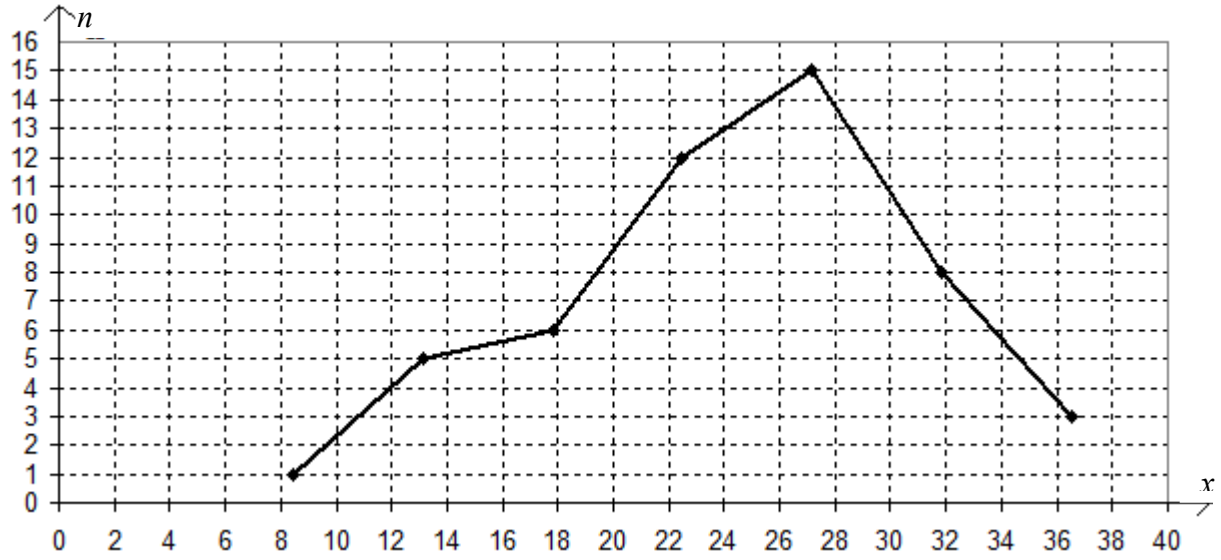

Рисунок 1.4. Полігон частот

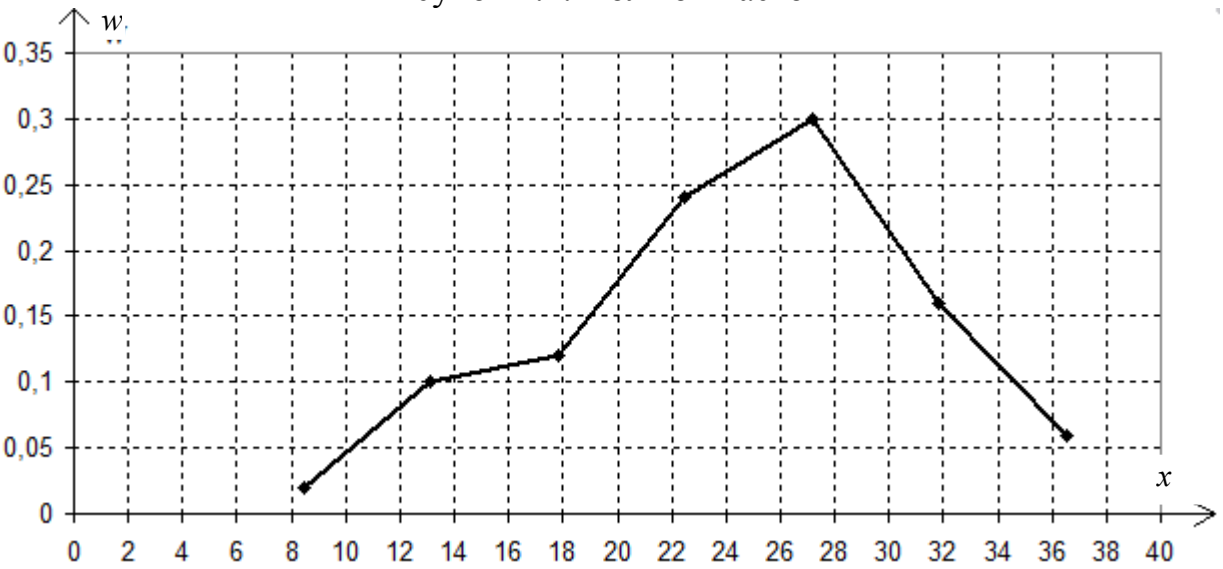

Рисунок 1.5. Полігон відносних частот

Для знаходження емпіричної функції розподілу обчислимо за формулою (1.4) суму відносних частот варіант, менших за x (тобто $F_n(x) = \sum_{x_i<x} w_i$). За x оберемо ліву границю кожного інтервала. Отримаємо:

якщо $x = a_1 = 6{,}11$, то $F_n(x) = \sum_{x_i<6{,}11} w_i = 0$, оскільки таких значень X немає;

якщо $x = a_2 = 10{,}79$, то $F_n(x) = \sum_{x_i<10{,}79} w_i = w_1 = 0{,}02$;

якщо $x = a_3 = 15{,}47$, то $F_n(x) = \sum_{x_i<15{,}47} w_i = w_1 + w_2 = 0{,}02 + 0{,}1 = 0{,}12$;

якщо $x = a_4 = 20{,}15$, то $F_n(x) = \sum_{x_i<20{,}15} w_i = w_1 + w_2 + w_3 = 0{,}02 + 0{,}1 + 0{,}12 = 0{,}24$;

якщо $x = a_5 = 24{,}83$, то $F_n(x) = \sum_{x_i<24{,}83} w_i = w_1 + w_2 + w_3 + w_4 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 = 0{,}48$;

якщо $x = a_6 = 29{,}51$, то $F_n(x) = \sum_{x_i<29{,}51} w_i = w_1 + w_2 + w_3 + w_4 + w_5 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 = 0{,}78;$

якщо $x = a_7 = 34{,}19$, то $F_n(x) = \sum_{x_i<34{,}19} w_i = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 + 0{,}16 = 0{,}94;$

якщо $x = a_8 = 38{,}87$, то $F_n(x) = \sum_{x_i<38{,}87} w_i = w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7 =$

$= 0{,}02 + 0{,}1 + 0{,}12 + 0{,}24 + 0{,}3 + 0{,}16 + 0{,}06 = 1.$

Останнє значення функції розподілу означає, що всі значення X менші за 38,87.

Емпіричну щільність розподілу обчислимо за формулою (1.5): $f_n(x) = \frac{w_i}{\Delta}$. Результати обчислень надано у табл. 1.9.

Таблиця 1.9

a_i	6,11	10,79	15,47	20,15	24,83	29,51	34,19	38,87
w_i	0,02	0,1	0,12	0,24	0,3	0,16	0,06	$\sum_{i=1}^{k} w_i = 1$
F_i	0	0,02	0,12	0,24	0,48	0,78	0,94	1
$[a_i; a_{i+1})$	[6,11; 10,79)	[10,79; 15,47)	[15,47; 20,15)	[20,15; 24,83)	[24,83; 29,51)	[29,51; 34,19)	[34,19; 38,87)	
f_i	0,0043	0,0214	0,0256	0,0513	0,0641	0,0342	0,0128	

За даними табл. 1.9 будуємо кумуляту (рис. 1.6) та графік емпіричної щільності розподілу (рис. 1.7).

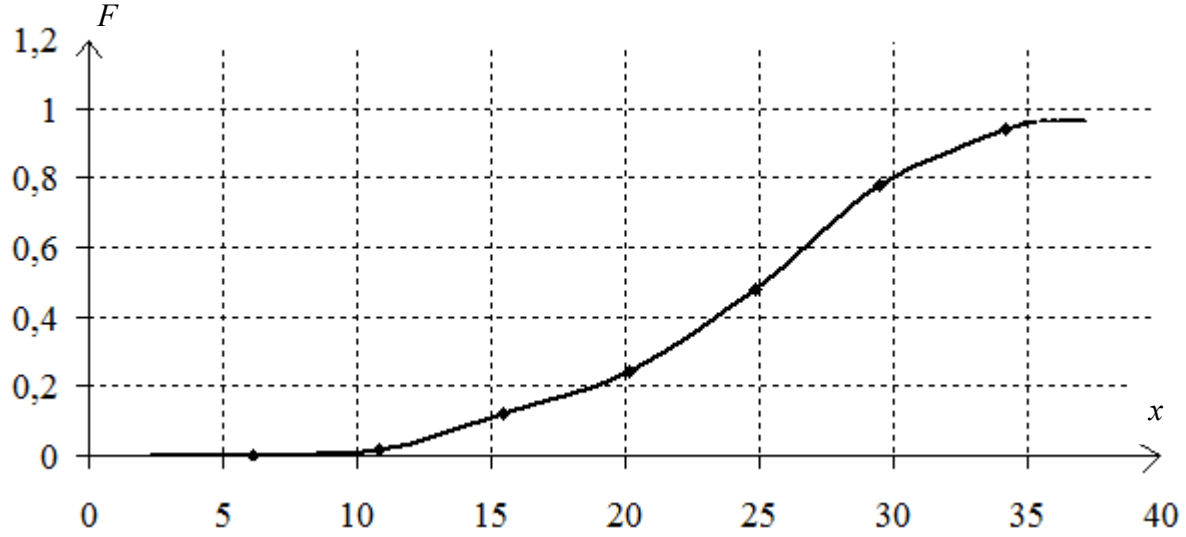

Рисунок 1.6. Графік кумуляти розподілу

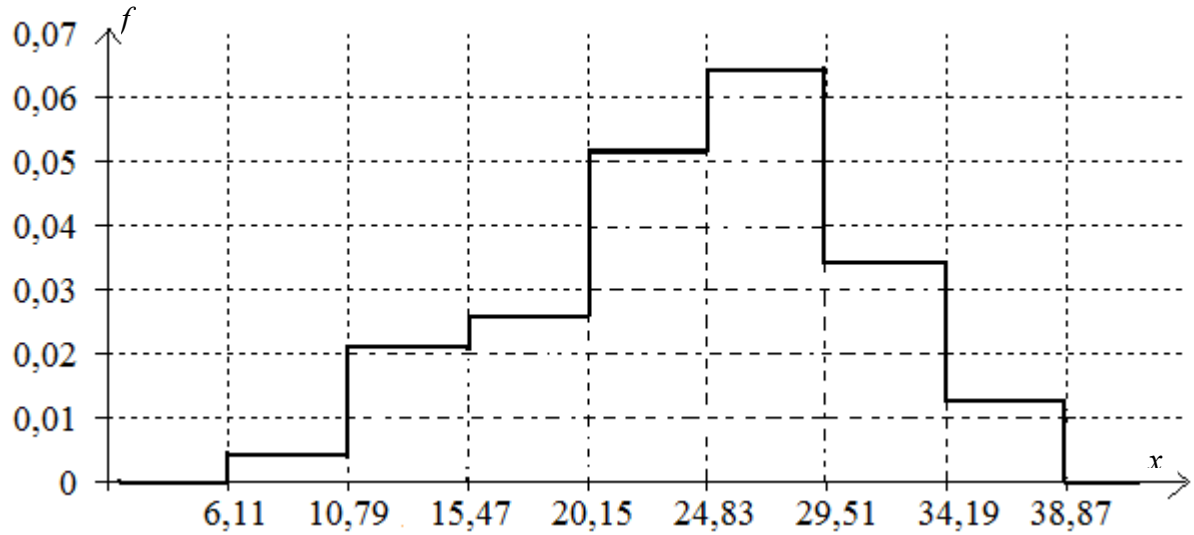

Рисунок 1.7. Графік емпіричної щільності розподілу

1.4. Числові характеристики статистичних рядів

1.4.1. Поняття про оцінки параметрів

Деяку ознаку X генеральної сукупності можна розглядати як випадкову величину. Тоді вибірка значень X – це емпіричний закон розподілу випадкової величини. Для дискретних і неперервних випадкових величин визначені числові характеристики, основними з яких є математичне сподівання, дисперсія і середнє квадратичне відхилення. Числові характеристики випадкових величин часто є параметрами їх розподілів. Аналогічно, числові характеристики визначені і для статистичних рядів, це – вибіркове середнє, вибіркове середнє геометричне, вибіркова дисперсія, вибіркове середнє квадратичне відхилення і т. ін.

У прикладних задачах за даними вибірки часто необхідно визначити закон розподілу випадкової величини, що є одним із основних завдань математичної статистики. При цьому вибіркове середнє вважається **оцінкою** (аналогом) математичного сподівання, вибіркова дисперсія – оцінкою дисперсії, вибіркове середнє квадратичне відхилення – оцінкою середнього квадратичного відхилення. При цьому виникає питання: наскільки правомірні такі оцінки?

Оцінки параметрів повинні відповідати таким вимогам.

Незсуненість. Це означає, що при проведенні великої кількості спостережень (вимірювань) з вибірками одного об’єму оцінка параметру, отримана з кожної вибірки, прямує до істинного значення цього параметру генеральної сукупності.

Спроможність. Зі збільшенням об’єму вибірки оцінка прямує до значення відповідного параметру генеральної сукупності з ймовірністю, що дорівнює 1.

Достатність. Оцінка містить всю необхідну інформацію.

Ефективність. Оцінки, отримані за вибірками однакового об’єму, мають

мінімальну дисперсію.

Зауваження. При використанні оцінок необхідно пам'ятати, що вони отримуються тільки за певних умов і, відповідно, дійсні тільки при виконанні цих умов.

Для оцінювання параметрів розподілу за даними вибірки використовується метод максимальної правдоподібності. Але він застосовується тільки тоді, коли відомий закон розподілу.

1.4.2. Числові характеристики рядів розподілу

Основною числовою характеристикою статистичного ряду є середнє арифметичне або вибіркове середнє.

Вибірковим середнім називається величина $\bar{x}$, яка обчислюється за формулою:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i w_i . \qquad (1.6)$$

У разі інтервального статистичного ряду за x_i вибирається середина *i*-го інтервала.

Якщо вибірка містить незгруповані дані, то вибіркове середнє розраховується за формулою:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (1.7)$$

Зауваження. Оскільки статистичний ряд є емпіричним законом розподілу величини *X*, то вибіркове середнє, зазвичай, вважається аналогом або оцінкою математичного сподівання випадкової величини *X*. Хоча це твердження вірне тільки для нормального закону розподілу.

Вибірковим середнім геометричним називається величина $\overline{x_G}$, яка обчислюється за формулою:

$$\overline{x_G} = \sqrt[n]{\prod_{i=1}^{n} x_i} . \qquad (1.9)$$

Середнє геометричне застосовується як центральна тенденція тоді, коли значення *X* змінюються з постійним співвідношенням між попереднім і наступним значеннями, тобто якщо $\frac{x_i}{x_{i-1}} = \frac{x_{i+1}}{x_i}$ (наприклад, збільшення капіталовкладень, експлуатаційні витрати і т. ін.).

Модою *Mo* називається значення величини *X*, яке має у вибірці найбільшу частоту. У випадку інтервального статистичного ряду мода розраховується за формулою:

$$Mo = x_{Mo} + \frac{\Delta\left(n_{Mo} - n_{Mo-1}\right)}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}}, \qquad (1.10)$$

де x_{Mo} – початок інтервала, якому відповідає найбільша частота (такий інтервал називається модальним);

Δ − величина інтервала;

n_{Mo} − частота у модальному інтервалі;

n_{Mo-1} , n_{Mo+1} − частоти в попередньому і наступному інтервалах відповідно.

Зауваження. Мода не застосовується тоді, коли гістограма або полігон частот показують наявність двох або більше вершин („піків").

Медіаною *Me* називається значення величини *X*, що розділяє вибірку, елементи якої розташовані у порядку зростання, на дві рівні за об'ємом частини.

Якщо це вибірка значень дискретної випадкової величини, то медіаною є те її значення, яке розташоване всередині. Якщо кількість членів ряду непарна, то це елемент з номером $\frac{n+1}{2}$. Якщо кількість елементів вибірки парна, то медіана дорівнює середньому арифметичному її членів з номерами $\frac{n}{2}$ та $\frac{n}{2}+1$.

Якщо розглядається вибірка неперервної випадкової величини, то медіана розраховується за формулою:

$$Me = X_{Me} + \frac{\Delta\left(\frac{n}{2} - n_x^{\max}\right)}{n_{Me}}, \qquad (1.11)$$

де X_{Me} – фактична нижня границя медіанного інтервала;

Δ − величина інтервала;

$n_x^{\max}$ – сума частот, що накопичена до початку медіанного інтервала;

n_{Me} – частота в медіанному інтервалі.

Зауваження. На значення медіани не впливають змінення значень крайніх елементів впорядкованої вибірки, тому її часто застосовують як центральну тенденцію тоді, коли крайні елементи вибірки значно відрізняються від інших її елементів.

Іноді виникає необхідність у більш дрібному поділі статистичного ряду. Тому, крім медіани виділяють **квартилі *Q₁*, *Q₂*=*Me*, *Q₃*** (1/4 ряду), **квінтилі *q₁*,…, *q₄*** (1/5 ряду), **децилі *d₁*, *d₂*,…, *d₉*** (1/10 ряду).

В інтервальному варіаційному ряді квартилі, всередині визначеного по накопичених частотах інтервала, розраховуються за формулами:

$$\text{нижній квартиль} - \ Q_1 = X_{Q_1} + \frac{\Delta\left(\frac{1}{4}n - n_{x_1}^{\max}\right)}{n_{Q_1}},$$

$$\text{верхній квартиль} - \ Q_3 = X_{Q_3} + \frac{\Delta\left(\frac{3}{4}n - n_{x_3}^{\max}\right)}{n_{Q_3}}, \qquad (1.11а)$$

де X_Q – фактична нижня границя квартильних інтервалів;

Δ − величина інтервала;

n − об'єм вибірки;

$n_{x_1}^{\max}$, $n_{x_3}^{\max}$ – суми частот, накопичені до початку відповідних квартильних інтервалів;

n_{Q_1}, n_{Q_3} – частоти в квартильних інтервалах.

Аналогічно, квінтилі та децилі інтервального ряду обчислюються за формулами:

$$\text{перший квінтиль} - \; q_1 = X_{q_1} + \frac{\Delta\left(\frac{1}{5}n - n_{x_1}^{\max}\right)}{n_{q_1}},$$

$$\text{другий квінтиль} - \; q_2 = X_{q_2} + \frac{\Delta\left(\frac{2}{5}n - n_{x_2}^{\max}\right)}{n_{q_2}},\ldots \qquad (1.11б)$$

$$\text{перший дециль} - \; d_1 = X_{d_1} + \frac{\Delta\left(\frac{1}{10}n - n_{x_1}^{\max}\right)}{n_{d_1}},$$

$$\text{другий дециль} - \; d_2 = X_{d_2} + \frac{\Delta\left(\frac{2}{10}n - n_{x_2}^{\max}\right)}{n_{d_2}},\ldots \qquad (1.11в)$$

1.4.3. Числові характеристики розсіювання

Варіаційним розмахом R називається різниця між максимальним і мінімальним елементом вибірки:

$$R = x_{\max} - x_{\min}. \qquad (1.12)$$

Вибірковою дисперсією S^2 називається середнє арифметичне квадратів відхилень варіант від їх вибіркової середньої:

$$S^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \sum_{i=1}^{k}(x_i - \bar{x})^2 w_i \qquad (1.13)$$

$$\text{або } S^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i. \qquad (1.14)$$

Дисперсія є показником розсіювання елементів вибірки відносно їх середнього значення. Вибіркова дисперсія, отримана за формулою (1.14), називається незсуненою оцінкою дисперсії генеральної сукупності.

Різниця дисперсій, отриманих за формулами (1.13) та (1.14) зазвичай невелика, однак може вплинути на точність оцінок. Тому, якщо відомо точне значення математичного сподівання, використовують формулу (1.13), в іншому випадку – формулу (1.14).

Якщо дані незгруповані, то дисперсію можна розрахувати за формулою:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (1.15)$$

В інтервальних рядах розподілу при розрахунку дисперсії вибирають значення, що знаходиться всередині інтервалу. При цьому допускається

помилка, яка збільшується із збільшенням інтервала та при малій кількості інтервалів. Враховуючи похибку, яка називається поправкою Шеппарда, дисперсія розраховується за формулою:

$$S^2_{\text{Шеппарда}} = S^2 - \frac{1}{12}\Delta, \tag{1.16}$$

де Δ − величина інтервала. Поправка Шеппарда використовується при об'ємі вибірки $n > 500$.

Вибірковим середнім квадратичним відхиленням S називається величина, що дорівнює кореню квадратному з вибіркової дисперсії:

$$S = \sqrt{S^2}. \tag{1.17}$$

Вибіркове середнє квадратичне відхилення теж є показником розсіювання елементів вибірки відносно їх середнього значення, але, на відміну від дисперсії, воно має ті ж одиниці вимірювання, що й елементи вибірки.

Коефіцієнтом варіації v називається величина, що дорівнює процентному відношенню вибіркового середнього квадратичного відхилення до модуля вибіркового середнього:

$$v = \frac{S}{|\bar{x}|} \cdot 100\% \ (\bar{x} \neq 0). \tag{1.18}$$

Якщо коефіцієнт варіації більший за 100%, то елементи вибірки неоднорідні і вона не може бути використана у подальших дослідженнях.

1.4.4. Характеристики форми ряду розподілу та його вершини

Початковим емпіричним моментом m_k^* називається величина, що обчислюється за формулою:

$$m_k^* = \frac{\sum_i x_i^k n_i}{n}. \tag{1.19}$$

Початковий емпіричний момент першого порядку $m_1^* = \frac{\sum_i x_i n_i}{n} = \bar{x}$ є вибірковим середнім.

Центральним емпіричним моментом m_k називається величина, що обчислюється за формулою:

$$m_k = \frac{\sum_i (x_i - \bar{x})^k n_i}{n}. \tag{1.20}$$

Центральний емпіричний момент першого порядку дорівнює нулю, а центральний момент другого порядку $m_2 = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n} = S^2$ − вибірковій дисперсії.

Умовним емпіричним моментом M_k називається величина, що обчислюється за формулою:

$$M_k = \frac{\sum_i \left(\frac{x_i - C}{h} \right)^k n_i}{n}, \tag{1.21}$$

де C − умовний нуль (варіанта, яка знаходиться в середині варіаційного ряду або має найбільшу частоту), h − крок (відстань між двома сусідніми варіантами).

За допомогою умовного моменту І-го порядку можна обчислити вибіркове середнє:

$$\overline{x} = M_1 h + C. \tag{1.22}$$

Центральні моменти довільного порядку виражаються через умовні моменти за допомогою формул:

$$m_0 = 1,$$

$$m_1 = 0,$$

$$m_2 = \left(M_2 - \left(M_1 \right)^2 \right) h^2 = S^2,$$

$$m_3 = \left(M_3 - 3M_2 M_1 + 2\left(M_1 \right)^3 \right) h^3,$$

$$m_4 = \left(M_4 - 4M_3 M_1 + 6M_2 \left(M_1 \right)^2 - 3\left(M_1 \right)^4 \right) h^4.$$

Симетричним називається ряд розподілу, в якому частоти варіант, рівновіддалених від центра розподілу, рівні між собою. У симетричних рядах вибіркове середнє, медіана і мода співпадають.

Асиметрією називається витягнутість однієї із віток графіка ряду розподілу. Асиметрія вважається помірною, якщо виконується умова:

$$\overline{x} - Mo \approx 3\left(\overline{x} - Me \right). \tag{1.23}$$

Коефіцієнт асиметрії As обчислюється за формулою:

$$As = \frac{m_3}{S^3}. \tag{1.24}$$

Додатне значення коефіцієнта асиметрії свідчить про наявність правосторонньої асиметрії, від'ємне − лівосторонньої.

Для оцінки «пологості» («крутості», «гостровершинності») графіка статистичного ряду використовують показник **ексцесу** e_x, який обчислюється за формулою:

$$e_x = \frac{m_4}{S^4} - 3. \tag{1.25}$$

Значення показника ексцесу для нормального закону розподілу дорівнює 3. Додатне значення показника ексцесу свідчить про вищу і гострішу вершину графіка ряду розподілу, ніж у нормальному законі розподілу. При від'ємному значенні − вершина нижча і полога.

Приклад 1.3. За даними вибіркового дослідження відомі ціни x_i певного товару у різних торгівельних організаціях (табл. 1.10). Знайти всі можливі числові характеристики за даними таблиці.

Таблиця 1.10

Організація	1	2	3	4	5	6	7	8
Ціна	100	110	115	125	140	145	145	150

Розв'язок. За незгрупованими даними табл. 1.10 можна знайти: вибіркове середнє за формулою (1.7), медіану, розмах варіації за формулою (1.12), дисперсію за формулою (1.15), вибіркове середнє квадратичне відхилення за формулою (1.17), коефіцієнт варіації за формулою (1.18). Кількість елементів вибірки $n = 8$. Вибіркове середнє

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{100+110+115+125+140+145+145+150}{8} = 128,75.$$

Кількість елементів вибірки парна, тому медіана дорівнює середньому арифметичному її членів з номерами $\frac{n}{2}$ та $\frac{n}{2}+1$: $x_{n/2} = x_4 = 125$; $x_{n/2+1} = x_5 = 140$;

$Me = \frac{x_4 + x_5}{2} = \frac{125+140}{2} = 132,5$; $Q_1 = \frac{x_2 + x_3}{2} = 112,5$, $Q_3 = 145$.

Розмах варіації $R = x_{\max} - x_{\min} = 150 - 100 = 50$.

Для розрахунку вибіркової дисперсії складемо таблицю 1.11.

Таблиця 1.11

x_i	100	110	115	125	140	145	145	150
$x_i - \bar{x}$	–28,75	–18,75	–13,75	–3,75	11,25	16,25	16,25	21,25
$\left(x_i - \bar{x}\right)^2$	826,563	351,56	189,06	14,063	126,56	264,06	264,06	451,56

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{826,563+351,56+189,06+14,063+126,56+264,06+264,06}{8} +$$

$$+\frac{451,56}{8} = 310,938.$$

Вибіркове середнє квадратичне відхилення $S = \sqrt{S^2} = \sqrt{310,938} \approx 17,63$.

Коефіцієнт варіації $v = \frac{S}{|\bar{x}|} \cdot 100\% = \frac{17,63}{128,75} \cdot 100\% \approx 13,69\%$.

Приклад 1.4. За даними вибіркового дослідження відомі ціни x_i певного товару у різних торгівельних організаціях (табл. 1.12). Визначити центральну тенденцію за даними таблиці.

Таблиця 1.12

Організація	1	2	3	4	5	6	7	8	9
Ціна	100	110	115	125	140	145	145	150	450

Розв'язок. Оскільки дані табл. 1.12 незгруповані, то оцінкою центральної тенденції може слугувати вибіркове середнє або медіана.

Знайдемо вибіркове середнє за формулою (1.7):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{100+110+115+125+140+145+145+150+450}{9} \approx 164,4 .$$

Кількість елементів вибірки непарна, тому медіана дорівнює її члену з номером $\frac{n+1}{2}$: $Me = x_{n+1/2} = x_5 = 140$.

Зобразимо дані таблиці графічно, вкажемо на діаграмі положення середнього і медіани (рис. 1.8).

На рис. 1.8 видно, що як центральну тенденцію вибірки слід взяти медіану.

Зауваження. Цей приклад показує, що у випадках наявності у вибірці даних, які сильно відрізняються один від одного, або даних, які сильно відрізняються від всіх останніх (так званих викидів), медіана є більш усталеною оцінкою центральної тенденції, ніж вибіркове середнє.

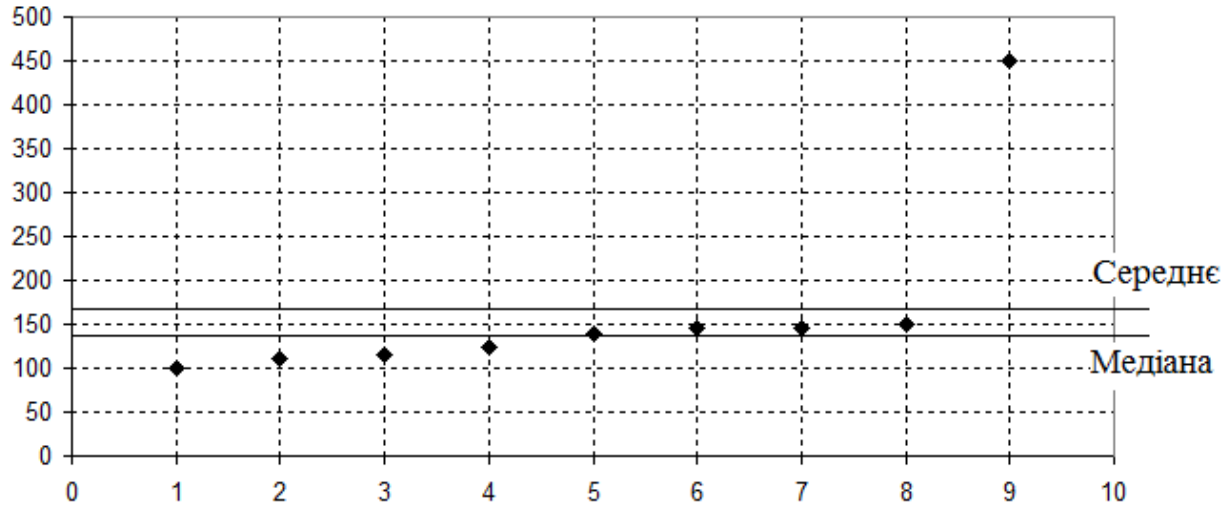

Рисунок 1.8. Положення середнього і медіани відносно вибіркових даних

Приклад 1.5. За даними вибіркового дослідження відома кількість людей, що відвідували лікарню протягом року. Дані згруповані залежно від віку відвідувачів (табл. 1.13). Знайти всі можливі числові характеристики за даними таблиці.

Таблиця 1.13

Вік	20-29	30-39	40-49	50-59	60-69
Кількість відвідувань	45	36	175	361	825

Розв'язок. Позначимо X – вік відвідувачів лікарні, n – загальна кількість відвідувань, n = 1442; n_i – кількість відвідувань залежно від віку; k – кількість досліджуваних вікових груп, k = 5. Тоді відповідно даним табл. 1.11 отримаємо інтервальний статистичний ряд (табл. 1.14).

Таблиця 1.14

$[a_i; a_{i+1}]$	20-29	30-39	40-49	50-59	60-69
n_i	45	36	175	361	825

Розрахуємо вибіркове середнє за формулою (1.6). Для зручності розрахунки оформимо у вигляді таблиці (табл. 1.15).

Таблиця 1.15

$[a_i; a_{i+1}]$	20-29	30-39	40-49	50-59	60-69	Суми
x_i	24,5	34,5	44,5	54,5	64,5	
n_i	45	36	175	361	825	1442
$x_i n_i$	1102,5	1242	7787,5	19675	53213	83019

Тоді $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{83019}{1442} \approx 57{,}57$.

Моду обчислимо за формулою (1.10), враховуючи, що:

x_{Mo} – початок модального інтервала (якому відповідає найбільша частота), $x_{Mo} = 60$;

n_{Mo} – частота у модальному інтервалі, $n_{Mo} = 825$;

n_{Mo-1} , n_{Mo+1} – частоти в попередньому і наступному інтервалах відповідно, $n_{Mo-1} = 361$, $n_{Mo+1} = 0$ (оскільки модальний інтервал є останнім);

Δ – довжина інтервала, $\Delta = 69-59 = 10$.

Отже, $Mo = x_{Mo} + \frac{\Delta\left(n_{Mo} - n_{Mo-1}\right)}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} = 60 + \frac{10(825-361)}{2\cdot 825 - 365 - 0} = 65$.

Медіану обчислимо за формулою (1.11), враховуючи, що:

X_{Me} – фактична нижня границя медіанного інтервала, $X_{Me} = 60$ (оскільки всього даних 1442, то їх половина $1442/2 = 721$; у статистичному ряді до початку останнього інтервала міститься 617 даних, тому медіана повинна знаходитися в останньому інтервалі);

$n_x^{\max}$ – сума частот, що накопичена до початку медіанного інтервала, $n_x^{\max} = 617$; n_{Me} – частота в медіанному інтервалі, $n_{Me} = 825$.

Отже, $Me = X_{Me} + \frac{\Delta\left(\frac{n}{2} - n_x^{\max}\right)}{n_m} = 60 + \frac{10\left(\frac{1442}{2} - 617\right)}{825} = 61{,}26$.

Верхній та нижній квартилі обчислимо за формулами (1.11а), враховуючи, що: X_{Q_1} – фактична нижня границя нижнього квартильного інтервалу, $X_{Q_1} = 50$ ($n = 1442$, $n/4 = 1442/4 = 360{,}5$; сума частот перших трьох інтервалів 256, а перших чотирьох − 617, тому нижній квартиль у четвертому інтервалі);

Δ − величина інтервала, $\Delta = 10$;

$n_{x_1}^{\max}$ – сума частот, накопичена до початку нижнього квартильного інтервала, $n_{x_1}^{\max} = 256$;

n_{Q_1} – частота квартильного інтервала, $n_{Q_1} = 361$.

Отже, $Q_1 = X_{Q_1} + \dfrac{\Delta\left(\frac{1}{4}n - n_{x_1}^{\max}\right)}{n_{Q_1}} = 50 + \dfrac{10\left(\frac{1442}{4} - 256\right)}{361} = 52{,}89$.

Аналогічно, верхній квартиль:

$$Q_3 = X_{Q_3} + \frac{\Delta\left(\frac{3}{4}n - n_{x_3}^{\max}\right)}{n_{Q_3}} = 60 + \frac{10(1081{,}5 - 617)}{825} = 65{,}63.$$

Перший квінтиль $q_1 = X_{q_1} + \dfrac{\Delta\left(\frac{1}{5}n - n_{x_1}^{\max}\right)}{n_{q_1}} = 50 + \dfrac{10\left(\frac{1442}{5} - 256\right)}{361} = 50{,}9$,

другий квінтиль $q_2 = X_{q_2} + \dfrac{\Delta\left(\frac{2}{5}n - n_{x_2}^{\max}\right)}{n_{q_2}} = 50 + \dfrac{10(576{,}8 - 256)}{361} = 58{,}9, \ldots$

Перший дециль $d_1 = X_{d_1} + \dfrac{\Delta\left(\frac{1}{10}n - n_{x_1}^{\max}\right)}{n_{d_1}} = 40 + \dfrac{10\left(\frac{1442}{10} - 81\right)}{175} = 43{,}6, \ldots$

Дисперсію обчислимо за формулами (1.14). Для зручності розрахунки оформимо у вигляді таблиці (табл. 1.16).

Таблиця 1.16

x_i	24,5	34,5	44,5	54,5	64,5	Суми
n_i	45	36	175	361	825	1442
$\left(x_i - \bar{x}\right)^2$	1093,6249	1190,25	1980,25	2970,25	4160,25	11394,62

Тоді $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \dfrac{11394{,}62}{1442 - 1} = 7{,}9$.

Вибіркове середнє квадратичне відхилення за формулою (1.17) дорівнює: $S = \sqrt{S^2} = 2{,}81$.

Коефіцієнт варіації за формулою (1.18) дорівнює: $v = \dfrac{S}{|\bar{x}|} \cdot 100\% = \dfrac{2{,}81}{57{,}57} \cdot 100\% \approx 4{,}89\%$.

Приклад 1.6. Є дані вибіркового дослідження про витрати часу (секунди) на обслуговування клієнтів у деякому банку (табл. 1.17). Знайти числові характеристики за даними таблиці та зробити висновок про можливі вид та форму ряду розподілу.

Таблиця 1.17

Витрати часу	180	300	420	540	660	780	900	1020	Сума
Кількість клієнтів	4	13	44	20	9	5	3	2	100

Розв'язок. Позначимо через X – витрати часу на обслуговування одного клієнта, n – загальну кількість клієнтів, $n = 100$; n_i – кількість клієнтів залежно від витраченого часу.

За даними табл. 1.17 можна знайти: моду, медіану, для знаходження інших числових характеристик використаємо метод моментів і формули 1.21, 1.22, 1.24, 1.25.

$Mo = x_3 = 420$.

Об'єм вибірки – 100, $\frac{n}{2} = 50$, $n_1 + n_2 = 4 + 13 = 17 < 50$, $n_1 + n_2 + n_3 = 4 + 13 + 44 = 61 > 50$, отже $Me = x_3 = 420$.

Аналізуючи значення x_i, вибираємо умовний нуль $C = 420$ та крок $h = 120$. Для зручності подальші розрахунки оформимо у вигляді таблиці (табл. 1.18).

Таблиця 1.18

x_i	180	300	420	540	660	780	900	1020	Сума
$\frac{x_i - C}{h}$	-2	-1	0	1	2	3	4	5	
n_i	4	13	44	20	9	5	3	2	100
$\left(\frac{x_i - C}{h}\right) n_i$	-8	-13	0	20	18	15	12	10	54
$\left(\frac{x_i - C}{h}\right)^2 n_i$	16	13	0	20	36	45	48	50	228
$\left(\frac{x_i - C}{h}\right)^3 n_i$	-32	-13	0	20	72	135	192	250	624
$\left(\frac{x_i - C}{h}\right)^4 n_i$	64	13	0	20	144	405	768	1250	2664

За формулою 1.20 знайдемо умовні емпіричні моменти I-IV порядків:

$$M_1 = \frac{\sum_i \left(\frac{x_i - C}{h}\right) n_i}{n} = \frac{54}{100} = 0,54;\ M_2 = \frac{\sum_i \left(\frac{x_i - C}{h}\right)^2 n_i}{n} = \frac{228}{100} = 2,28;$$

$$M_3 = \frac{\sum_i \left(\frac{x_i - C}{h}\right)^3 n_i}{n} = \frac{624}{100} = 6,24;\ M_4 = \frac{\sum_i \left(\frac{x_i - C}{h}\right)^4 n_i}{n} = \frac{2664}{100} = 26,64.$$

Тоді вибіркове середнє $\overline{x} = M_1 h + C = 0,54 \cdot 120 + 420 = 484,8$. Значення вибіркового середнього не співпадає із модою та медіаною, аналіз даних ряду розподілу свідчить про наявність нормального виду розподілу, отже, варто знайти коефіцієнти асиметрії та ексцесу.

Використаємо формули 1.22 для знаходження центральних емпіричних моментів:

$S^2 = m_2 = \left(M_2 - \left(M_1\right)^2\right)h^2 = \left(2,28 - 0,54^2\right)\cdot 120^2 = 28632,96$;

$m_3 = \left(M_3 - 3M_2M_1 + 2\left(M_1\right)^3\right)h^3 = \left(6,24 - 3\cdot 2,28\cdot 0,54 + 2\cdot\left(0,54\right)^3\right)\cdot 120^3 = 4944374,78$;

$m_4 = \left(M_4 - 4M_3M_1 + 6M_2\left(M_1\right)^2 - 3\left(M_1\right)^4\right)h^4 = (26,64 - 4\cdot 6,24\cdot 0,54 + 6\cdot 2,28\cdot\left(0,54\right)^2 -$

$-3\cdot\left(0,54\right)^4)\cdot 120^4 = 3503470853$.

Так як дисперсія: $S^2 = 28632,96$, то вибіркове середнє квадратичне відхилення $S = \sqrt{S^2} = \sqrt{28632,96} \approx 169,213$.

Коефіцієнт варіації $v = \frac{S}{\left|\overline{x}\right|}\cdot 100\% = \frac{169,213}{484,8}\cdot 100\% \approx 34,9\%$.

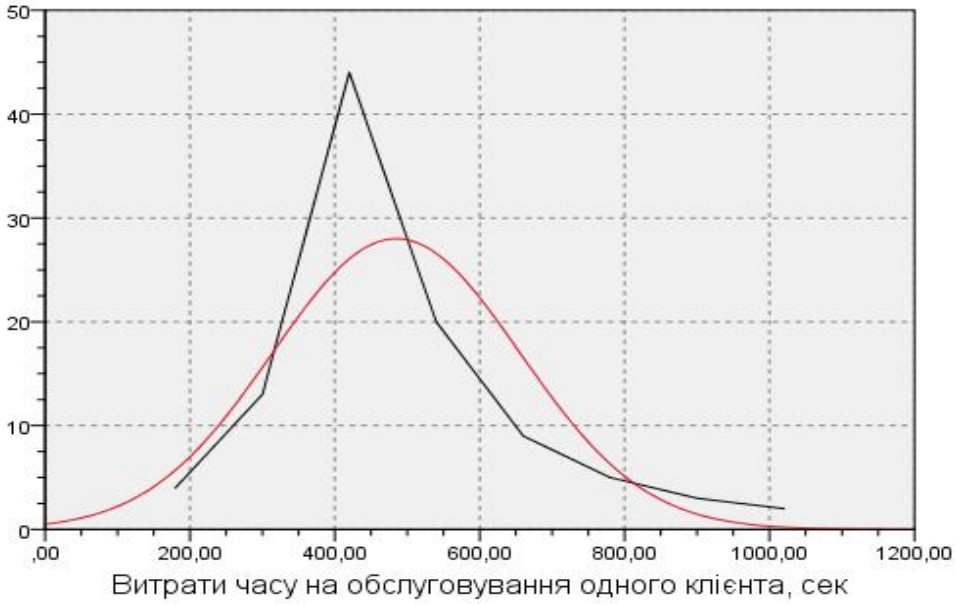

Рисунок 1.9. Форма і вид ряду розподілу відносно нормальної кривої

Значення коефіцієнта асиметрії $As = \frac{m_3}{S^3} = \frac{4944374,78}{169,213^3} \approx 1,02$ свідчить про наявність правосторонньої асиметрії (рис. 1.9).

Показник ексцесу $e_x = \frac{m_4}{S^4} - 3 = \frac{3503470853}{169,213^4} - 3 \approx 4,27$ характеризує більш гостру вершину графіка даних ряду розподілу порівняно із нормальною кривою (рис. 1.9).

1.5. Довірчі інтервали і довірча ймовірність

Однією з основних задач математичної статистики є оцінка числових характеристик (параметрів) генеральної сукупності за вибірковими даними.

Для вибірки можна обчислити такі числові характеристики, як: вибіркове

середнє, мода, медіана, вибіркова дисперсія та вибіркове середнє квадратичне відхилення. Для генеральної сукупності часто визначаються не самі ці параметри, а довірчі інтервали.

Довірчим інтервалом для певного параметру генеральної сукупності називається такий числовий інтервал, в межах якого знаходиться цей параметр. Ймовірність, з якою довірчий інтервал покриє істинне значення параметра, називається **довірчою ймовірністю** або **рівнем надійності** і позначається γ.

Значення довірчої ймовірності обирає дослідник залежно від того, яку ступінь точності розрахунків вимагає дослідження. Зазвичай це значення знаходиться в інтервалі від 0,9 до 0,999. Якщо вимоги точності дуже високі, то для довірчої ймовірності обирається значення 0,999; якщо підвищені – 0,99; звичайні – 0,95; знижені – 0,9.

Довірчі інтервали розраховуються з урахуванням певних вимог до генеральної сукупності. Зазвичай це вимога нормального розподілу її даних.

1.5.1. Довірчий інтервал для генерального середнього при відомій генеральної дисперсії

Нехай X – генеральна сукупність, що підкоряється нормальному закону розподілу; σ^2 – відома генеральна дисперсія; $\{x_1, x_2, ..., x_n\}$ – вибірка з генеральної сукупності об'ємом n; $\overline{x}$ – вибіркове середнє. Потрібно знайти довірчий інтервал для генерального середнього a із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$\overline{x} - z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}} < a < \overline{x} + z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \qquad (1.26)$$

де значення $z_{\frac{1-\gamma}{2}}$ знаходиться з таблиці (табл. 1.19) або за допомогою вбудованої функції Excel НОРМСТОБР(γ). Величина $z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ є шириною довірчого інтервалу.

Таблиця 1.19

Значення $z_{\frac{1-\gamma}{2}}$

γ	0,4	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,001
$z_{\frac{1-\gamma}{2}}$	0,253	0,675	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,090

Приклад 1.7. Автомат, що фасує чай в пачки, працює зі стандартним відхиленням $\sigma = 5$ г. Проведено вибірку об'ємом $n = 30$ пачок. Середня вага пачки чаю у вибірці $\overline{x} = 101$ г. Знайти довірчий інтервал для середньої ваги пачки чаю в генеральній сукупності із рівнем надійності $\gamma = 0{,}95$. Знайти об'єм вибірки, якщо потрібна ширина довірчого інтервалу ± 1 грам.

Розв'язок. Оскільки γ=0,95, то $\frac{1-\gamma}{2}=\frac{1-0,95}{2}=0,025$. З табл. 1.19 знайдемо $z_{\frac{1-\gamma}{2}}=z_{0,025}=1,96$.

Тоді ширина довірчого інтервалу: $z_{\frac{1-\gamma}{2}}\cdot\frac{\sigma}{\sqrt{n}}=1,96\cdot\frac{5}{\sqrt{30}}\approx 1,79$ і за формулою (1.26) маємо довірчий інтервал:

$\bar{x}-1,79<a<\bar{x}+1,79$; $101-1,79<a<101+1,79$; $99,21<a<102,79$.

Отже, середня вага пачки чаю знаходиться в інтервалі від 99,21 до 102,79 грам.

Знайдемо об'єм вибірки, необхідний для того, щоб ширина довірчого інтервалу дорівнювала 1 грам, тобто $z_{\frac{1-\gamma}{2}}\cdot\frac{\sigma}{\sqrt{n}}=1$. Знайдемо n з отриманого рівняння: $1,96\cdot\frac{5}{\sqrt{n}}=1\Rightarrow\sqrt{n}=1,96\cdot 5=9,8\Rightarrow n=9,8^2=96,04$. Отже, мінімальний об'єм вибірки для отримання довірчого інтервалу шириною 1 грам дорівнює 97 пачкам.

Зауваження. У прикладі для розрахунку нового довірчого інтервалу потрібно знаходити вибіркове середнє для вибірки об'ємом 97 пачок, а не середнє арифметичне середніх для вибірок об'єму 30 і 67 пачок.

1.5.2. Довірчий інтервал для генерального середнього при невідомій генеральній дисперсії

Нехай X – генеральна сукупність, що підкоряється нормальному закону розподілу; генеральна дисперсія σ^2 невідома; $\{x_1,x_2,...,x_n\}$ – вибірка з генеральної сукупності об'ємом n; $\bar{x}$ – вибіркове середнє; S – вибіркове середнє квадратичне відхилення. Потрібно знайти довірчий інтервал для генерального середнього a із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$\bar{x}-t_{\frac{1-\gamma}{2},n-1}\cdot\frac{S}{\sqrt{n-1}}<a<\bar{x}+t_{\frac{1-\gamma}{2},n-1}\cdot\frac{S}{\sqrt{n-1}}, \qquad (1.27)$$

де значення $t_{\frac{1-\gamma}{2},n-1}$ знаходиться з таблиці розподілу Стьюдента, яка є у статистичних довідниках, або за допомогою вбудованої функції Excel СТЬЮДРАСПОБР ($\frac{1-\gamma}{2}$, n - 1). Величина $t_{\frac{1-\gamma}{2},n-1}\cdot\frac{S}{\sqrt{n-1}}$ є шириною довірчого інтервалу.

Приклад 1.8. Автомат фасує чай в пачки. Проведено вибірку об'ємом n = 30 пачок. Середня вага пачки чаю у вибірці $\bar{x}$ = 101 г, вибіркове стандартне відхилення S = 4 г. Знайти довірчий інтервал для середньої ваги пачки чаю в

генеральної сукупності із рівнем надійності $\gamma = 0{,}95$. Знайти об'єм вибірки, якщо ширина довірчого інтервалу ± 1 грам.

Розв'язок. Оскільки γ=0,95, то $\frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025$; $n = 30$, тоді $n - 1 = 29$. За допомогою Excel знайдемо $t_{\frac{1-\gamma}{2},n-1} = t_{0{,}025;29}$. Натиснемо f_x у командному рядку, виберемо категорію ***Статистические*** і функцію СТЬЮДРАСПОБР; задамо параметри 0,025 і 29 (див. рис. 1.10, зміст командного рядку). Отримаємо $t_{0{,}025;29} = 2{,}3638$.

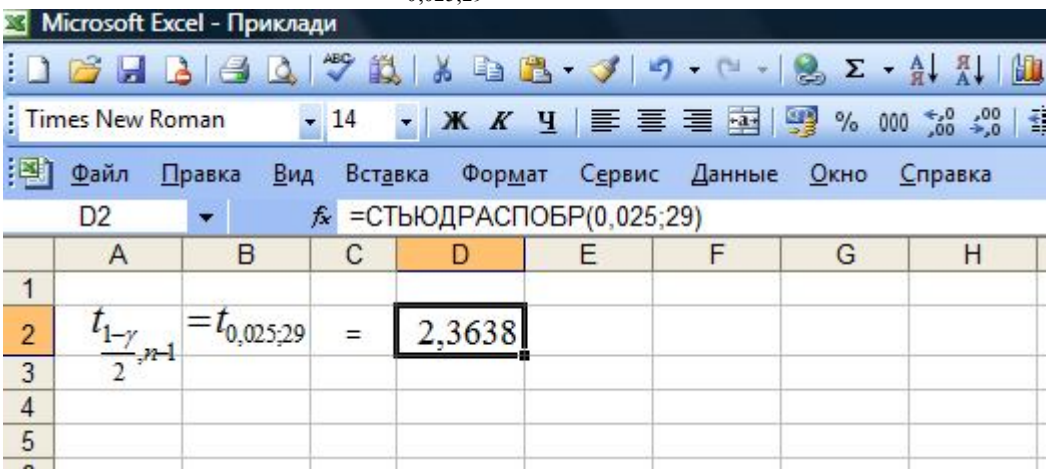

Рис. 1.10. Знаходження $t_{\frac{1-\gamma}{2},n-1}$ за допомогою функцій Excel

Тоді ширина довірчого інтервала $t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}} = 2{,}3638 \cdot \frac{4}{\sqrt{29}} \approx 1{,}75$ і довірчий інтервал за формулою (1.27):

$$\bar{x} - 1{,}75 < a < \bar{x} + 1{,}75; \quad 101 - 1{,}75 < a < 101 + 1{,}75; \quad 99{,}25 < a < 102{,}75.$$

Отже, середня вага пачки чаю знаходиться в інтервалі від 99,25 до 102,75 грам.

Знайдемо об'єм вибірки, необхідний для того, щоб ширина довірчого інтервалу дорівнювала 1 грам, тобто $t_{\frac{1-\gamma}{2},n-1} \cdot \frac{S}{\sqrt{n-1}} = 1$. Знайдемо n з отриманого рівняння:

$$2{,}3638 \cdot \frac{4}{\sqrt{n-1}} = 1 \Rightarrow \sqrt{n-1} = 2{,}3638 \cdot 4 \approx 9{,}46 \Rightarrow$$

$$\Rightarrow n - 1 = 9{,}46^2 = 89{,}49 \Rightarrow n = 90{,}49 \approx 91.$$

Отже, мінімальний об'єм вибірки для отримання довірчого інтервалу шириною 1 грам дорівнює 91 пачці.

1.5.3. Довірчий інтервал для генеральної частки

В прикладних дослідженнях часто потрібно визначити частку об'єктів, що мають певну властивість.

Частка об'єктів генеральної сукупності, що має певну властивість, називається **генеральною часткою**. Частка об'єктів вибірки, що має певну властивість, називається **вибірковою часткою**.

Нехай X – генеральна сукупність, що підкоряється нормальному закону розподілу; $\{x_1, x_2, ..., x_n\}$ – вибірка з генеральної сукупності об'єму n; m – кількість елементів вибірки, що мають задану властивість; $w = \frac{m}{n}$ – вибіркова частка. Потрібно знайти довірчий інтервал для генеральної частки W із заданим рівнем надійності γ.

Шуканий довірчий інтервал знаходиться за формулою:

$$w - z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}} < W < w + z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}}, \qquad (1.28)$$

де значення $z_{\frac{1-\gamma}{2}}$ знаходиться з табл. 1.18 або за допомогою вбудованої функції Excel НОРМСТОБР(γ). Величина $z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}}$ є шириною довірчого інтервалу.

Зауваження. Формула (1.28) використовується тоді, коли $nw \geq 5, \quad n(1-w) \geq 5$.

Приклад 1.9. Проведено вибірку об'ємом $n = 2000$ одиниць продукції. Серед обраних 150 одиниць виявилися бракованими. Знайти довірчий інтервал для генеральної частки бракованих виробів із рівнем надійності $\gamma = 0{,}95$.

Розв'язок. Оскільки $\gamma = 0{,}95$, то $\frac{1-\gamma}{2} = \frac{1-0{,}95}{2} = 0{,}025$. З табл. 1.18 знайдемо $z_{\frac{1-\gamma}{2}} = z_{0{,}025} = 1{,}96$.

Знайдемо вибіркову частку бракованих виробів:

$$m = 150;\ w = \frac{m}{n} = \frac{150}{2000} = 0{,}075.$$

Перевіримо можливість знаходження довірчого інтервалу:

$nw = 2000 \cdot 0{,}075 = 150 \geq 5, \quad n(1-w) = 2000(1-0{,}075) = 2000 \cdot 0{,}925 = 1850 \geq 5$.

Тоді ширина довірчого інтервала:

$$z_{\frac{1-\gamma}{2}} \cdot \sqrt{\frac{w(1-w)}{n}} = 1{,}96\sqrt{\frac{0{,}075(1-0{,}075)}{2000}} \approx 0{,}012,$$

за формулою (1.28) отримаємо довірчий інтервал:

$$w - 0{,}012 < W < w + 0{,}012;$$
$$0{,}075 - 0{,}012 < W < 0{,}075 + 0{,}012;$$
$$0{,}063 < W < 0{,}087.$$

Отже, доля бракованих виробів в генеральній сукупності знаходиться в межах від 0,063 до 0,087, тобто складає від 6,3% до 8,7% від обсягу продукції.

1.6. Визначення числових характеристик і довірчих інтервалів з використанням табличного процесору Microsoft Excel

Більшість числових характеристик у випадку незгрупованих даних можна обчислити з використанням табличного процесора Microsoft Excel. Основні вбудовані функції Excel, що застосовуються для таких розрахунків, представлено у таблиці 1.20. Щоб викликати потрібну функцію, слід нажати кнопку f_x у командному рядку, обрати категорію ***Статистические*** та ім'я функції.

Найчастіше використовуються такі функції:

− **наибольший** (масив даних, *k*) – видає *k*-те найбільше значення в ряді даних;

− **наименьший** (масив даних, *k*) – видає *k*-те найменше значення в ряді даних.

Ширину довірчого інтервалу для генерального середнього можна знайти за допомогою вбудованої статистичної функції Excel ДОВЕРИТ (*альфа, станд_откл, размер*). Параметр *альфа* – це так званий рівень значущості, $\alpha = 1 - \gamma$; параметр *станд_откл* – це вибіркове середнє квадратичне відхилення *S*; параметр *размер* – це об'єм вибірки.

Таблиця 1.20

Статистичні функції Excel

Числові характеристики	**Назва функції**
Середнє	СРЗНАЧ (масив даних)
Середнє геометричне	СРГЕОМ (масив даних)
Мода	МОДА (масив даних)
Медіана	МЕДИАНА (масив даних)
Дисперсія	ДИСП (масив даних)
Середнє квадратичне відхилення	СТАНДОТКЛОН (масив даних)
Мінімальне значення	МИН (масив даних)
Максимальне значення	МАКС (масив даних)
Частота	ЧАСТОТА (масив даних; масив інтервалів)

Приклад 1.10. За даними вибіркового дослідження відома заробітна платня (у грн.) 20-ти службовців певної компанії (табл. 1.21). Знайти за допомогою вбудованих статистичних функцій Excel всі можливі числові характеристики за даними таблиці. Знайти довірчий інтервал для генерального середнього – середньої заробітної платні службовців компанії.

Таблиця 1.21

3560	2190	2390	3400
2180	2400	3350	2340
2900	2570	3300	3150
3680	3250	2250	3240
2180	2600	2870	3050

Розв'язок. Запишемо в лист Excel вхідні дані і числові характеристики, які можна знайти (рис. 1.11). Для знаходження характеристик введемо: в чарунку I3 формулу „=СРЗНАЧ(B2:B21)"; в чарунку I4 формулу „=МЕДИАНА(B2:B21)"; в чарунку I5 формулу „=ДИСП(B2:B21)"; в чарунку I6 формулу „=СТАНДОТКЛОН(B2:B21)"; в чарунку I7 формулу „=МАКС(B2:B21)"; в чарунку I8 формулу „=МИН(B2:B21)".

Для знаходження довірчого інтервалу для генерального середнього знайдемо за допомогою функції ДОВЕРИТ його ширину (див. рис. 1.11, командний рядок). Параметрами візьмемо $\alpha = 1 - \gamma = 1 - 0{,}95 = 0{,}05$; замість другого параметру надамо посилання на чарунку I6, що містить розраховане значення середнього квадратичного відхилення; *размер* – це об'єм вибірки, що дорівнює 20.

Для знаходження початку інтервала запишемо в чарунку I10 формулу „=I3–I9"; для знаходження кінця – формулу „=I3+I9" в чарунку I11.

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

I9 =ДОВЕРИТ(0,05;I6;20)

	A	B	C	D	E	F	G	H	I
1		Вихідні дані		Числові характеристики					
2		3560							
3		2180		Середнє					2842,5
4		2900		Медіана					2885
5		3680		Дисперсія					257914
6		2180		Середнє квадратичне відхилення					507,853
7		2190		Максимальне значення					3680
8		2400		Мінімальне значення					2180
9		2570		Ширина довірчого інтервалу					222,572
10		3250		Початок довірчого інтервалу					2619,93
11		2600		Кінець довірчого інтервалу					3065,07
12		2390							
13		3350							
14		3300							
15		2250							
16		2870							
17		3400							
18		2340							
19		3150							
20		3240							
21		3050							

Рисунок 1.11. Розрахунок числових характеристик

1.7. Побудова гістограми засобами Microsoft Excel

Excel надає два способи побудови гістограми.

Для побудови гістограми першим способом необхідно:

1) Внести в лист Excel вхідні дані і інтервали, за якими ці дані будуть групуватися.

2) Знайти частоти попадання даних в інтервали за допомогою функції ЧАСТОТА, для чого:

− виділити діапазон чарунок (на одну більше, ніж інтервалів), в яких будуть записані частоти;

− викликати f_x – ***Статистические*** – ЧАСТОТА;

− ввести посилання на чарунки, що містять вхідні дані та інтервали;

− натиснути **Ctrl+Shift+Enter**.

3) Викликати ***Вставка – Диаграмма – Гистограмма***, з'явиться діалогове вікно (рис. 1.12).

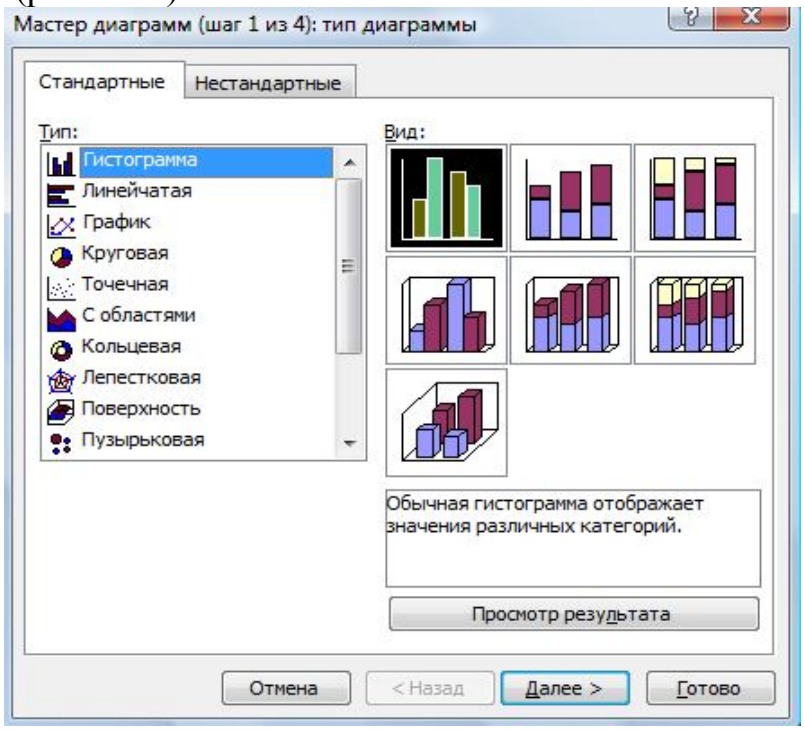

Рисунок 1.12. Діалогове вікно майстра діаграм

4) Надати необхідні для побудови гістограми параметри:

− діапазон вхідних даних, спосіб їх групування (за рядками або стовпцями) та імена рядків даних, якщо це потрібно;

− якщо імена рядків задано, відмітити *Добавить легенду* і вказати її розміщення;

− якщо потрібно, додати *Имена рядов*, або (та) *Имена категорий*, або (та) *Значения*;

− якщо потрібно, заповнити *Заголовок*, *Линии сетки*, *Оси*, *Таблицу данных*.

Для побудови гістограми другим способом необхідно:

1) Внести в лист Excel вхідні дані.

2) Обрати в меню ***Сервис – Анализ данных – Гистограмма***, з'явиться діалогове вікно (рис. 1.13).

3) Задати необхідні для побудови гістограми параметри:

входной интервал – задати посилання на чарунки, в яких знаходяться вхідні дані;

интервал карманов (параметр не є обов'язковим) – задати діапазон чарунок і набор граничних значень у порядку зростання; якщо параметр не введений, то буде автоматично створений набір відрізків, рівномірно розподілених між мінімальним і максимальним значеннями даних;

выходной интервал – ввести посилання на верхню ліву чарунку діапазону, в який буде надано гістограму, або відмітити параметр *Новый рабочий лист* або *Новая рабочая книга*;

интегральный процент – якщо параметр відмічено, то будуть розраховані накопичені частоти і побудований їх графік;

вывод графика – якщо параметр відмічено, то буде створено автоматичну діаграму, при цьому обов'язково задається значення *Новая книга*.

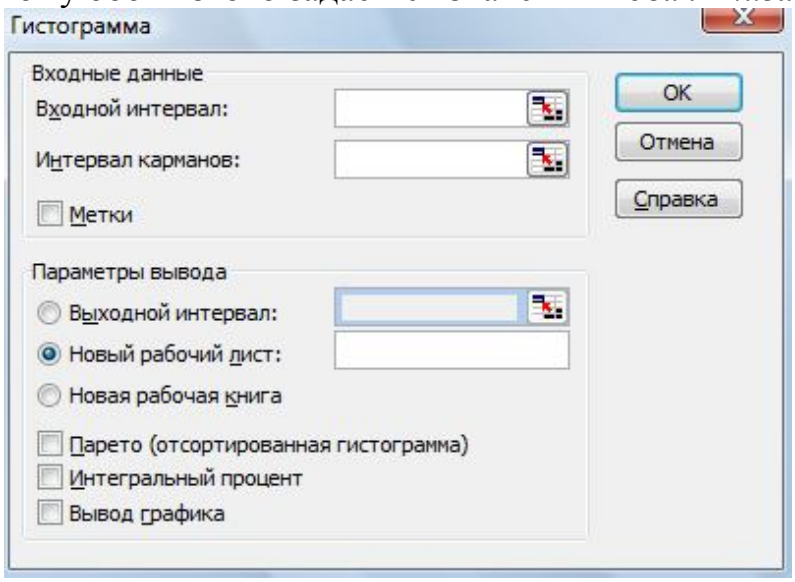

Рисунок 1.13. Діалогове вікно для побудови гістограми

Приклад 1.11. За даними вибіркового дослідження відома кількість родин з дітьми дошкільного віку у селах деякої області (табл. 1.22). Побудувати за допомогою Excel гістограму за даними таблиці.

Таблиця 1.22

Кількість родин з дітьми дошкільного віку у селах деякої області

27	36	34	46	43	28	29	37	40	43
40	33	50	37	41	32	27	43	34	32
30	41	54	42	47	35	49	49	54	36
36	51	36	24	35	25	33	38	38	36
29	51	32	36	53	30	55	44	46	38
29	44	48	30	34	46	47	36	37	36
30	58	42	46	46	29	38	44	40	30
35	35	63	47	37	29	53	41	42	41

Розв'язок. Запишемо в лист Excel вхідні дані завдання (рис. 1.14), стовпчик з границями інтервалів в чарунках M2 : M11 (задається тільки початок інтервалів). Розрахуємо частоти попадання в інтервали (див. зміст командного рядка рис. 1.14).

1			Кількість родин з дітьми дошкільного віку									Інтервали	Частоти
2	27	36	34	46	43	28	29	37	40	43		24	1
3	40	33	50	37	41	32	27	43	34	32		28	4
4	30	41	54	42	47	35	49	49	54	36		32	13
5	36	51	36	24	35	25	33	38	38	36		36	17
6	29	51	32	36	53	30	55	44	46	38		40	11
7	29	44	48	30	34	46	47	36	37	36		44	13
8	30	58	42	46	46	29	38	44	40	30		48	9
9	35	35	63	47	37	29	53	41	42	41		52	5
10												56	5
11												60	1
12													1

Рисунок 1.14. Вхідні дані для побудови гістограми

Викличемо ***Вставка – Диаграмма – Гистограмма***, задамо діапазон даних, тобто розраховані частоти і вкажемо групування за стовпцями (див. рис. 1.15, виділення діапазону чарунок, що містять частоти).

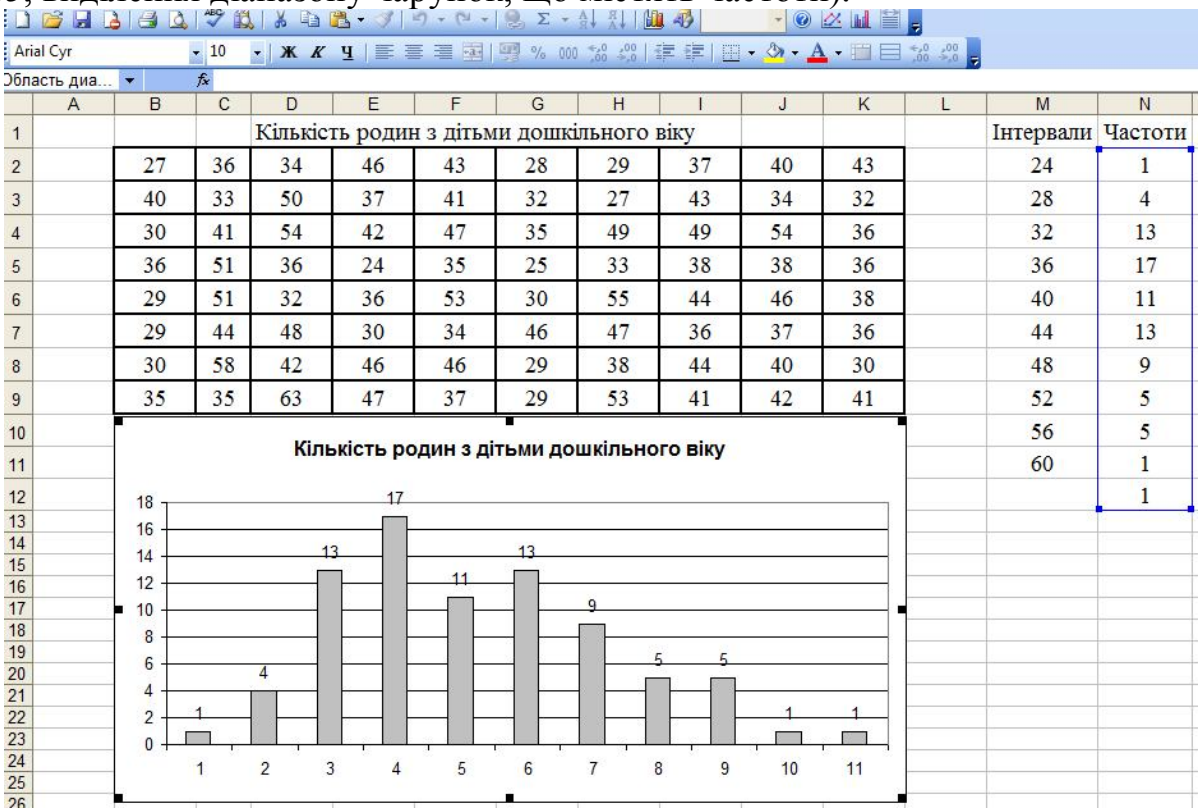

Рис. 1.15. Лист Excel з вхідними даними і гістограмою

Для зручності читання діаграми додамо *Заголовок* та *Значения*. Заберемо відмітку *Легенда*, оскільки імена рядів не було надано – розглядається тільки один тип даних.

Після побудови діаграми можна, у разі необхідності, змінити шрифти, ширину стовпців гістограми, колір стовпців і фону. Для внесення змін потрібно двічі натиснути лівою кнопкою миші на відповідне поле гістограми.

Зауважимо, що на горизонтальній осі надаються не границі інтервалів, а їх порядковий номер.