ЛЕКЦІЯ 3
ВАРІАЦІЙНИЙ АНАЛІЗ

1. Варіаційна статистика: основні поняття.
2. Розподіл частот та табулювання даних.
3. Графічне представлення емпіричних даних.

1. Варіаційна статистика

Варіаційною статистикою називають обчислення числових та функціональних характеристик емпіричного розподілу.

Варіаційний ряд – це ряд даних, впорядкованих у порядку їхнього зростання чи зменшення. Він може бути побудований з кількісних або порядкових вибірок.

Ряд чисел, які характеризують розподіл одиниць досліджуваної сукупності залежно від величини ознаки, називають ***варіаційним рядом***.

Якщо групування здійснено за інтервалами зміни ознаки, то таке групування називається ***інтервальним***. Подавши результат групування рядом варіант або інтервалів варіацій, розміщених у зростаючій послідовності, і низкою відповідних частот, дістанемо ***варіаційний ряд (дискретний або інтервальний)***.

Дискретним або перервним варіаційним рядом називаються розташовані в порядку зростання варіанти

$$x_1, x_{2,\dots}, x_k$$

які являють собою окремі ізольовані одне від іншого значення варійованої ознаки x . Якщо кількість варіант у дискретному ряду занадто велика або ознака x генеральної сукупності є неперервною випадковою величиною, то виникає потреба побудови інтервального варіаційного ряду.

Інтервальним (неперервним) варіаційним рядом називається ряд, в якому значення варіант задані у вигляді інтервалів, тобто значення ознаки X можуть відрізнятись одне від одного на яку завгодно малу величину.

Частотою значення ознаки (або інтервалу) називають кількість членів сукупності з деякою варіантою або відповідно кількість членів сукупності, варіанти яких лежать у даному інтервалі.

Рядом розподілу називають ряд чисел, які характеризують розподіл одиниць досліджуваної сукупності.

Нехай у даній статистичній сукупності вивчається деяка ознака (вона змінюється з переходом від одного члена статистичної сукупності до іншого). Зміну цієї ознаки називають її **варіацією**, а значення ознаки у даного члена статистичної сукупності – його **варіантою**.

Відповідність між варіантами варіаційного ряду та їх частотами (або відносними частотами) називається **статистичним розподілом вибірки.**

Маємо варіаційний ряд: $X_1, X_2, X_3, \dots, X_k$,

де X_i — числа, які показують зміну (варіацію) ознаки, що вивчається, і називаються **варіантами**;

і — номер варіанти (і = 1,2,3, …к).

Приклад.

При обстеженні студентів 1-го курсу за віком було зафіксовано такі дані: 17,18,18,18,18,19,20,20,20,21,21,21,17,18,18,18,19,20,20,20,21,21,21,24.

Якщо впорядкувати ці дані у зростаючому або спадному порядку, то отримаємо **ранжований ряд**. Числа, які показують, скільки разів (як часто) зустрічаються окремі значення варіант, називаються **частотами**.

Позначимо частоту і-ї варіанти X_i через n_i, тоді ранжований дискретний варіаційний ряд запишеться у вигляді:

Варіаційний ряд у загальному вигляді			Варіаційний ряд для прикладу 1		
Номер: варіанти	Значення варіанти	Частота варіанти	Номер варіанти	Значення варіанти	Частота варіанти
1	X_1	n_1	1	17	2
2	X_2	n_2	2	18	7
.	.	.	3	19	2
.	.	.	4	20	6
.	.	.	5	21	6
k	X_k	n_k	6	24	1

2. Розподіл частот і табулювання даних.

Нехай із генеральної сукупності вилучена вибірка обсягу n. Досліджується деяка ознака (наприклад, рівень інтелекту, час реакції тощо). Тоді кожен елемент вибірки може набувати різних значень досліджуваної ознаки, які позначають x_1 x_2, ..., x_k, где $k \leq n$. Значення ознаки називають варіантом x_i де *i* - порядковий номер варіанти.

Послідовність варіант, упорядкована за зростанням, називається варіаційним рядом. Число появ варіанти x_i називають частотою варіанти і позначають n_i.

Наприклад, в результаті дослідження отримані наступні дані: 8, 5, 7, 8, 5, 8, 6.

Представимо їх у вигляді варіаційного ряду: 5, 5, 6, 7, 8, 8, 8.

Досліджувана ознака приймає чотири значення $x_1 = 5$, $x_2 = 6$, $x_3 = 7$, $x_4 = 8$, які мають наступну частоту: $n_1=2$, $n_2= 1$, $n_3= 1$, $n_4=3$. Сума частот усіх варіант дорівнює обсягу вибірки

$n_1+n_2+\ldots+n_k=n$

або

$$\sum_{i=1}^{k} n_i = n$$

Для попереднього прикладу $n_1+n_2+\ldots+n_k=2+1+1+3=7$

Відношення частоти варіанти n_i, до обсягу вибірки n називається відносною частотою варіанти і позначається w_i:

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

Сума всіх відносних частот дорівнює 1: $W_1 + W_2+...+ W_K= 1$

Для попереднього прикладу $W_1 + W_2+ ...+ W_k= \frac{2}{7}+\frac{1}{7}+\frac{1}{7}+\frac{3}{7}=\frac{7}{7}=1$

Приклад .

Провели експеримент. 60 осіб протягом певного часу мали дати відповіді (так або ні) на запитання. Підрахували кількість негативних відповідей. Далі розташували їх в порядку неспадання і отримали 7 груп спостережень:

К-ть негат. відповідей	**Кількість осіб**
0	8
1	17
2	16
3	10
5	6
6	2
7	1

Індекс	i	1	2	3	4	5	6	7
К-ть негат. відповідей	x_i	0	1	2	3	4	5	7
Частота	m_i	8	17	16	10	6	2	1
Відносна частота	W_i	$\frac{8}{60}$	$\frac{17}{60}$	$\frac{16}{60}$	$\frac{10}{60}$	$\frac{6}{60}$	$\frac{2}{60}$	$\frac{1}{60}$

Первинна обробка даних, отриманих в результаті вимірювання, полягає в їх описі, впорядкування, табулюванні і поданні у вигляді, зручному для подальшої обробки. Для цього вибірку представляють у вигляді статистичного розподілу, яке може бути задано двома способами:

а) у вигляді розподілу частот (відносних частот) - переліку варіант і відповідних їм частот (відносних частот);

б) у вигляді інтервального розподілу (розподілу згрупованих частот) - послідовності інтервалів і відповідних їм частот (відносних частот).

Розподіл частот, як правило, використовується в разі, якщо вимірювана змінна є дискретною, а інтервальний розподіл - якщо змінна безперервна.

Приклад.

В результаті емпіричного дослідження отримані наступні дані:

1, 2, 1, 3, 5, 6, 7, 1, 2, 4, 5, 6, 3.

Задати статистичний розподіл вибірки.

Розв'язання.

Визначимо обсяг вибірки: n= 13.

Побудуємо варіаційний ряд: 1, 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 6, 7. Задамо статистичний розподіл вибірки у вигляді частот і відносних частот:

x_i	1	2	3	4	5	6	7
n_i	3	2	2	1	2	2	1
w_i	$\frac{3}{13}$	$\frac{2}{13}$	$\frac{2}{13}$	$\frac{1}{13}$	$\frac{2}{13}$	$\frac{2}{13}$	$\frac{1}{13}$

Контроль: 3+2+2+1+2+2+1=13;

$\frac{3}{13}+\frac{2}{13}+\frac{2}{13}+\frac{1}{13}+\frac{2}{13}+\frac{2}{13}+\frac{1}{13}=\frac{13}{13}=1$

Якщо досліджувана змінна приймає велику кількість різних значень, то зручніше використовувати статистичний розподіл у вигляді інтервального розподілу. Для цього роблять табулювання даних, тобто представляють вихідну вибірку у вигляді таблиці відповідної структури. Табулювання даних здійснюється в чотири етапи:

1 етап - визначення розмаху вибірки R. Для цього з максимального значення вибірки віднімають мінімальну: $\mathbf{R = x_{max} - x_{min}}$

2 етап - визначення ширини інтервалу групування даних h. Перш ніж шукати ширину інтервалу, необхідно визначити кількість інтервалів групування

$$k \approx \sqrt{n}$$

n – об'єм вибірки

Після цього в якості необхідної кількості інтервалів вибирається ціле число, отримане в результаті округлення значення √n в бік збільшення;

Більш точною вважається формула, запропонована Стерджессом

$$k = 1 + 3{,}322 lg_n$$

Для вибірок об'єму n >1000 слід віддавати перевагу формулі Стерджесса

Ширина інтервалу групування h виходить шляхом ділення розмаху вибірки на кількість інтервалів:

$$h = \frac{R}{k}$$

3 етап - визначення меж часткових інтервалів групування даних. При цьому ліва межа першого інтервалу повинна бути менше або дорівнювати x_{mjn}. Кожна наступна межа виходить з попередньої шляхом додавання ширини інтервалу. Права межа останнього інтервалу повинна бути більше або дорівнювати x_{max}:

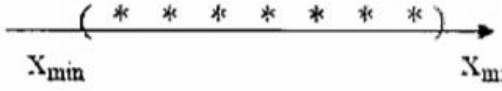


Рис. 2.1 – Графическое представление границ частичных интервалов группирования данных

4 етап - безпосередньо табулювання даних. При цьому підраховується, скільки елементів вибірки потрапило в кожен частковий інтервал. Значення, що потрапляють точно на межу інтервалу, враховуються один раз. Результатом табулювання даних є

таблиця, що складається з трьох стовпців, перший з яких містить кордону часткових інтервалів, другий - частоти, третій - відносні частоти.

Приклад.

В результаті вимірювання швидкості читання в класі з 38 учнів були отримані наступні результати:

90, 66, 106, 84, 105, 83, 104, 82, 97, 97, 59, 95, 78, 70, 47, 95. 100, 69, 44, 80, 75, 75, 51, 109, 89, 58, 59, 72, 74, 75, 81, 71, 68, 112, 62, 91, 93, 84.

Задати статистичний розподіл вибірки.

Розв'язання.

Скористаємося інтервальним розподілом частот.

1) визначимо розмах вибірки:

$x_{max}=112$, $x_{min}=44$

$R = x_{max}-x_{min}= 112-44=68$

2) визначимо ширину інтервалу групування даних h:

n=38

$k\approx\sqrt{n}=\sqrt{38}\approx6,2\approx7$

$h=\frac{R}{k}=\frac{68}{7}\approx9,7\approx10$

Ліву межу першого часткового інтервалу вибираємо рівною X_{min} = 44, всі наступні межі отримуємо з попередньої, додатком ширини інтервалу групування h=10.

Межа часткових інтервалів	Частота	Відносна частота
[44; 54)	3	$\frac{3}{38}$
[54; 64)	4	$\frac{4}{38}$
[64; 74)	6	$\frac{6}{38}$
[74; 84)	9	$\frac{9}{38}$
[84; 94)	6	$\frac{6}{38}$
[94; 104)	5	$\frac{5}{38}$
[104; 114)	5	$\frac{5}{38}$
Сума частот	38	$\frac{38}{38}=1$

Для контролю правильності групування потрібно обчислити суму частот, яка дорівнює обсягу вибірки. Аналіз інтервального розподілу дозволяє зробити висновок, що основна частина учнів (21 осіб) читає зі швидкістю 64-94 слова в хвилину. Але є учні (7 осіб), які читають з невисокою швидкістю, а також школярі (10 осіб) з високою швидкістю читання.

3. Графічне представлення емпіричних даних

Графічне представлення результатів дослідження дозволяє проводити деякі узагальнення вихідних даних. Найчастіше використовується два основних способи графічного представлення даних: полігон частот (відносних частот), гістограма частот.

Полігоном частот (полігоном розподілу) називають ламану, відрізки якої з'єднують точки (x_1, n_1), (x_2, n_2), ... , (x_i, n_i). Для побудови полігону частот на осі абсцис відкладають варіанти x_i, а на осі ординат - відповідні їм частоти n_i. Отримані точки (x_i, n_i) з'єднують відрізками. Полігон частот дозволяє в графічному вигляді представити варіювання досліджуваної ознаки.

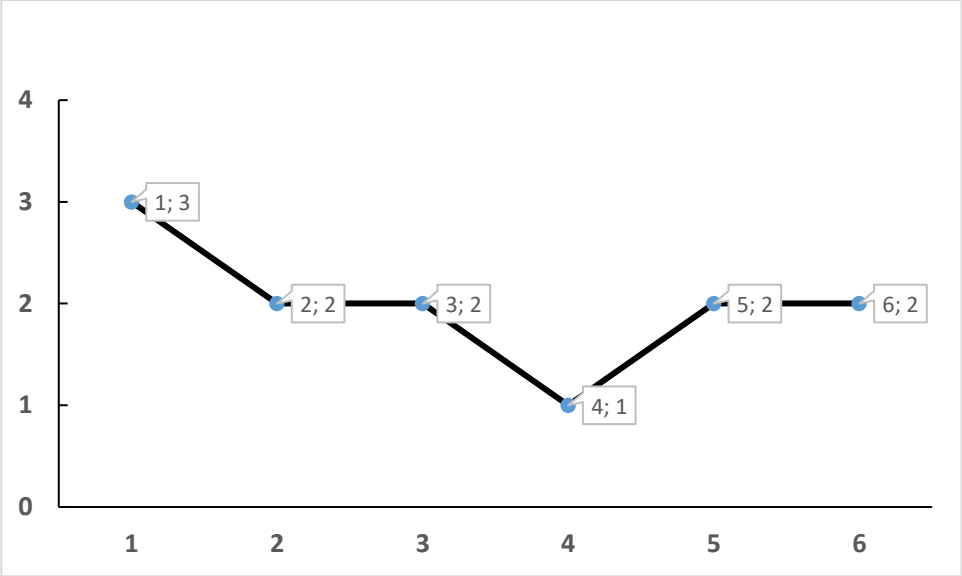
Полігоном відносних частот називають ламану, відрізки якої з'єднують точки (x_1, w_1), (x_2, w_2), ... , (x_i, w_i). Для побудови полігону відносних частот на осі абсцис відкладають x_i, а на осі ординат – відповідні їм частоти W_j. Отримані точки (x_i, w_i) з'єднують відрізками.

Побудувати полігон частот для даних з прикладу (сл 16).

Рішення. Скористаємося розподілом частот, отриманих в прикладі 1.

x_i	1	2	3	4	5	6	7
n_i	3	2	2	1	2	2	1
w_i	$\frac{3}{13}$	$\frac{2}{13}$	$\frac{2}{13}$	$\frac{1}{13}$	$\frac{2}{13}$	$\frac{2}{13}$	$\frac{1}{13}$

Побудуємо точки з координатами: (1; 3), (2; 2), (3; 2), (4; 1), (5; 2), (6; 2), (7; 1) і з'єднаємо їх відрізками.



Полігон частот, побудований на основі статистичного розподілу емпіричних даних.

Гістограмою частот (гістограмою) називають ступінчасту фігуру, що складається з прямокутників, підставами яких служать часткові інтервали шириною h, а висотою - частота n.

При побудові гістограми розподілу варіаційного ряду з нерівними інтервалами по осі ординат завдають не частоти, а щільність розподілу ознаки у відповідних інтервалах.

Щільність розподілу – це частота, розрахована на одиницю ширини інтервалу.

Для побудови гістограми частот на осі абсцис відкладають часткові інтервали, а над ними будують прямокутники висотою n_i.

Межі часткових інтервалів	[44; 54)	[54; 64)	[64; 74)	[74; 84)	[84; 94)	[94; 104)	[104; 114)
Частота	3	4	6	9	6	5	5