Лекція 11

ЗМІСТОВИЙ МОДУЛЬ 2. МАТЕМАТИЧНА СТАТИСТИКА

Первинне опрацювання статистичних даних

План

1. Генеральна і статистична сукупності. Статистичний розподіл вибірки.
2. Емпірична функція розподілу.
3. Полігон і гістограма частот та відносних частот.
4. **Чисельні характеристики статистичного розподілу вибірки: вибіркове середнє, вибіркова дисперсія, вибіркове середнє квадратичне відхилення, вибіркові початковий і центральний моменти, мода, асиметрія, ексцес.**
5. **Застосування статистичного розподілу вибірки в економіці.**

Предмет **математичної статистики** полягає в розробці методів збору та обробки статистичних даних для одержання наукових та практичних висновків.

Теоретичним забезпеченням математичної статистики є теорія ймовірності.

Задачі математичної статистики можна умовно розділити на дві групи.

Перша група задач визначає способи збору та групування результатів статистичних досліджень, які можна отримати в процесі спостережень природних та соціальних явищ або за допомогою спеціально поставлених для цього експериментів.

Друга група задач визначає методи аналізу статистичних даних, які включають:

1. оцінку невідомої ймовірності події, оцінку невідомої функції розподілу, оцінку параметрів відомої функції розподілу, оцінку залежності випадкової величини від одної або декількох інших;
2. перевірку статистичних гіпотез про вигляд невідомого розподілу, про числові значення параметрів розподілу, вигляд якого відомий, про порівняння різних розподілів та їх параметрів і т.д.

На підставі висновків статистичних досліджень є можливість прийняття вірних рішень, навіть в умовах неповної інформації про досліджувані явища та неповної визначеності їх параметрів.

Математична статистика виникла (**XVII ст**.) та почала розвиватись паралельно з теорією імовірностей. Подальшим розвитком (кінець XIX – початок XX ст.) математична статистика зобов'язана П.Л.Чебишову, А.А.Маркову, О. М.Ляпунову, а також К.Гауссу, Ф.Гальтону, К. Пірсону та іншим.

У XX ст. найбільший вклад у математичну статистику зробили В.І.Романовський, Е.Е. Слуцький, А.Н.Колмогоров, Стьюдент (псевдонім У. Госсета), Е.Пірсон, Ю.Нейман, А.Вальд, А.В.Скороход, В.С.Королюк та інші вчені.

Генеральна та вибіркова сукупності

Для розв’язування першої групи задач статистичних досліджень вивчають сукупність однорідних об’єктів відносно деякої якісної або кількісної ознаки. При цьому можливі два підходи: або суцільне без винятків вивчення всіх об’єктів, або вивчення вибраної певним чином частини цих об’єктів.

При вивченні частини об’єктів висновки результату досліджень поширюють на всю сукупність об’єктів.

Генеральною сукупністю називають всю сукупність об'єктів, які підлягають статистичному вивченню. (Якщо деталей дуже багато або перевірка пов’язана з руйнуванням деталі (наприклад, випробування деталі на міцність), тоді цей спосіб перевірки не доцільний.

Вибірковою сукупністю або вибіркою називають сукупність випадково вибраних об'єктів, які підлягають статистичному вимірюванню.

Вибірка – це частина генеральної сукупності.

Об'ємом сукупності (вибіркової або генеральної) називають число об'єктів цієї сукупності.

Наприклад, якщо з 5000 виробів для дослідження взято 50, тоді об’єм генеральної сукупності *N = 5000, а об’єм вибірки n = 50.*

Приклад 1.

Спостерігають величину врожаю пшениці на 10 дослідницьких ділянках. Результати спостережень представлені у Таблиці 1 (ознака N – номер ділянки, ознака X – врожай у центнерах з га)

Таблиця 1

N	1	2	3	4	5	6	7	8	9	10
X	15,2	19,1	17,1	20,8	18,4	16,4	22	20,4	17,6	18,2

Таке зведення називають ***рядом варіант*** або ***простим статистичним рядом.***

Вибірка може бути **повторною,** коли об'єкт повертають до сукупності перед відбором наступного, або **безповторною,** коли об'єкт не повертають до сукупності. Найчастіше використовують безповторні вибірки.

Вибірка повинна бути **репрезентативною,** тобто правильно представляти генеральну сукупність. Ця вимога виконується, коли відбір об'єктів ведеться випадково але має достатній об'єм вибірки.

Способи відбору

1. Вибір, який не потребує розділення генеральної сукупності на частини. До цього виду вибору відносять:

– простий випадковий безповторний відбір;

– простий випадковий повторний відбір.

2. Вибір, при якому генеральна сукупність розділяється на частини (розшарований випадковий відбір). До цього виду вибору відносять:

– типовий відбір;

– механічний відбір;

– серійний відбір.

Типовим називають відбір, при якому об'єкти відбирають не з усієї генеральної сукупності, а лише з її типових частин. Наприклад, якщо вироби виготовлені на різних станках, то відбір проводять лише з виробів кожного станка окремо.

Типовий відбір доцільно використовувати тоді, коли однакові вироби виготовляють на станках, серед яких є більш та менш досконалі, або у випадку виготовлення однакових . виробів різними підприємствами.

Механічним називають відбір, при якому генеральна сукупність механічно поділяється на стільки частин, скільки має бути об'єктів у вибірці. З кожної частини випадковим чином відбирають один об'єкт. Наприклад, якщо потрібно перевірити 25% усіх виготовлених станком-автоматом виробів, то відбирають кожен четвертий виріб. Щоб механічний відбір був репрезентативним, треба враховувати специфіку технологічного процесу.

Серійним називають відбір, при якому об'єкти із генеральної сукупності відбирають не по одному, а серіями, які і досліджують. Серійний відбір використовують тоді, коли ознака, яку досліджують, мало змінюється в різних серіях.

В економічних дослідженнях іноді використовують ***комбінований відбір.*** Наприклад, спочатку поділяють генеральну сукупність на серії однакового об'єму, випадковим чином відбирають декілька серій і, нарешті, з кожної серії випадковим чином беруть окремі об'єкти.

Організація даних: статистичний розподіл вибірки

У математичній статистиці замість слова *«дані»* вживається термін *«варіанти»*. Числову характеристику варіанти при цьому називають ***ознакою***.

Нехай із генеральної сукупності взята вибірка об'єктів $\{x_1, x_2, ..., x_n\}$ об'єму n, для вивчення ознаки X. Тобто, значення

$x_1, x_2, ..., x_n$ є *варіанти ознаки* X.

Першим кроком обробки є впорядкування варіант. Розглянемо цей процес на прикладі.

Приклад 1. У Таблиці В наведена вибірка середньомісячної платні 100 співробітників фірми *N*. Треба впорядкувати вибірку.

Таблиця В

Вибірка середньомісячної платні 100 співробітників фірми *N*

338	348	304	314	326	314	324	304	342	308
336	304	302	338	314	304	320	321	322	321
312	323	336	324	312	312	364	356	362	302
322	310	334	292	362	381	304	366	298	304
381	368	304	298	368	290	340	328	316	322
302	314	292	342	321	322	290	332	298	296
296	298	324	338	352	326	318	304	332	322
360	312	331	331	304	316	332	282	342	338
342	322	324	325	302	328	354	330	316	324
334	350	334	324	332	340	324	314	326	323

У нашому прикладі ознака є число, що виражає середньомісячну зарплату співробітників фірми *N*. Отже, у Таблиці В наведено 100 значень варіант.

Розмістимо дані Таблиці Б у порядку зростання (див. Таблицю В.1).

Таблиця В.1

Впорядкована вибірка середньомісячної платні 100 співробітників фірми (у порядку зростання)

282	298	304	314	321	323	326	332	340	356
290	302	304	314	321	324	326	334	340	360
290	302	304	314	321	324	328	334	342	362
292	302	304	314	322	324	328	334	342	362
292	302	308	314	322	324	330	336	342	364
296	304	310	316	322	324	331	336	342	366
296	304	312	316	322	324	331	338	348	368
298	304	312	316	322	324	332	338	350	368
298	304	312	318	322	325	332	338	352	381
298	304	312	320	323	326	332	338	354	381

Розподіл частот

Нехай у нашій вибірці із n варіант x_1, x_2, ..., x_m ознака X прийняла значення $x_1 - n_1$ раз, значення $x_2 - n_2$ раз, ..., значення $x_m - n_m$ раз.

Додатне число, що вказує, скільки раз та чи інша варіанта зустрічається в таблиці даних, називається ***частотою***.

Ряд

$$n_1,\ n_2,\ ...,\ n_m$$

називається ***рядом частот***. Відмітимо, що сума усіх частот повинна дорівнювати об'єму вибірки

$$\sum_{i=1}^{m} n_i = n. \qquad (1)$$

Статистичний розподіл вибірки встановлює зв'язок між рядом варіант, що зростає або спадає, і відповідними частотами. Він може бути представлений таблицею

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

де n — об'єм вибірки, $n = n_1 + n_2 + ... + n_m$.

Статистичний розподіл вибірки, заданий цією таблицею, також називають ***простим*** чи ***незгрупованим статистичним розподілом*** або ***розподілом частоти варіанти*** x_i ***(рядом розподілу частоти варіанти*** x_i***)***.

Приклад 2. Для вивчення потреб у певних розмірах взуття власник магазину спостерігає розміри взуття, проданого на протязі дня:

40, 35, 37, 39, 40, 41, 36, 42, 40, 39, 36, 43, 43, 41, 38, 37, 36, 42, 40, 38.

Статистичний розподіл цієї вибірки (розподіл частоти розміру взуття) буде мати такий вигляд

Розмір взуття x_i	35	36	37	38	39	40	41	42	43
Кількість n_i	1	3	2	2	2	4	2	2	2

Контроль: $1 + 3 + 2 + 2 + 2 + 4 + 2 + 2 + 2 = 20$, $n = 20$.

Наведемо подальші удосконалення вибірки, що фігурує у Прикладі 1 (Таблиця В.1), перетворивши її в розподіл частот середньомісячної платні співробітників фірми N (Таблиця В.2).

Таблиця В.2

Розподіл частоти середньомісячної платні співробітників фірми N

x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
282	1	314	5	328	2	350	1
290	2	316	3	330	1	352	1
292	2	318	1	331	2	354	1
296	2	320	1	332	4	356	1
298	4	321	3	334	3	360	1
302	4	322	6	336	2	362	2
304	9	323	2	338	4	364	1
308	1	324	7	340	2	366	1
310	1	325	1	342	4	368	2
312	4	326	3	348	1	381	2
	30		32		25		13
Разом:							100

Подальший крок в обробці даних, що призводить до суттєвого спрощення дослідження, є їх *згрупування*.

Як видно з Таблиці В.2 максимальне та мінімальне значення варіанти будуть

$$x_{\max} = 381, \quad x_{\max} = 282.$$

Різниця цих чисел

$$R = x_{\max} - x_{\max}$$

називається ***розмахом варіант***. У нашому випадку $R = 99$.

Введемо для варіанти інтервали зміни платні.

280–290	310–320	340–350	370–380
290–300	320–330	350–360	380–390
300–310	330–340	360–370	

(1)

Кожний інтервал називається ***класом інтервалів*** або ***класом***. Всього маємо $k = 11$ класів платні.

Використовуючи дані Таблиці В.2, просумуємо частоти для кожного класу інтервалів (1), причому значення x_i, що знаходяться на границі класів, заносимо до того класу, що є слідуючим до класу, де це число зустрілось вперше. Результат перепишемо її у вигляді Таблиці В.3.

Таблиця В.3

Згрупований розподіл частоти середньомісячної платні співробітників фірми *N*

Інтервали платні	Частоти n_i	
280–290	1	(=1)
290–300	10	(=2+2+2+4)
300–310	14	(=4+9+1)
310–320	14	(=1+4+5+3+1)
320–330	25	(=1+3+6+2+7+1+3+2)
330–340	16	(=1+2+4+3+2+4)
340–350	7	(=2+4+1)
350–360	4	(=1+1+1+1)
360–370	7	(=1+2+1+1+2)
370–380	0	(=0)
380–390	2	(=2)
Разом:	100	

Таблиця вигляду В.3, яка встановлює зв'язок між згрупованим рядом варіант, що зростає або спадає, та сумами їхніх частот по класах, називається ***згрупованим розподілом частоти варіанти*** x_i.

Для кожного класу маємо верхню та нижню границі, наприклад, для першого і другого класу інтервалів маємо

$$x_{1\min} = 280, \quad x_{1\max} = 290$$

$$x_{2\min} = 290, \quad x_{2\max} = 300.$$

Шириною класу h_i називається число одиниць виміру у цьому класі, тобто різниця

$$h_i = x_{i\max} - x_{i\min}.$$

У нашому випадку ширина класів однакова і дорівнює $h = 10$.

Згрупований розподіл накопиченої частоти

Часто поряд з розподілом частоти варіанти необхідно мати розподіл ***накопиченої (кумулятивної) частоти.*** Розподіл накопиченої частоти одержується послідовним додаванням частот чергового інтервалу, починаючи з першого і кінчаючи останнім (див. Таблицю В.4).

Таблиця В.4

Згрупований розподіл накопиченої частоти середньомісячної платні співробітників фірми *N*

Інтервали платні	Частоти n_i	Платня		Накопичені частоти F_i	
280–290	1	менше ніж	290	1	(=1)
290–300	10	менше ніж	300	11	(=1+10)
300–310	14	менше ніж	310	25	(=11+14)
310–320	14	менше ніж	320	39	(=25+14)
320–330	25	менше ніж	330	64	(=39+25)
330–340	16	менше ніж	340	80	(=64+16)
340–350	7	менше ніж	350	87	(=80+7)
350–360	4	менше ніж	360	91	(=87+4)
360–370	7	менше ніж	370	98	(=91+7)
370–380	0	менше ніж	380	98	(=98+0)
380–390	2	менше ніж	390	100	(=98+2)
Разом:	100				

Розподіл відносної частоти (частості) вибірки

Нерідко замість значень частот використовуються відносні частоти. Нехай існує m частот $n_1, \ldots, n_m \left(n_1 + \ldots + n_m = n\right)$. Відношення частоти n_i варіанти x_i до об'єму вибірки n

$$W_i = \frac{n_i}{n}$$

називається ***відносною частотою*** або ***частостью***, причому, сума усіх відносних частот

$$\sum_{i=1}^{m} W_i = 1. \qquad (2)$$

Залежність між упорядкованим рядом варіант і відповідними їм відносними частотами або частостями також називається ***статистичним розподілом вибірки***, тобто маємо табличне представлення розподілу

x_i	x_1	x_2	…	x_m
W_i	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_m}{n}$

де n – об'єм вибірки і $x_1 < x_2 < \ldots < x_m$.

Приклад 3. Заданий розподіл частот вибірки

x_i	4	6	9
n_i	10	3	7

Знайти розподіл частостей.

Розв'язання. Об'єм вибірки $n = 10 + 3 + 7 = 20$. Частостями будуть

$$W_1 = \frac{10}{20} = 0{,}5; \quad W_2 = \frac{3}{20} = 0{,}15; \quad W_3 = \frac{7}{20} = 0{,}35.$$

Тому розподіл частостей цієї вибірки буде

x_i	4	6	9
W_i	0,5	0,15	0,35

Згруповані розподіли відносної та накопиченої відносної частот середньомісячної платні співробітників фірми N

Інтервали платні	Відносні частоти $W_i = n_i / n$	Платня	Накопичені відносні частоти F_i/n
280–290	0.01	менше ніж 290	0.01
290–300	0.10	менше ніж 300	0.11
300–310	0.14	менше ніж 310	0.25
310–320	0.14	менше ніж 320	0.39
320–330	0.25	менше ніж 330	0.64
330–340	0.16	менше ніж 340	0.80
340–350	0.07	менше ніж 350	0.87
350–360	0.04	менше ніж 360	0.91
360–370	0.07	менше ніж 370	0.98
370–380	0.00	менше ніж 380	0.98
380–390	0.02	менше ніж 390	1
Разом:	100		

Згрупований розподіл щільності частоти і щільності відносної частоти (частості)

Якщо поділити всі частоти (другий стовпчик) Таблиці В.3 на ширину інтервалу $h = 10$, то отримаємо ***розподіл щільності частоти*** **вибірки**

$$\frac{n_i}{h}.$$

Якщо поділити всі відносні частоти (другий стовпчик) Таблиці В.5 на ширину інтервалу $h = 10$, то отримаємо ***розподіл щільності відносної частоти (частості)*** вибірки

$$\frac{W_i}{h}.$$

Згрупований розподіл частоти (n_i), відносної частоти (W_i), щільності частоти (n_i/h), щільності відносної частоти (W_i/h), накопиченої частоти (F_i) і накопиченої відносної частоти (F_i/n) вибірки середньомісячної платні 100 співробітників фірми N

x	n_i	W_i	$\frac{n_i}{h}$	$\frac{W_i}{h}$	x	F_i	$\frac{F_i}{n}$
280–290	1	0,01	0,1	0,001	менше ніж 290	1	0,01
290–300	10	0,10	1,0	0,010	менше ніж 300	11	0,11
300–310	14	0,14	1,4	0,014	менше ніж 310	25	0,25
310–320	14	0,14	1,4	0,014	менше ніж 320	39	0,39
320–330	25	0,25	2,5	0,025	менше ніж 330	64	0,64
330–340	16	0,16	1,6	0,016	менше ніж 340	80	0,80
340–350	7	0,07	0,7	0,007	менше ніж 350	87	0,87
350–360	4	0,04	0,4	0,004	менше ніж 360	91	0,91
360–370	7	0,07	0,7	0,007	менше ніж 370	98	0,98
370–380	0	0,00	0,0	0,000	менше ніж 380	98	0,98
380–390	2	0,02	0,2	0,002	менше ніж 390	100	1,00
Разом:	100		1,00				

2. Емпірична функція розподілу та її властивості

Нехай є статистичний розподіл частот деякої ознаки X. Позначимо через n загальну кількість спостережень, тобто об'єм вибірки; n_x – кількість спостережень, при яких спостерігались ознаки X менше x.

Тоді відносна частота (або частість) події $X < x$ дорівнює $\frac{n_x}{n}$.

Якщо x змінюється, то може змінюватись відносна частота, тобто $\frac{n_x}{n}$ є функція від x. Ця функція знаходиться емпіричним (дослідним) шляхом, тому її називають ***емпіричною***.

Означення 1. ***Емпіричною функцією розподілу (або функцією розподілу вибірки)*** *називають функцію $F^*(x)$, яка визначає для кожного значення x частість події $X < x$.*

Математично це означення має вигляд

$$F^*(x) = \frac{n_x}{n},$$

де n_x – кількість варіант, які менше від x, n – об'єм вибірки.

Таким чином, щоб знайти, наприклад, $F^*(x_3)$, треба кількість варіант, що менше x_3, поділити на об'єм вибірки, тобто

$$F^*(x_3) = \frac{n_1 + n_2}{n}.$$

Зауваження. *Інтегральну функцію розподілу $F(x)$ генеральної сукупності у математичній статистиці називають* ***теоретичною функцією розподілу****. Вона відрізняється від емпіричної функції розподілу $F^*(x)$ тим, що визначає імовірність події $X < x$, а не частість цієї події.*

З теореми Бернуллі випливає, що частість

$$F^*(x) = \frac{n_x}{n} \quad \text{події} \quad X < x$$

прямує до імовірності

$$F(x) = P(X < x)$$

цієї події. Тому $F(x)$ та $F^*(x)$ мало відрізняються одна від одної.

Доцільно використовувати $F^*(x)$ для наближеного представлення функції розподілу $F(x)$ генеральної сукупності.

Емпірична функція розподілу $F^*(x)$ має такі властивості:

1. $0 \le F^*(x) \le 1$.
2. $F^*(x)$ – зростаюча функція.
3. $$F^*(x) = \begin{cases} 0, & x \le x_1; \\ 1, & x > x_m, \end{cases}$$

де x_1 – найменша варіанта, x_m – найбільша варіанта.

Приклад 1. Знайти емпіричну функцію розподілу за статистичним розподілом вибірки

Таблиця 5

x_i	2	6	10
n_i	12	18	30

та побудувати її графік.

Розв'язаним. Об'єм цієї вибірки буде $n = 12 + 18 + 30 = 60$. Найменша варіанта дорівнює 2, тому $F^*(x) = 0$ для $x \le 2$. Найбільша варіанта дорівнює 10, тому $F^*(x) = 1$ для $x > 10$. Значення $x < 6$, тобто $X = \{x_1 = 2\}$, спостерігалось 12 разів, тому $F^*(x) = \frac{12}{60} = 0,2$ при $2 < x \ge 6$.

Значення $X < 10$, тобто $X = \{x_1 = 2\}$ та $X = \{x_2 = 6\}$, спостерігались $12 + 18 = 30$ разів, тому $F^*(x) = \frac{30}{60} = 0,5$ при $6 < x \le 10$.

Тобто, простий статистичний розподіл частоти, що заданий Таблицею 5, замінюється згрупованим розподілом частоти (див. Таблицю 6).

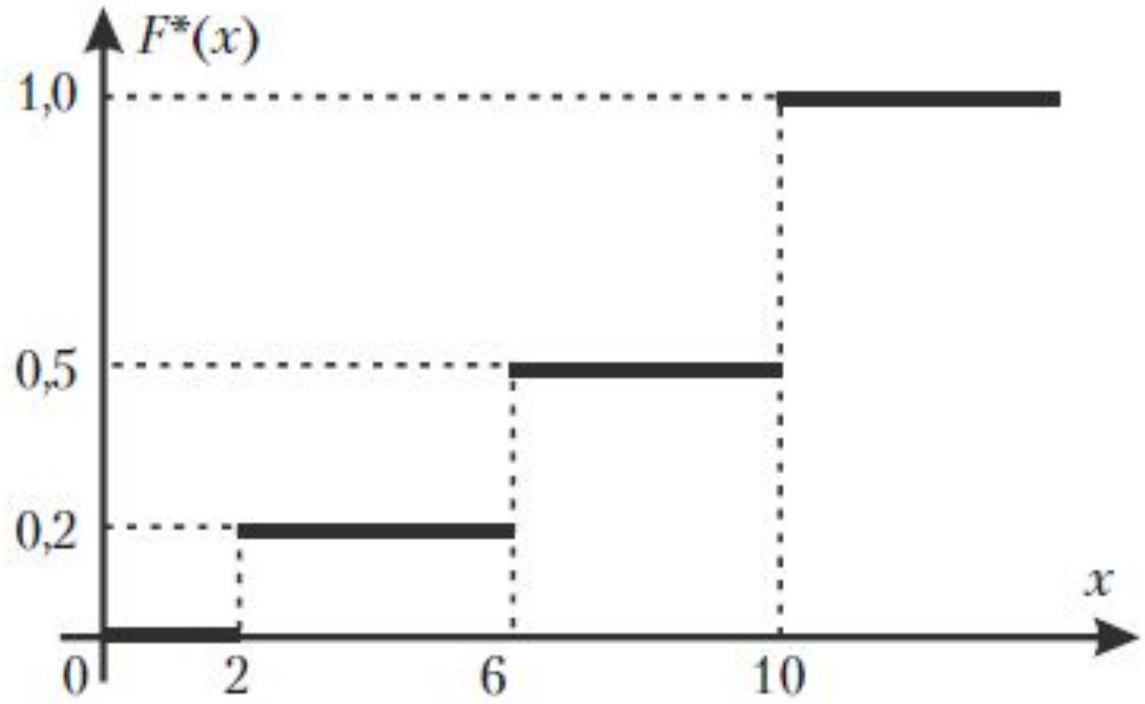
Таблиця 6

Варіанта x_i	Частота n_i	Варіанта x_i	Накопичені частоти F_i
$x \le 2$	0	менше ніж 2	0
$2 < x \le 6$	12	менше ніж 6	12
$6 < x \le 10$	18	менше ніж 10	30
$10 < x$	30	більше ніж 10	60
Разом:	60		

Тут же побудований розподіл накопиченої частоти. Таким чином, одержали емпіричну функцію розподілу вигляду

$$F^*(x) = \begin{cases} 0, & x \le 2; \\ 0{,}2, & 2 < x \le 6; \\ 0{,}5, & 6 < x \le 10; \\ 1, & x > 10. \end{cases}$$

Графік цієї функції зображено на мал. 18.



Мал. 18.

3. Полігон і гістограма частот та відносних частот

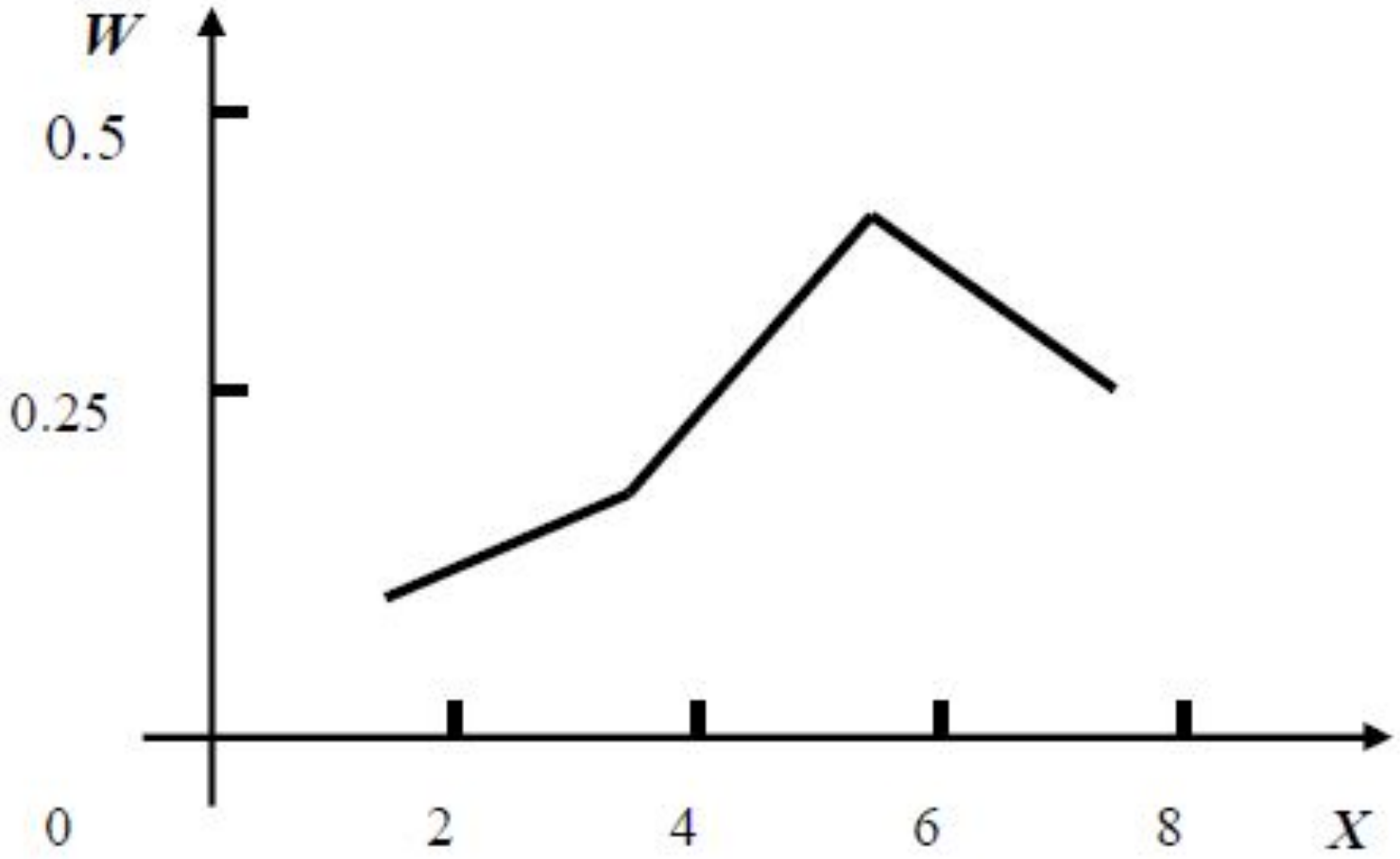
Для наочності часто будують графічні зображення статистичного розподілу, зокрема полігон та гістограму.

Полігоном частот називають ламану лінію, відрізки якої з'єднують точки $(x_1; n_1)$, $(x_2; n_2)$ і т.д. Для побудови полігона частот на вісі абсцис відкладають варіанти x_i, а на вісі ординат відповідні їх частоти n_i; точки $(x_1; n_1)$ з'єднують ламаною лінією.

Полігоном відносних частот називають ламану лінію, відрізки якої з'єднують точки $(x_1; w_1)$, $(x_2; w_2)$ і т.д.

Приклад: Побудуємо полігон відносних частот для розподілу:

X_i	1,5	3,5	5,5	7,5
W_i	0,1	0,2	0,4	0,3


W
0.5
0.25
0
2
4
6
8
X

Гістограми – ступінчасті фігури, що складаються з прямокутників, основами яких є рівні інтервали варіант шириною ***h***, а висоти рівні відношенню $\frac{n_1}{h}$, або $\frac{w_1}{h}$. Перші називають гістограмами частот, другі – гістограмами відносних частот.

Для побудови гістограм, на вісі абсцис відкладають паралельні вісі абсцис відрізки на відстані $\frac{n_1}{h}$, або $\frac{w_1}{h}$ по вісі ординат.

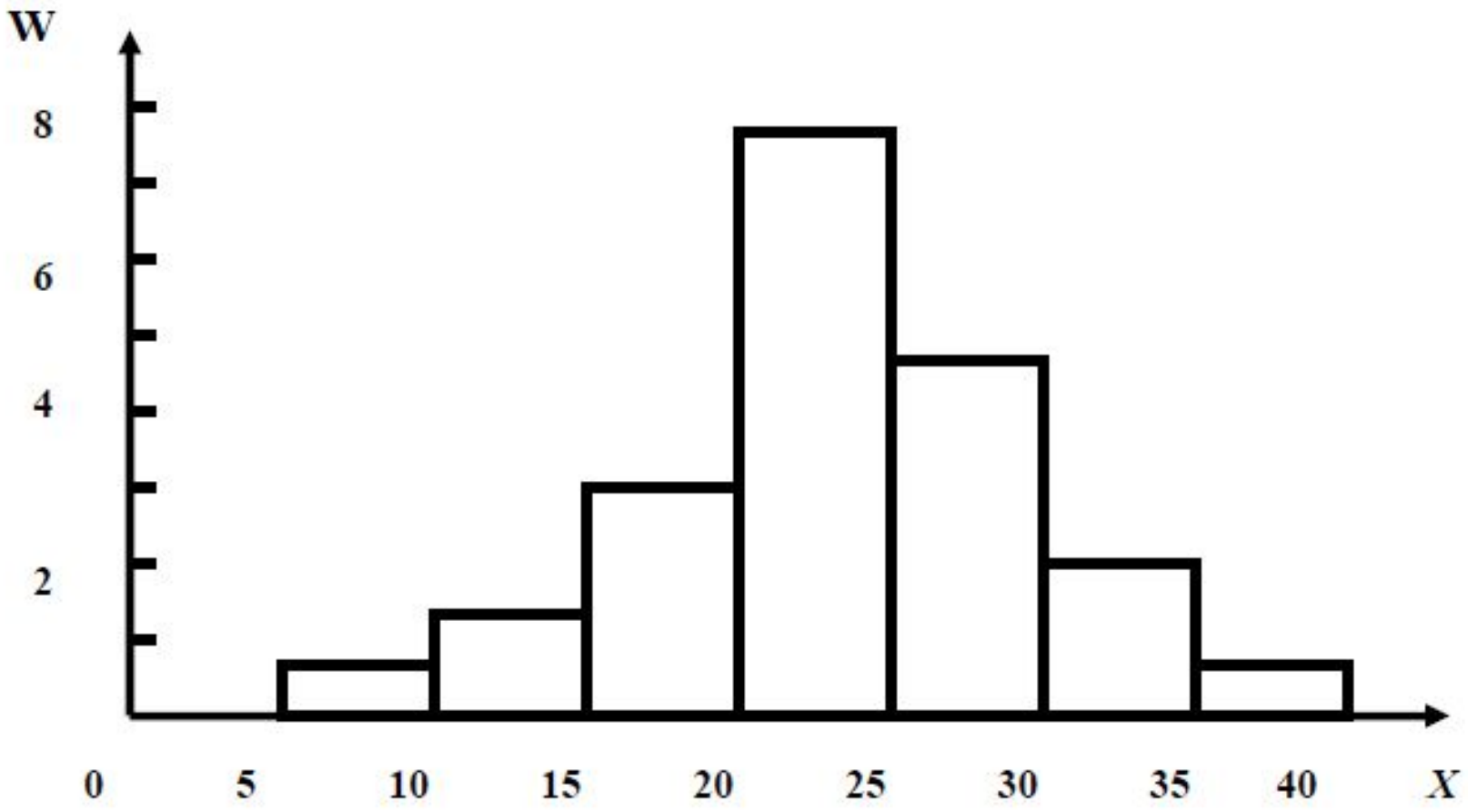
Гістограми мають такі властивості. Площа під гістограмою частот дорівнює об'єму вибірки, а площа гістограми відносних частот дорівнює одиниці.

Полігони частіше будують при невеликому об'ємі вибірки, а гістограми при великому об'ємі.

Приклад: Побудувати гістограму частот розподілу для об'єму n = 100 наведеному в таблиці.

Частинний інтервал $h = 5$	Сума частот варіант n_i	Густина частоти n_i / h
5-10	4	0,8
10-15	6	1,2
15-20	16	3,2
20-25	36	7,2
25-30	24	4,8
30-35	10	2,0
35-40	4	0,8

Відповідна табличним даним гістограма має вигляд:



Запитання для самоперевірки:

1. Надайте визначення генеральної та вибіркової сукупності
2. Надайте означення варіанті, варіаційному ряду, частоті, відносній частоті варіант.
3. Дайте означення дискретного статистичного розподілу вибірки і вкажіть його характеристики
4. Що називається інтервальним статистичним розподілом вибірки?
5. Назвіть характеристики інтервального статистичного розподілу та надайте формули для їх обчислення
6. Що являє собою полігон частот і відносних частот? Гістограма частот і відносних частот?
7. Асіметрія і ексцес статистичного розподілу вибірки
8. Що називається емпіричною функцією (комулятою)? Властивості комуляти.