Лекція 1

ВИБІРКОВИЙ МЕТОД, ЙОГО ХАРАКТЕРИСТИКИ ТА ПРИКЛАДИ ЗАСТОСУВАННЯ В СФЕРІ ІТ

1.1. Генеральна та вибіркова сукупності

Розрізняють два види спостережень*:* загальне, коли вивчаються всі об'єкти, і незагальне, вибіркове, коли вивчається частина об'єктів. Прикладом загального спостереження є перепис населення, що охоплює все населення країни. Вибірковим спостереженням є, наприклад, соціологічні дослідження, які охоплюють частину населення країни, області, району, дослідження ефективності того чи іншого програмного продукту і т.ін.

Вся сукупність об'єктів (спостережень), що підлягають вивченню, називається **генеральною сукупністю**. В математичній статистиці поняття генеральної сукупності трактується як сукупність всіх уявних спостережень, які могли б бути проведені за даним реальним комплексом умов. Поняття генеральної сукупності в певному значенні є аналогічним до поняття випадкової величини (закону розподілу ймовірностей, ймовірносному простору).

Та частина об'єктів, яка відібрана для безпосереднього вивчення з генеральної сукупності, називається **вибірковою сукупністю**, або **вибіркою**.

Кількість об'єктів (спостережень) в генеральній чи вибірковій сукупності називається їх **обсягом**. Генеральна сукупність може мати як скінченний, так і нескінченний обсяг.

Вибірку можна розглядати як деякий емпіричний аналог генеральної сукупності.

Сутність вибіркового методу полягає в тому, щоб за деякою частиною генеральної сукупності (за вибіркою) видавати міркування про її властивості в цілому.

Основні переваги вибіркового методу

1. Дозволяє суттєво економити на витраті ресурсів (матеріальних, трудових, часових).
2. Є єдино можливим у випадку нескінченної генеральної сукупності чи у випадку, коли дослідження пов'язане зі знищенням об'єктів, за якими ведеться спостереження (наприклад, дослідження довговічності мікрочипів, граничних режимів роботи електронних пристроїв, тощо).

Випадковість відбору елементів у вибірку досягається дотриманням принципу рівної можливості всім елементам генеральної сукупності бути відібраними. Вибірка називається **репрезентативною**, якщо вона достатньо добре відтворює генеральну сукупність.

Розрізняють наступні **види вибірок**:

1. **просто випадкова вибірка:** об'єкти відбирають по одному зі всієї генеральної сукупності;
2. **механічна вибірка:** генеральна сукупність довільним чином ділиться на стільки груп , скільки об'єктів повинно входити у вибірку, і із кожної групи беруть один об'єкт;
3. **типова вибірка:** об'єкти відбирають не з усієї генеральної сукупності, а тільки з її частини, яку обирають за деякою ознакою;

4. **серійна вибірка:** об’єкти відбирають не по одному, а групами (серіями) зі всієї генеральної сукупності.

Використовують два способи утворення вибірки:

- **повторний відбір** (за схемою поверненої кулі), коли кожен елемент, випадково відібраний і досліджений, повертається у загальну сукупність і може бути повторно відібраним;
- **безповторний відбір** (за схемою неповерненої кулі) , коли відібраний елемент не повертається у загальну сукупність.

Математична теорія вибіркового методу базується на аналізі просто випадкової вибірки. Теоретичною основою застосування вибіркового методу є **закон великих чисел** і теорія ймовірностей, згідно з якими відмінності між аналогічними характеристиками генеральної та вибіркової сукупностей можна зменшити зі збільшенням обсягу вибірки. Вірогідна оцінка всієї досліджуваної сукупності за результатами вибіркового спостереження можлива лише за умов:

1) кількість відібраних одиниць для спостереження повинна бути досить великою;
2) відбір одиниць для вибіркового спостереження має бути таким, щоб кожна одиниця сукупності мала однакову можливість потрапити у вибірку.

Основний недолік вибіркового методу – похибки дослідження, чи похибки репрезентативності (представництва).

Встановлення статистичних закономірностей починається з відомостей про те, які значення прийняла в результаті спостережень ознака, що нас цікавить, яку називатимемо **«випадкова величина *X*».**

Позначимо:

x_i – значення ознаки (випадкової величини X);

N та n – об'єми генеральної та вибіркової сукупностей;

N_i та n_i – кількість елементів генеральної та вибіркової сукупностей зі значеннями ознаки x_i.

Середнє арифметичне розподілу ознаки в генеральній та вибірковій сукупностях називаються відповідно **генеральним** та **вибірковим середнім**, а дисперсії цих розподілів – **генеральною** та **вибірковою дисперсіями**.

Відношення кількості елементів генеральної та вибіркової сукупностей, які мають певну ознаку A, до їх обсягів, називаються відповідно **генеральною** і **вибірковою частками**.

Найважливішим завданням вибіркового методу є **оцінка параметрів** (характеристик) генеральної сукупності за даними вибірки.

1.2. Дискретний статистичний розподіл вибірки

Нехай із генеральної сукупності проведена вибірка $(x_1, x_2, ..., x_k)$. Елементи вибірки $(x_1, x_2, ..., x_k)$ називаються **варіантами.** Перший крок до опрацювання наявного статистичного матеріалу – це його впорядкування: розташування варіант в порядку зростання (спадання), тобто **ранжування** варіант ряду. Ранжований в порядку зростання (чи спадання) ряд варіант із відповідними їм вагами (частотами та частостями) називається **варіаційним рядом**.

Якщо варіанта x_1 спостерігалась n_1 разів, x_2 - n_2 разів,…, x_k - n_k разів

$\left(\sum_{i=1}^{k} n_i = n\right)$, то дискретний статистичний розподіл має вид:

$X = x_i$	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
$\omega_i = \frac{n_i}{n}$	ω_1	ω_2	…	ω_k

Для графічного зображення варіаційних рядів найчастіше використовують **полігон, гістограму, кумулятивну криву.**

Полігон використовують для зображення дискретного варіаційного ряду. Він є ламаною лінією, в якої кінці відрізків мають координати $(x_i, n_i),\ i = 1, 2, ..., m.$

Гістограма використовується тільки для зображення інтервальних варіаційних рядів і представляє собою ступінчасту фігуру із прямокутників з основами, які дорівнюють інтервалам значення ознаки $k_i = x_{i+1} - x_i,\ i = 1, 2, ..., m$, і висотами рівними частотам (частостям) $n_i (w_i)$ інтервалів. Якщо з'єднати середини верхніх основ прямокутників відрізками прямої то можна одержати полігон розподілу. Варіаційний ряд називатимемо **неперервним** (інтервальним), якщо варіанти можуть відрізнятись одна від одної на дуже малу величину. При вивченні варіаційних рядів використовують також поняття **накопиченої частоти** ($n_i^{нак}$). Накопичена частота показує, скільки спостерігалось варіант зі значенням ознаки меншим за x. Відношення накопиченої частоти $n_i^{нак}$ до загальної кількості спостережень n назвемо **накопиченою частістю** $w_i^{нак}$. **Кумулятивна крива (кумулята)** – крива накопичених частот (частостей).

Для дискретного ряду кумулята представляє ламану, з'єднану точками $(x_i, n_i^{нак})$ або $(x_i, w_i^{нак}), i = 1,2,...,m.$ Для інтервалів варіаційного ряду ламана починається з точки, абсциса якої дорівнює початку першого інтервалу, а ордината – накопиченій частоті (частості), яка дорівнює нулю. Інші точки цієї ламаної відповідають кінцям інтервалів.

1.3. Емпірична функція розподілу

Емпіричною функцією розподілу $F^*(x)$ називається відносна частота (частість) того, що ознака (випадкова величина *X*) прийме значення, менше заданого *x*, тобто: $F^*(x) = w(X < x) = n_x^{нак}$. Тобто, для даного *x* емпірична функція розподілу представляє накопичену частість $w_x^{нак} = n_x^{нак} / n$. Отже,

$$F^*(x) = \begin{cases} 0, & x \le x_1 \\ \dfrac{n_1}{n}, & x_1 < x \le x_2 \\ \dfrac{n_1 + n_2}{n}, & x_2 < x \le x_3 \\ \ldots \\ \sum\limits_{i=1}^{k-1} \dfrac{n_i}{n}, & x_{k-1} < x \le x_k \\ 1, & x > x_k \end{cases}. \qquad (1.1)$$

Емпірична функція розподілу відіграє фундаментальну роль у математичній статистиці.

На відміну від емпіричної функції розподілу вибірки $F^*(x)$ функцію розподілу $F(x)$ генеральної сукупності у математичній статистиці

називають **теоретичною** функцією розподілу.

Теорема 1.1. Нехай $F^*(x)$ - емпірична функція розподілу, яку побудовано за вибіркою $X=(x_1,x_2,...,x_n)$ і $F(x)$ - відповідна теоретична функція розподілу. Тоді для будь-якого $x\ (-\infty<x<\infty)$ і довільного $\varepsilon>0$

$$\lim_{n\to\infty} P\left(\left|F^*(x)-F(x)\right|<\varepsilon\right)=1.$$

Тому $F^*(x)$ доцільно використовувати для наближеного представлення функції розподілу $F(x)$ генеральної сукупності. Функція $F^*(x)$ володіє всіма властивостями функції $F(x)$.

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

1) $0\le F^*(x)\le 1.$

2) $F^*(x)$ - неспадна.

3) $F^*(x)$ - неперервна зліва.

1.4. Приклади

1. Побудувати полігон відносних частот за вибіркою, яку задано дискретним варіаційним рядом:

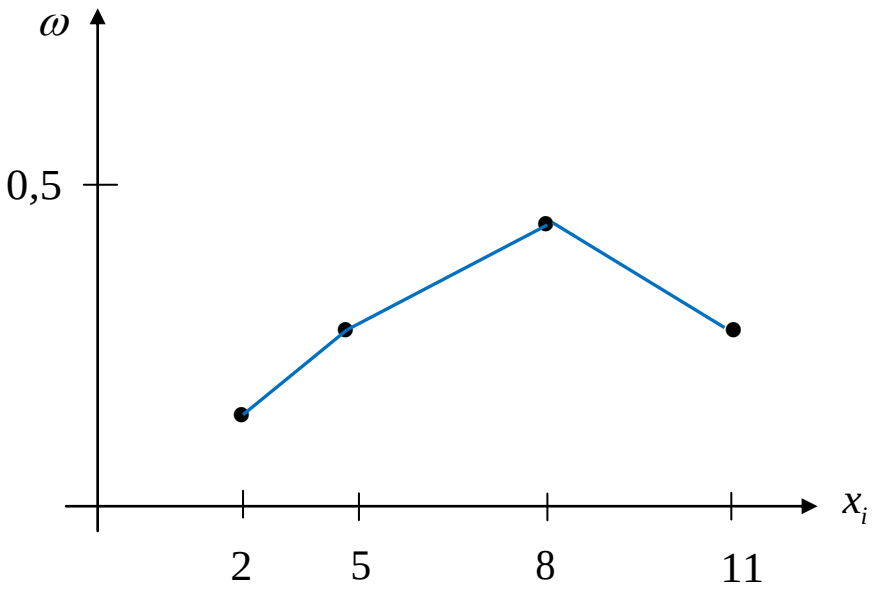
x_i	**2**	**5**	**8**	**11**
n_i	**4**	**8**	**10**	**8**

Розв'язання: для побудови полігону відносних частот потрібно знайти обсяг вибірки: $n=\sum_{i=1}^{4} n_i = 4+8+10+8=30$.

Знаходимо відносні частоти:

$$\omega_1 = \frac{n_1}{n} = \frac{4}{30} = \frac{2}{15};\ \omega_2 = \frac{n_2}{n} = \frac{8}{30} = \frac{4}{15};\ \omega_3 = \frac{n_3}{n} = \frac{10}{30} = \frac{1}{3};\ \omega_4 = \frac{n_4}{n} = \frac{8}{30} = \frac{4}{15}$$

В декартовій системі координат будуємо точки з координатами (x_i, ω_i) і з'єднуємо їх ламаною лінією:



2. У цеху працює 60 робітників. Для вивчення продуктивності праці випадковим чином обрали 10 робітників і підрахували кількість деталей, які виготовив кожен з цих робітників за зміну. Одержали вибірку: 17, 20, 17, 22, 24, 20, 20, 24, 22, 20. Знайти : варіаційний ряд, статистичний розподіл вибірки.

Розв'язання. Нехай випадкова величина X – це кількість деталей, які виготовляє один робітник за зміну. Обсяг вибірки дорівнює 10, а генеральна сукупність має 60 значень.

Варіаційний ряд: 17,17,20,20,20,20,22,22,24,24.

У вибірці всього 4 різних значення: 17,20,22 і 24.

Частоти цих значень: $n_1 = 2;\ n_2 = 4;\ n_3 = 2;\ n_4 = 2$.

Відносні частоти: $\omega_1 = 0,2;\ \omega_2 = 0,4;\ \omega_3 = 0,2;\ \omega_4 = 0,2$.

Статистичний розподіл вибірки:

x_i	**17**	**20**	**22**	**24**
n_i	**2**	**4**	**2**	**2**
ω_i	**0,2**	**0,4**	**0,2**	**0,2**

3. Побудувати статистичний розподіл частот і полігон частот для вибірки відхилень від цілі точки падіння снаряда в метрах: -10, 20,10, 20, -50, -20, -30, 40, -20, -30, -10, 10, 20, -40, 50, -10, 10,50, -50, 20.

Розв'язання. Обсяг вибірки дорівнює 20.

Варіаційний ряд: -50,-50, -40,-30, -20,-20, -10,-10,-10,10,10,10,20,20,20,20,30,40,50,50.

Маємо 10 різних значень: -50,-40,-30,-20,-10,10,20,30,40,50.

Частоти цих значень:

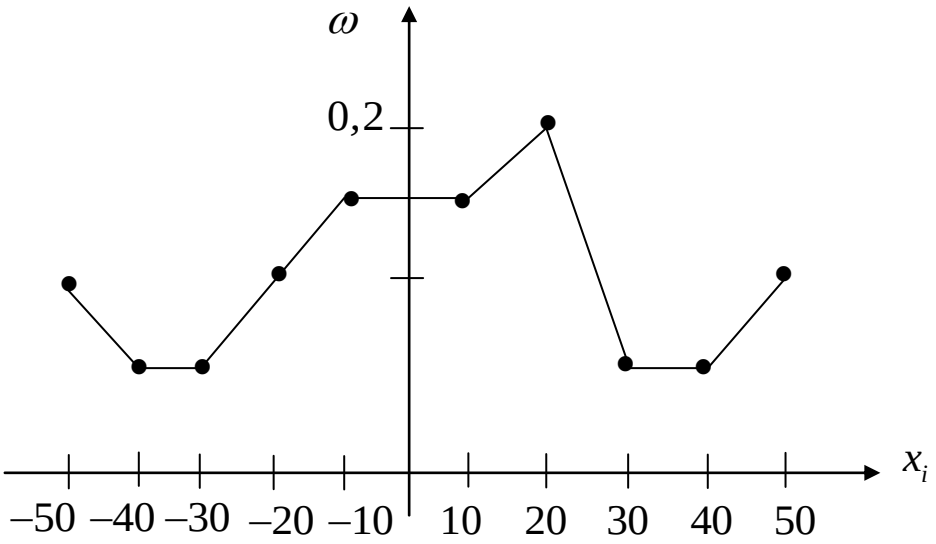
$\omega_1 = 0,1;\ \omega_2 = 0,05;\ \omega_3 = 0,05;\ \omega_4 = 0,1;\ \omega_5 = 0,15;$

$\omega_6 = 0,15;\ \omega_7 = 0,2;\ \omega_8 = 0,05;\ \omega_9 = 0,05;\ \omega_{10} = 0,1;$

Шуканий статистичний розподіл:

x_i	-50	-40	-30	-20	-10	10	20	30	40	50
n_i	2	1	1	2	3	3	4	1	1	2
ω_i	0,1	0,05	0,05	0,1	0,15	0,15	0,2	0,05	0,05	0,1

Полігон частот:



4. Знайти емпіричну функцію розподілу за статистичним розподілом вибірки:

x_i	**2**	**5**	**7**	**9**	**12**
n_i	**3**	**4**	**8**	**6**	**3**

Розв'язання.

Обсяг вибірки: 3+4+8+6+3=24.

Найменше значення дорівнює 2, отже $F^*(x)=0$ для всіх $x\le 2$.

Якщо значення $x<5$, то $F^*(x)=\frac{3}{24}=\frac{1}{8}$ для всіх $2<x\le 5$.

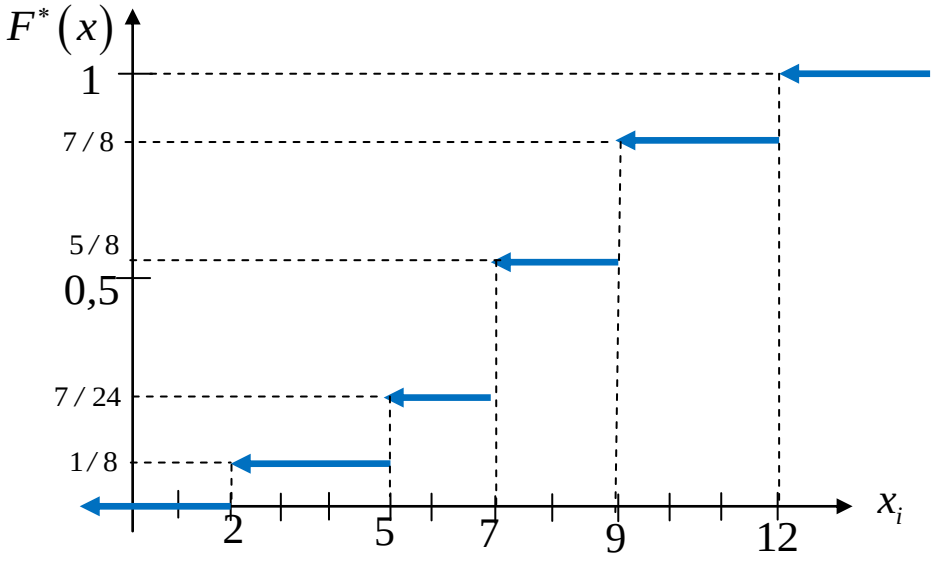
Аналогічно, для всіх $5<x\le 7$, $F^*(x)=\frac{3+4}{24}=\frac{7}{24}$.

Для $7<x\le 9$, $F^*(x)=\frac{3+4+8}{24}=\frac{15}{24}=\frac{5}{8}$.

Для $9<x\le 12$, $F^*(x)=\frac{3+4+8+6}{24}=\frac{21}{24}=\frac{7}{8}$.

Для всіх $x>12$, $F^*(x)=1$. Отже,

$$F^*(x)=\begin{cases}0, \ при \ \ x\le 2\\ \frac{1}{8}, \ при \ \ 2<x\le 5\\ \frac{7}{24}, при \ 5<x\le 7\\ \frac{5}{8}, при \ 7<x\le 9\\ \frac{7}{8}, \ при \ 9<x\le 12\\ 1, \ при \ x>12\end{cases}$$



5. Записати вибірку 4,3,6,5,7,4,5,5,6,4,5,5,6,5,3,5,6,4,5,6 у вигляді варіаційного та статистичного рядів, побудувати емпіричну функцію розподілу, кумуляту, гістограму та полігон частот вибірки.

Розв'язання.

Обсяг вибірки: $n=20$.

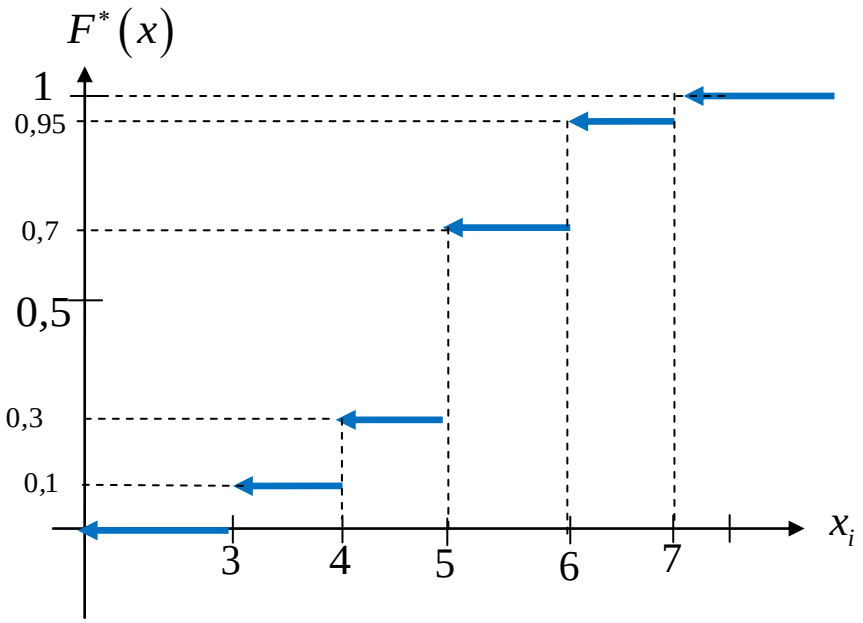
Варіаційний ряд: 3,3,4,4,4,4,5,5,5,5,5,5,5,5,6,6,6,6,6,7.

Статистичний розподіл:

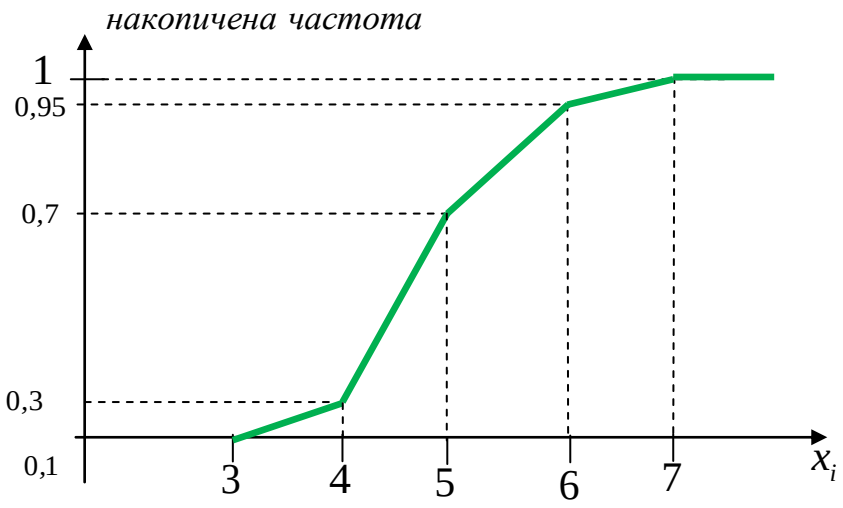
x_i	**3**	**4**	**5**	**6**	**7**
n_i	**2**	**4**	**8**	**5**	**1**
ω_i	**0.1**	**0,2**	**0,4**	**0,25**	**0,05**

Функція розподілу: $F^*(x) = \begin{cases} 0, \ при \ \ x \le 3 \\ 0.1, \ при \ \ 3 < x \le 4 \\ 0.3, \ при \ \ 4 < x \le 5 \\ 0.7, \ при \ \ 5 < x \le 6 \\ 0.95, \ при \ \ 6 < x \le 7 \\ 1, \ при \ \ x > 7 \end{cases}$.

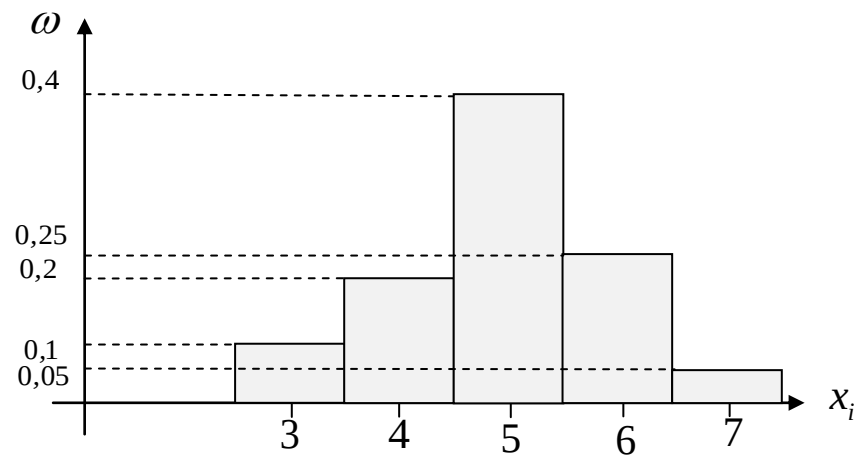
Графік функції розподілу:



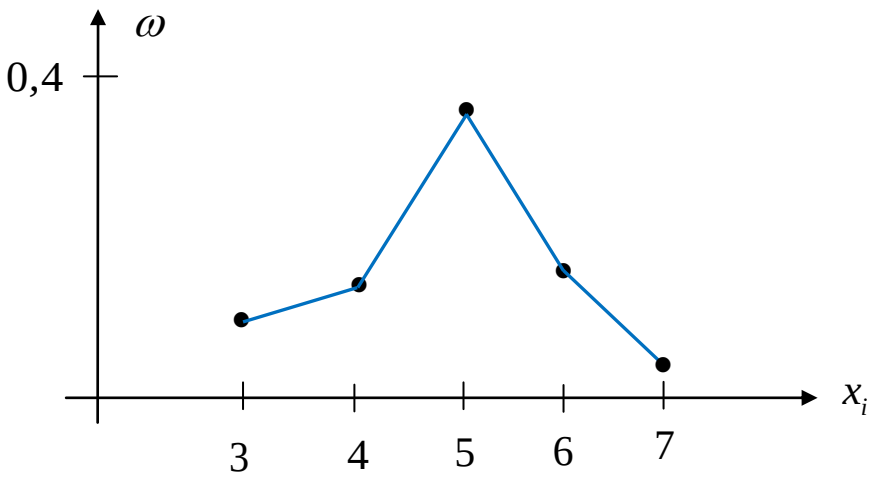
Кумулята:



Гістограма (побудова): область значень випадкової величини розбивають на рівні інтервали довжиною h . Потім підраховують суму частот тих значень випадкової величини, які належать кожному з одержаних відрізків. Потім будують прямокутник з основою h і висотою яка дорівнює відносній частоті варіанти. Для нашого прикладу гістограма має вид:



Полігон:



Числові характеристики вибірки.

1.5. Середні величини

Середні величини є одними із важливих характеристик варіаційних рядів. В математичній статистиці розрізняють декілька видів середніх величин: арифметичну, геометричну, гармонічну, квадратичну, кубічну та інші. Всі перераховані типи середніх можуть бути обчислені для випадку, коли кожна із варіант варіаційного ряду зустрічається тільки один раз (середня буде називатися **простою** або **незваженою**). Якщо значення варіант повторюються різне число разів, то обчисленні середні значення називають **зваженними**. Для характеристики варіаційного ряду один із перерахованих типів середніх вибирають не довільно, а в залежності від особливостей явища, що вивчається, і мети, для якої середнє обчислюється. Середні величини є значущими при дослідженні вибірки, якщо сама вибірка є однорідною.

Формули для обчислень всіх типів середніх величин можна одержати із формули для обчислення **степеневого середнього.** Якщо варіанти $x_1, x_2, ..., x_n$ зустрічаються у вибірці один раз або однакову кількість разів, то степенева середня обчислюється за формулою:

$$\overline{x}_{cm} = \sqrt[m]{\frac{\sum_{i=1}^{k} x_i^m}{n}}$$

$m-$ показник степеня,який визначає тип середньої.

Якщо варіанти $x_1, x_2, ..., x_n$ повторюються різне число разів, то степенева середня обчислюється за формулою зваженої степеневої середньої порядку m:

$$\overline{x}_{cm} = \sqrt[m]{\frac{\sum_{i=1}^{k} x_i^m n_i}{\sum_{i=1}^{k} n_i}}$$

де $n_i -$ частота варіанти x_i; $k -$ кількість варіант; $n = \sum_{i=1}^{k} n_i -$ обсяг вибірки.

При $m = -1$ отримаємо незважену і зважену **середні гармонічні**:

$$\overline{x}_{гар} = \frac{n}{\sum_{i=1}^{k} \frac{1}{x_i}} — \text{незважена;} \quad \overline{x}_{гар} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} — \text{зважена.}$$

Середня гармонічна обчислюється тоді, коли середня використовується для розрахунку сум доданків, які обернено пропорційні величині цієї ознаки, тобто коли знаходять суму не самих варіант, а обернених до них

величин $\frac{1}{x_1}, \frac{1}{x_2}, ..., \frac{1}{x_n}$.

Приклад 1.5. Зібрані данні про роботу 30 програмістів протягом 7 годин. Обчислити середню гармонічну зважену.

Витрати часу на виготовлення одиниці коду (хв), x_i	Кількість програмістів, n_i	$\frac{n_i}{x_i}$
20	4	0,2
18	6	$\frac{1}{3} \approx 0,333$
16	8	0,5
15	9	0,6
14	3	$\frac{3}{14} \approx 0,214$

Розв'язання. За формулою для зваженого гармонічного :

$$\overline{x}_{гар} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{30}{0,2+0,333+0,5+0,6+0,214} \approx 16,24 \blacktriangleright$$

Якщо $m = 1$, то одержимо незважену і зважену **середні арифметичні**:

$$\overline{x}_a = \frac{\sum_{i=1}^{k} x_i}{n} \text{—незважена;} \quad \overline{x}_a = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} \text{—зважена.}$$

де x_i - варіанти дискретного ряду або середини інтервалів інтервального варіаційного ряду; n_i - відповідні їм частоти; k - кількість варіант, які не

повторюються, або кількість інтервалів: $n = \sum_{i=1}^{m} n_i$.

Основні властивості середньої арифметичної, аналогічні властивостям математичного сподівання випадкової величини :

1. Середня арифметична сталої дорівнює самій сталій. .

2. Якщо усі варіанти збільшити (зменшити) в одне і те ж число раз, то середня арифметична збільшиться (зменшиться) у стільки ж разів:

$$\overline{kx} = k\bar{x},$$

або $$\frac{\sum_{i=1}^{m}(kx_i)n_i}{n} = k\frac{\sum_{i=1}^{m}(x_i)n_i}{n} .$$

3. Якщо усі варіанти збільшити (зменшити) на одне і те саме число, то середня арифметична збільшиться (зменшиться) на це число:

$$\overline{x+c} = \bar{x} + c \quad \text{або} \quad \frac{\sum_{i=1}^{m}(x_i + c)n_i}{n} = \frac{\sum_{i=1}^{m} x_i n_i}{n} + c.$$

4. Середня арифметична відхилень варіантів від середньої арифметичної дорівнює нулю: $\overline{x - \bar{x}} = 0$ або $\sum_{i=1}^{m}(x_i - \bar{x})n_i = 0.$

5. Середня арифметична алгебраїчної суми декількох ознак дорівнює такій самій сумі середніх арифметичних цих ознак :

$$\overline{x+y} = \bar{x} + \bar{y}.$$

6. Якщо ряд складається з декількох груп, загальна середня дорівнює середньому арифметичному групових середніх, причому вагами є обсяги груп:

$$\bar{x} = \frac{\sum_{i=1}^{m} \bar{x}_i n_i}{n}.$$

Приклад 1.6. При дослідженні виробництва мікрочипів, були одержані наступні данні ваги 20 чипів (г): 2,5; 3; 3,2; 2,5; 3; 3; 3,2; 3,3; 2,7; 2,9; 2,8; 3,1; 3,6; 3; 3,4; 3; 3,6; 3,1; 2,8; 2,3. Побудувати варіаційний ряд, статистичний розподіл відносних частот, накреслити полігон відносних частот, знайти вибіркову середню.

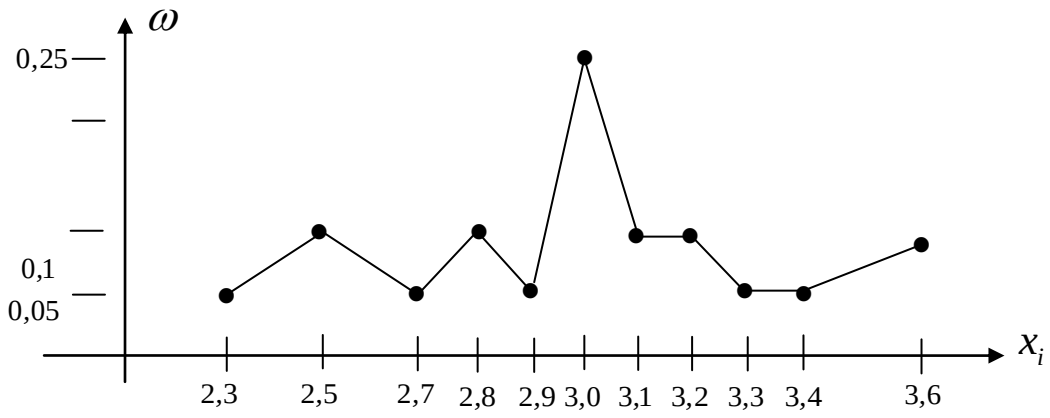
Розв'язання.

Варіаційний ряд: 2,3;2,5; 2,5; 2,7; 2,8; 2,8; 2,9; 3; 3; 3; 3; 3; 3,1; 3,1; 3,2; 3,2; 3,3; 3,4; 3,6; 3,6. Маємо 11 різних значень.

Статистичний розподіл:

x_i	2,3	2,5	2,7	2,8	2,9	3,0	3,1	3,2	3,3	3,4	3,6
n_i	1	2	1	2	1	5	2	2	1	1	2
ω_i	0,05	0,1	0,05	0,1	0,05	0,25	0,1	0,1	0,05	0,05	0,1

Полігон:



Вибіркова середня арифметична зважена:

$$\overline{x}_a = \frac{2{,}3 + 2{,}5 \cdot 2 + 2{,}7 + 2{,}8 \cdot 2 + 2{,}9 + 3{,}0 \cdot 5 + 3{,}1 \cdot 2 + 3{,}2 \cdot 2 + 3{,}3 + 3{,}4 + 3{,}6 \cdot 2}{20} = 3{,}0$$

Відповідь: при виробництві найчастіше зустрічаються чіпи вагою 3 гр.

Групова і загальна середня

Нехай всі значення кількісної ознаки X, генеральної чи вибіркової, розділені на декілька груп.

Груповою середньою називають середню арифметичну значень ознаки даної групи.

Загальною середньою $\overline{x}$ називають середню арифметичну значень ознаки, які належать всій сукупності. Загальна середня дорівнює середньому арифметичному групових середніх.

Приклад 1.7. Знайти загальну середню сукупності, яка складається із заданих груп.

x_{1i}	2	5	8	x_{2i}	4	7	9	x_{3i}	3	6	10
n_{1i}	6	14	10	n_{2i}	6	9	5	n_{3i}	7	12	6

Розв'язання.

Групові середні:

$$\overline{x}_1 = \frac{2 \cdot 6 + 5 \cdot 14 + 8 \cdot 10}{6 + 14 + 10} = \frac{162}{30} = 5{,}4;$$

$$\overline{x}_2 = \frac{4 \cdot 6 + 7 \cdot 9 + 9 \cdot 5}{6 + 9 + 5} = \frac{132}{20} = 6{,}6;$$

$$\bar{x}_3 = \frac{3\cdot 7 + 6\cdot 12 + 10\cdot 6}{7+12+6} = \frac{153}{25} = 6{,}12.$$

Загальна середня: $\bar{x} = \frac{5{,}4\cdot 30 + 6{,}6\cdot 20 + 6{,}12\cdot 25}{30+20+25} = \frac{447}{75} \approx 5{,}96$ ►

Приклад 1.8. Нехай ноутбуки від трьох виробників розподілені за вартістю наступним чином:

Вартість ноутбука, грн	Кількість ноутбуків			
	Виробник ноутбуків			Всього
	№1	№2	№3	
7500	7	1	0	8
8500	12	5	0	17
9500	15	9	4	28
10500	6	18	8	32
11500	0	12	32	44
12500	0	5	16	21
Всього	40	50	60	150

Обчислити середню вартість ноутбука

- по кожному виробнику;
- по всьому торговельному осередку.

Розв'язання.

1) Групові середні вартості:

$$\bar{x}_1 = \frac{7500\cdot 7 + 8500\cdot 12 + 9500\cdot 15 + 10500\cdot 6}{40} = 9000{,}0\,(грн)$$

$$\bar{x}_2 = \frac{7500\cdot 1 + 8500\cdot 5 + 9500\cdot 9 + 10500\cdot 18 + 11500\cdot 12 + 12500\cdot 5}{50} =$$

$$= 10500{,}0\,(грн)$$

$$\overline{x}_3 = \frac{9500 \cdot 4 + 10500 \cdot 8 + 11500 \cdot 32 + 12500 \cdot 16}{60} = 11500{,}0\left(грн\right)$$

2) Загальна середня:

$$\overline{x} = \frac{9000 \cdot 40 + 10500 \cdot 50 + 11500 \cdot 60}{150} = 10500\left(грн\right) \blacktriangleright$$

Якщо у формулу $\overline{x}_{ст} = \sqrt[m]{\frac{\sum_{i=1}^{n} x_i^m}{n}}$ підставити $m = 2$, то отримаємо **незважену середню квадратичну**: $\overline{x}_{кв} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$.

Якщо у формулу $\overline{x}_{ст} = \sqrt[m]{\frac{\sum_{i=1}^{k} x_i^m n_i}{\sum_{i=1}^{k} n_i}}$ підставити $m = 2$, то отримаємо **зважену середню квадратичну**: $\overline{x}_{кв} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i}}$.

Середня квадратична використовується для розрахунку тільки тоді, коли варіанти є відхиленнями фактичних величин від їх середніх арифметичних або від норми.

Якщо ж у ці формули підставити $m = 0$, то одержимо незважену та зважену середні геометричні:

- $\overline{x}_{геом} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ - незважена середня геометрична;

- $\overline{x}_{геом} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot ... \cdot x_k^{n_k}}$ - зважена середня геометрична;

Середня геометрична використовується переважним чином для вивчення динаміки. Середні коефіцієнти і темпи росту розраховуються за формулами середньої геометричної.

Середню кубічну використовують для узагальнення ознак, виражених лінійними розмірами об'ємних фігур.

Приклад 1.9. Нехай діаметри плодів яблуні дорівнюють 4,6,6,7,7 см. Обчислити середній діаметр плодів.

Розв'язання.

$$\overline{x}_{куб} = \sqrt[3]{\frac{\sum_{i=1}^{n} x_i^3}{n}} = \sqrt[3]{\frac{4^3 + 6^3 + 6^3 + 7^3 + 7^3}{5}} = 6{,}18(см).$$

За формулою середньої арифметичної:

$$\overline{x}_a = \frac{4+6+6+7+7}{5} = 6(см).$$

Обчислимо середній об'єм плода в обох випадках:

$$V_{куб} = \frac{1}{6}\pi(6{,}18)^3 = 123{,}52(см^3);$$

$$V_a = \frac{1}{6}\pi(6)^3 = 113{,}04(см^3).$$

Знайдемо об'єм всіх п'яти плодів:

$V = \frac{1}{6}\pi(4^3 + 6^3 + 6^3 + 7^3 + 7^3) = 618{,}58(см^3)$, тоді середній об'єм становитиме $V_{сер} = \frac{618{,}58}{6} = 123{,}72(см^3)$. З обчислень можна зробити

висновок, що середня кубічна дає більш правильну характеристику ознаки.►

Окрім розглянутих середніх величин, які називаються **аналітичними**, в статистичному аналізі застосовують структурні, або порядкові, середні.
З них найчастіше застосовуються **медіана і мода**.

Медіаною $\tilde{M}e$ варіаційного ряду називається значення ознаки, яке припадає на середину ранжируваного ряду спостережень.

Для дискретного варіаційного ряду з непарною кількістю членів медіана дорівнює центральній варіанті, а для ряду з парним - півсумі двох центральних варіант.

Для інтервального варіаційного ряду знаходиться медіанний інтервал, на який припадає середина ряду, а значення медіани на цьому інтервалі знаходять за допомогою лінійної інтерполяції. Зауважимо, що медіана може бути приблизно знайдена за допомогою кумуляти як значення ознаки, для якої $w_x^{нак} = 1/2.$

Придатність медіани як міри центральної тенденції полягає в тому, що на неї не впливає зміна крайніх членів варіаційного ряду, якщо будь-який із них менший медіани, залишається меншим від неї, а будь-який більший медіани, продовжує бути більшим від неї. Медіану краще застосовувати (ніж середнє арифметичне) для ряду, у якого крайні варіанти в порівнянні з іншими виявилися надмірно великими або малими.

Медіана має таку **властивість**: сума абсолютних величин відхилень елементів вибірки від медіани менша, ніж від будь-якої іншої величини:

$$\sum_{i=1}^{n} \left| x_i - \tilde{M}e \right| < \sum_{i=1}^{n} \left| x_i - a \right|; \ a \neq \tilde{M}e .$$

Цю властивість медіани можна використати при проектуванні розміщення зупинок громадського транспорта, заправок тощо.

Приклад 1.10. На одному з відрізків залізничної гілки планується зробити зупинку пасажирського потяга. Розподіл найближчих населених пунктів із чисельністю їх населення наведено в таблиці:

На якому кілометрі залізниці розташовано населенний пункт, км	8	10	13	17	24	26	30	34
Чисельність населення, тис. чол	3	4	2	7	10	1	3	7

На якому кілометрі залізниці потрібно розташувати зупинку, щоб сумарна відстань, яку долатимуть потенційні пасажири, до цієї зупинки, була найменшою.

Розв'язання. Використаємо основну властивість медіани. Знайдемо медіану заданої вибірки:

варіаційний ряд:

$$8,8,8,10,10,10,10,13,13,\underbrace{17,...,17}_{7\text{ разів}},\underbrace{24,...,24}_{10\text{ разів}},26,30,30,30,\underbrace{34,...,34}_{7\text{ разів}}$$

$x_{19} = 24$ є медіаною цього варіаційного ряду. Отже, зупинку потрібно зробити на 24-ому кілометрі.►

Модою $\tilde{M}o$ **варіаційного ряду** називається варіанта, якій відповідає найбільша частота.

Для інтервального ряду знаходиться модальний інтервал, який має найбільшу частоту, а значення моди на цьому інтервалі визначають за допомогою лінійної інтерполяції. Проте, простіше моду можна знайти

графічним шляхом за допомогою гістограми. Особливість моди як міри центральної тенденції полягає в тому, що вона не змінюється при зміні крайніх членів ряду, тобто має певну стійкість до варіації.

1.6. Показники варіації

Середні величини, розглянуті вище, не відображають мінливості (варіації) значень ознаки. Найпростішим (і дуже наближеним) показником варіації є варіаційний розмах *R*, який дорівнює різниці між найбільшою і найменшою варіантами ряду: $R = x_{max} - x_{min}$.

Відхиленням називають різницю $x_i - \overline{x}$ між значенням ознаки і загальною середньою.

Теорема 1.2. Сума добутків відхилень на відповідні частоти дорівнює нулю: $\sum_i n_i \left(x_i - \overline{x} \right) = 0$.

Доведення: $\sum_i n_i \left(x_i - \overline{x} \right) = \sum_i n_i x_i - \sum_i n_i \overline{x} = \overline{x} n - n\overline{x} = 0$ ■

Наслідок. Середнє значення відхилення дорівнює нулю.

Середнім лінійним відхиленням варіаційного ряду називається середнє арифметичне абсолютних величин відхилень варіант від загального середнього арифметичного:

$$d = \frac{\sum_{i=1}^{m} \left| x_i - \overline{x} \right| n_i}{n}$$

Генеральною дисперсією $D_г$ називають середнє арифметичне квадратів відхилень значень ознаки генеральної сукупності від їх середнього значення $\overline{x}_г$.

Якщо всі значення $x_1, x_2, ..., x_n$ ознаки генеральної сукупності обсягу N різні, то $D_г = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}_г\right)^2$. Якщо всі значення ознаки мють відповідні частоти $N_1, N_2, ..., N_k$, причому $N_1 + N_2, +.. + N_k = N$, то $D_г = \frac{1}{N}\sum_{i=1}^{k} N_i \left(x_i - \overline{x}_г\right)^2$.

Генеральним середнім квадратичним відхиленням (стандартом) називають корінь квадратний з генеральної дисперсії: $\sigma_г = \sqrt{D_г}$.

Дисперсією s^2 варіаційного ряду (вибірковою дисперсією) називається середнє арифметичне квадратів відхилень варіанти від її середньої арифметичної:

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2 n_i}{n} = \sum_{i=1}^{m}(x_i - \overline{x})^2 w_i$$

Для незгрупованого ряду ($n_i = 1$) з формули маємо:

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{n}.$$

Дисперсію s^2 називають **емпіричною** або **вибірковою**, підкреслюючи, що вона (на відміну від дисперсії випадкової величини σ^2) знаходиться за дослідними або статистичними даними.

Бажано в якості міри варіації (розсіювання) мати характеристику, яка виражається в тих самих одиницях, що і значення ознаки. Такою характеристикою є **середнє квадратичне відхилення** s **—** арифметичне значення кореня квадратного з дисперсії

$$s = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 n_i}{n}}.$$

В деяких випадках використовують таку характеристику, як **коефіцієнт варіації**, який дорівнює відсотковому відношенню середнього квадратичного відхилення до середнього арифметичного:

$$\tilde{v} = \frac{s}{\bar{x}} \cdot 100\% \ (\bar{x} \neq 0)$$

Якщо коефіцієнт варіації ознаки, набуває тільки позитивних значень, є високим (наприклад, більше 100%), то, як правило, це свідчить про неоднорідність значень ознаки.

Груповою дисперсією називають дисперсію значень ознаки, що належить групі відносно групової середньої:

$$D_{гр} = \frac{1}{N_j}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}_{гр}\right)^2.$$

Якщо відома дисперсія кожної групи, то можна знайти їх середню арифметичну.

Внутрішньогруповою дисперсією називають середню арифметичну дисперсійну зважену за обсягами груп:

$$D_{внгр} = \frac{1}{n}\sum_{i} N_j D_{jгр},$$

де N_j - обсяг групи, $n = \sum_j N_j$ - обсяг всієї сукупності.

Міжгруповою дисперсією називають дисперсію групових середніх відносно загальної середньої:

$$D_{міжгр} = \frac{1}{n}\sum_j N_j \left(\overline{x}_j - \overline{x}\right)^2 .$$

Загальною дисперсією називають дисперсію значень ознаки всієї сукупності відносно загальної середньої:

$$D_{заг} = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 .$$

Теорема 1.3. Якщо сукупність складається із декількох груп, то загальна дисперсія дорівнює сумі внутрішньогрупової і міжгрупової дисперсій:

$$D_з = D_{внгр} + D_{міжгр}.$$

Основні властивості дисперсії, аналогічні властивостям дисперсії випадкової величини:

1. Дисперсія сталої дорівнює нулю.
2. Якщо усі варіанти збільшити (зменшити) в k разів, то дисперсія збільшиться (зменшиться) в k^2 разів:

$$s_{kx}^2 = k^2 s_x^2 \text{ або } \frac{\sum_{i=1}^{m}\left(x_i k - \overline{x}k\right)^2 n_i}{n} = k^2 \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}\right)^2 n_i}{n}.$$

3. Якщо всі варіанти збільшити (зменшити) на одне і те саме число, то дисперсія не зміниться:

$$s_{x+c}^2 = s_x^2 \text{ або } \frac{\sum_{i=1}^{m}\left(\left(x_i + c\right) - \left(\overline{x} + c\right)\right)^2 n_i}{n} = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x}\right)^2 n_i}{n}$$

4. Дисперсія дорівнює різниці між середнім арифметичним квадратів варіант і квадратом середнього арифметичного:

$$s^2 = \overline{x^2} - \overline{x}^2$$

5. Якщо ряд складається з декількох груп спостережень, то загальна дисперсія дорівнює сумі середнього арифметичного групових дисперсій і міжгрупової дисперсії (правило додавання дисперсій):

$$s^2 = \overline{s_i^2} + \delta^2$$

де s^2 — загальна дисперсія (дисперсія всього ряду);

$\overline{s_i^2} = \dfrac{\sum_{i=1}^{l} s_i^2 n_i}{n}$ — середнє арифметичне групових дисперсій,

де $s_i^2 = \dfrac{\sum_{j=1}^{m} (x_j - \overline{x}_i)^2 n_j}{n_i}$; $\delta^2 = \dfrac{\sum_{i=1}^{l} (\overline{x_i} - \overline{x})^2 n_i}{n}$ — міжгрупова дисперсія.

Приклад 1.12. Знайти внутрішньогрупову, міжгрупову і загальну дисперсії для груп прикладу 1.7.

x_{1i}	2	5	8	x_{2i}	4	7	9	x_{3i}	3	6	10
n_{1i}	6	14	10	n_{2i}	6	9	5	n_{3i}	7	12	6

Розв'язання. Групові середні, які було знайдено в прикладі 2.3:

$\overline{x}_1 = 5{,}4;\ \overline{x}_2 = 6{,}6;\ \overline{x}_3 = 6{,}12.$

Знайдемо середні значення квадратів ознак:

$$\overline{x}_1^2 = \frac{2^2 \cdot 6 + 5^2 \cdot 14 + 8^2 \cdot 10}{30} = \frac{1041}{30} = 33{,}8;$$

$$\overline{x}_2^2 = \frac{4^2 \cdot 6 + 7^2 \cdot 9 + 9^2 \cdot 5}{20} = \frac{942}{20} = 47{,}1;$$

$$\overline{x}_3^2 = \frac{3^2 \cdot 7 + 6^2 \cdot 12 + 10^2 \cdot 6}{25} = \frac{1095}{25} = 43{,}8.$$

За формулою $D_{гр} = \frac{1}{N_j} \sum_{i=1}^{k} n_i \left(x_i - \overline{x}_{гр} \right)^2$ обчислюємо групові дисперсії:

$$D_{1гр} = \overline{x}_1^2 - \left(\overline{x}_1 \right)^2 = 33{,}8 - \left(5{,}4 \right)^2 = 4{,}64;$$

$$D_{2гр} = \overline{x}_2^2 - \left(\overline{x}_2 \right)^2 = 47{,}1 - \left(6{,}6 \right)^2 = 47{,}1 - 43{,}56 = 3{,}54;$$

$$D_{3гр} = \overline{x}_3^2 - \left(\overline{x}_3 \right)^2 = 43{,}8 - \left(6{,}12 \right)^2 = 43{,}8 - 37{,}45 = 6{,}35.$$

Внутрішньгрупова дисперсія: за формулою $D_{внгр} = \frac{1}{n} \sum_i N_j D_{jгр}$:

$$D_{внгр} = \frac{4{,}64 \cdot 30 + 3{,}54 \cdot 20 + 6{,}35 \cdot 25}{30 + 20 + 25} = 4{,}92.$$

Міжгрупова дисперсія: за формулою $D_{міжгр} = \frac{1}{n} \sum_j N_j \left(\overline{x}_j - \overline{x} \right)^2$:

загальна середня з прикладу 2.3 $\overline{x} = 5{,}96$, тоді

$$D_{міжгр} = \frac{30 \cdot \left(5{,}4 - 5{,}96 \right)^2 + 20 \left(6{,}6 - 5{,}96 \right)^2 + 25 \left(6{,}12 - 5{,}96 \right)^2}{75} = 0{,}24.$$

Загальна дисперсія: за формулою $D_{заг} = \frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \overline{x} \right)^2$;

$$D_{заг} = \frac{6 \left(2 - 5{,}96 \right)^2 + 14 \left(5 - 5{,}96 \right)^2 + 10 \left(8 - 5{,}96 \right)^2 + ... + 6 \left(10 - 5{,}96 \right)^2}{75} =$$

$$= 5{,}16$$

Або за формулою $D_з = D_{внгр} + D_{міжгр}$: $D_з = 4{,}92 + 0{,}24 = 5{,}16$.►

Приклад 1.13. Задано вибіркові данні для однієї і тієї ж ознаки в трьох серіях спостережень. Об'єднати групи в одну вибірку, знайти розмах вибірки, моду, медіану і коефіцієнт варіації.

x_{1i}	2	4	8	x_{2i}	4	8	9	x_{3i}	2	4	9
n_{1i}	6	14	10	n_{2i}	6	9	5	n_{3i}	7	12	6

Розв'язання. Всього обсяг вибірки

$n = 6+14+10+6+9+5+7+12+6 = 75$.

Статистичний розподіл:

x_i	2	4	8	9
n_i	13	32	19	11

Розмах вибірки: $R = 9-2 = 7$.

Медіана:

$$2,\underbrace{..,2}_{13\text{ разів}},4,\underbrace{..4}_{32\text{ рази}},8,\underbrace{..,8}_{19\text{ разів}},9,9,9,9,9$$

оскільки вибірка містить парну кількість варіант, то $\tilde{Me} = x_{38} = 4$.

Мода: найчастіше у вибірці зустрічається значення 4. Отже, $\tilde{Mo} = 4.$

Для знаходжння коефіцієнта варіації $\tilde{\nu} = \frac{s}{\overline{x}} \cdot 100\% \ (\overline{x} \neq 0)$ потрібно знайти арифметичну середню та середнє квадратичне відхилення вибірки.

Середня арифметична зважена: $\overline{x} = \frac{2 \cdot 13 + 4 \cdot 32 + 8 \cdot 19 + 9 \cdot 6}{75} = 4{,}8.$

Дисперсія:

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{n} = \frac{(2-4,8)^2 + (4-4,8)^2 + (8-4,8)^2 + (9-4,8)^2}{75} =$$
$$= 0,4848$$

Середнє квадратичне відхилення: $s = \sqrt{0,4848} = 0,696$.

Коефіцієнт варіації: $\tilde{v} = \frac{0,696}{4,8} \cdot 100\% = 14,5\%$. Такий показник свідчить про те, що наша вибірка однорідна і репрезентативна.►

1.7. Початкові і центральні моменти варіаційного ряду

Середня арифметична і дисперсія варіаційного ряду є частковими випадками більш загального поняття – моментів варіаційного ряду.

Початковий момент $\tilde{v}_k$ k**-го порядку** варіаційного ряду визначаються за формулою:

$$\tilde{v}_k = \frac{\sum_{i=1}^{m} x_i^k n_i}{n}$$

Очевидно, що $\tilde{v}_1 = \bar{x}$, тобто середнє арифметичне є моментом першого порядку варіаційного ряду.

Центральний момент μ_k k**-го порядку** варіаційного ряду визначається за формулою:

$$\mu_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k n_i}{n}$$

За допомогою моментів розподілу можна не тільки описати центральну тенденцію, розсіювання, але й інші особливості варіаційної ознаки. Очевидно, що використовуючи властивість середньої арифметичної, $\tilde{\mu}_1 = 0,\ \tilde{\mu}_2 = s^2$, тобто центральний момент першого порядку для будь-якого розподілу дорівнює 0, а другий момент є дисперсією варіаційного ряду.

Коефіцієнтом асиметрії варіаційного ряду називається число

$$\tilde{A} = \frac{\tilde{\mu}_3}{s^3} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^3 n_i}{ns^3}.$$

Якщо $\tilde{A} = 0$, тоді розподіл має симетричну форму. При $\tilde{A} > 0$ $(\tilde{A} < 0)$ кажуть про позитивну (правосторонню) або негативну (лівосторонню) асиметрію.

Ексцесом варіаційного ряду називається число

$$\tilde{E} = \frac{\tilde{\mu}_4}{s^4} - 3 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^4 n_i}{ns^4} - 3.$$

Ексцес є показником «крутизни» варіаційного ряду в порівнянні із нормальним розподілом. Якщо $\tilde{E} > 0$ $(\tilde{E} < 0)$, то полігон варіаційного ряду має більш круту (пологу) вершину в порівняння з нормальною кривою.

Контрольні запитання

1. Що називається варіаційним рядом?
2. Як складається емпірична функція розподілу?
3. Що таке «кумулята»?
4. Як побудувати гістограму?
5. Побудова графіка функції розподілу, особливості.
6. Означення середнього арифметичного варіаційного ряду.
7. Чим відрізняються такі характеристики варіаційного ряду як мода і медіана?
8. Яка характеристика варіаційного ряду є аналогом математичного сподівання випадкової величини?
9. Назвіть області застосування методів математичної статистики в ІТ.