Заняття №10

Методи статистичного опису результатів спостережень

Означення 1. Сукупність спостерігаємих випадкових величин $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ будемо називати **вибіркою**. Реалізацію вибірки ξ позначимо через $x' = (x_1, x_2, ..., x_n)$. Якщо всі елементи вибірки $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ є незалежними і однаково розподілені, як деяка випадкова величина ξ_0, то таку вибірку називають вибіркою з **генеральної сукупності** з розподілом $F_{\xi_0}(\cdot)$.

Означення 2. Кількість спостережень над випадковою величиною n називається **об'ємом вибірки**.

Означення 3. Нехай $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ - вибірка з генеральної сукупності з розподілом $F_{\xi_0}(\cdot)$ і $x' = (x_1, x_2, ..., x_n)$ - значення ξ, що спостерігались. Кожній реалізації x вибірки ξ можна поставити у відповідність упорядковану послідовність

$$x_{(1)} \le x_{(2)} \le ... \le x_{(n)},$$

де $x_{(1)} = \min(x_1, x_2, ..., x_n)$, $x_{(2)}$ - друге за величиною значення з $x_1, x_2, ..., x_n$ і т.д.; $x_{(n)} = \max(x_1, x_2, ..., x_n)$. Позначимо через $\xi_{(k)}$ випадкову величину, яка для кожної реалізації вибірки ξ набуває значення $x_{(k)}$, $k = 1, 2, ..., n.$

За вибіркою ξ визначимо нову послідовність випадкових величин $\xi_{(1)}, \xi_{(2)}, ..., \xi_{(n)}$, які називаються **порядковими статистиками** вибірки. При цьому $\xi_{(k)}$ - k-та порядкова статистика, а $\xi_{(1)}, \xi_{(n)}$ - мінімальна та максимальне значення вибірки відповідно.

З визначення порядкових статистик випливає, що вони задовольняють нерівності

$$\xi_{(1)} \le \xi_{(2)} \le ... \le \xi_{(n)}. \quad (1)$$

Послідовність (1) називається **варіаційним рядом** вибірки.

Зазначимо, що різниця між найбільшим та найменшим елементами реалізації вибірки називається **розмахом вибірки**

$$W = x_{(n)} - x_{(1)}.$$

Означення 4. Нехай у вибірці об'єму n елемент x_i зустрічається m_i разів, тоді число m_i називається **частотою** елемента x_i. Сукупність пар $(x_i, m_i), i = 1, 2, ..., s$ називається **статистичним рядом** і записується у вигляді таблиці

x_1	x_2	....	x_s
m_1	m_2	....	m_s

Зазначимо, що s - це кількість різних (без повторень) елементів вибірки. При цьому очевидними є такі співвідношення

1) $s \le n$,

2) $m_1 + m_2 + ... + m_s = n$.

Приклад 1.
Записати у вигляді варіаційного і статистичного рядів наступну вибірку:
5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4. Вказати значення для n та s.

1) Варіаційний ряд вибірки

$$2 \le 2 \le 2 < 3 < 4 \le 4 < 5 \le 5 \le 5 < 7 \le 7 \le 7 \le 7 < 10 \le 10.$$

2) Статистичний ряд

x_i	2	3	4	5	7	10
m_i	3	1	2	3	4	2

3) $n = 15, \quad s = 6.$

Групування даних

Якщо дані, які ми спостерігаємо відповідають випадковій величині абсолютно неперервного типу, то для подальшої їх обробки може знадобитися групована вибірка. Запишемо у вигляді алгоритму побудову групованої вибірки.

1. Перший крок.

Визначимо кількість інтервалів групування. Для цього застосуємо формулу Стерджесса:

$$s = 1 + [\log_2 n].$$

На практиці цю формулу зручно використовувати у такому вигляді

$$s = 1 + [3,322 \cdot \lg n].$$

2. Другий крок.

Визначення довжини інтервалу групування:

$$h = \frac{W}{s-1} = \frac{x_{(n)} - x_{(1)}}{s-1}.$$

3. Третій крок.

За початок першого інтервалу доцільно взяти таке число

$$y_0 = x_{(1)} - \frac{h}{2},$$

Кінець першого інтервалу та початок другого інтервалу визначається за такою формулою

$$y_1 = y_0 + h.$$

Продовжуючи далі, будемо мати:

$$y_2 = y_1 + h,$$

.....

$$y_s = y_{s-1} + h.$$

У результаті роботи алгоритму, одержимо таблицю

Інтервал	$[y_0, y_1)$	$[y_1, y_2)$	.....	$[y_{s-1}, y_s)$
Частота m_i	m_1	m_2	.....	m_s

Частота $m_i, i = 1,2,...,s$ визначається, як кількість елементів початкової вибірки, що потрапляють в i-тий інтервал.

Зазначимо, що даний алгоритм носить рекомендаційний характер і відповідно існують інші підходи до групування даних.

Важливою є задача переходу від групованої вибірки до групованого статистичного ряду. При цьому під групованим статистичним рядом будемо розуміти сукупність пар $(y_i^*, m_i), i = 1,2,...,s,$ де y_i^*, m_i - відповідно середина та частота i-го інтервалу. В результаті одержимо таблицю

y_1^*	y_2^*	....	y_s^*
m_1	m_2	....	m_s

З цією таблицею можна працювати як зі звичайним статистичним рядом.

Приклад 2.

У таблиці задано вибірку, яка відповідає доходам мешканців містечка. Потрібно за згаданим вище алгоритмом згрупувати ці дані.

№	Дохід	№	Дохід
1	3820	13	6660
2	9470	14	5490
3	3490	15	5980
4	7790	16	6250
5	4210	17	8390
6	3870	18	3630
7	4490	19	6090
8	9620	20	10450
9	6200	21	6800
10	6350	22	6470
11	7430	23	9160
12	7670	24	5110

Розв'язання.

Визначаємо кількість інтервалів групування

$$s = 1 + [3,322 \lg 24] = 5.$$

Визначаємо довжину інтервалу групування

$$h = \frac{x_{(n)} - x_{(1)}}{s-1} = \frac{10450 - 3490}{4} = 1740, \Rightarrow \frac{h}{2} = 870.$$

Інтервал	[2620, 4360)	[4360, 6100)	[6100, 7840)	[7840, 9580)	[9580,11320)
m_i	5	5	9	3	2

Емпірична функція розподілу

Визначимо для кожного дійсного z випадкову величину $\mu_n(z)$, яка дорівнює кількості елементів вибірки $\xi' = (\xi_1, \xi_2, ..., \xi_n)$, значення яких не перевищують z:

$$\mu_n(z) = \left| \left\{ j : \xi_j \le z \right\} \right|,$$

де $|\cdot|$ - кількість елементів скінченної множини. Функція, яка задається рівністю $F_n(z) = \dfrac{\mu_n(z)}{n}$, називається **емпіричною функцією розподілу**. Функцію розподілу $F_{\xi_0}(z)$ випадкової величини ξ_0, що спостерігається, називається **теоретичною функцією розподілу**.

Теорема (В.І. Глівенко, 1933) Нехай $F_n(z)$ - емпірична функція розподілу, яка побудована за вибіркою $\xi' = (\xi_1, \xi_2, ..., \xi_n)$, $F_{\xi_0}(z)$ - відповідна теоретична функція розподілу. Тоді

$$P\left(\lim_{n\to\infty} \sup_{-\infty<z<\infty} \left| F_n(z) - F_{\xi_0}(z) \right| = 0 \right) = 1.$$

Емпірична функція розподілу будується наступним чином:

1) Якщо всі елементи вибірки є різними, тобто $s = n$, тоді

$$F_n^*(x) = \begin{cases} 1, & x \ge x_{(n)} \\ k/n, & x_{(k)} \le x < x_{(k+1)} \\ 0, & x < x_{(1)} \end{cases}.$$

2) Якщо дані представлені у вигляді статистичного ряду

y_1	y_2	....	y_s
m_1	m_2	....	m_s

$$F_n^*(x) = \begin{cases} 1, & x \ge y_s \\ \dfrac{m_1 + m_2 + ... + m_k}{n}, & y_k \le x < y_{k+1} \\ 0, & x < y_1 \end{cases}$$

Приклад 3.

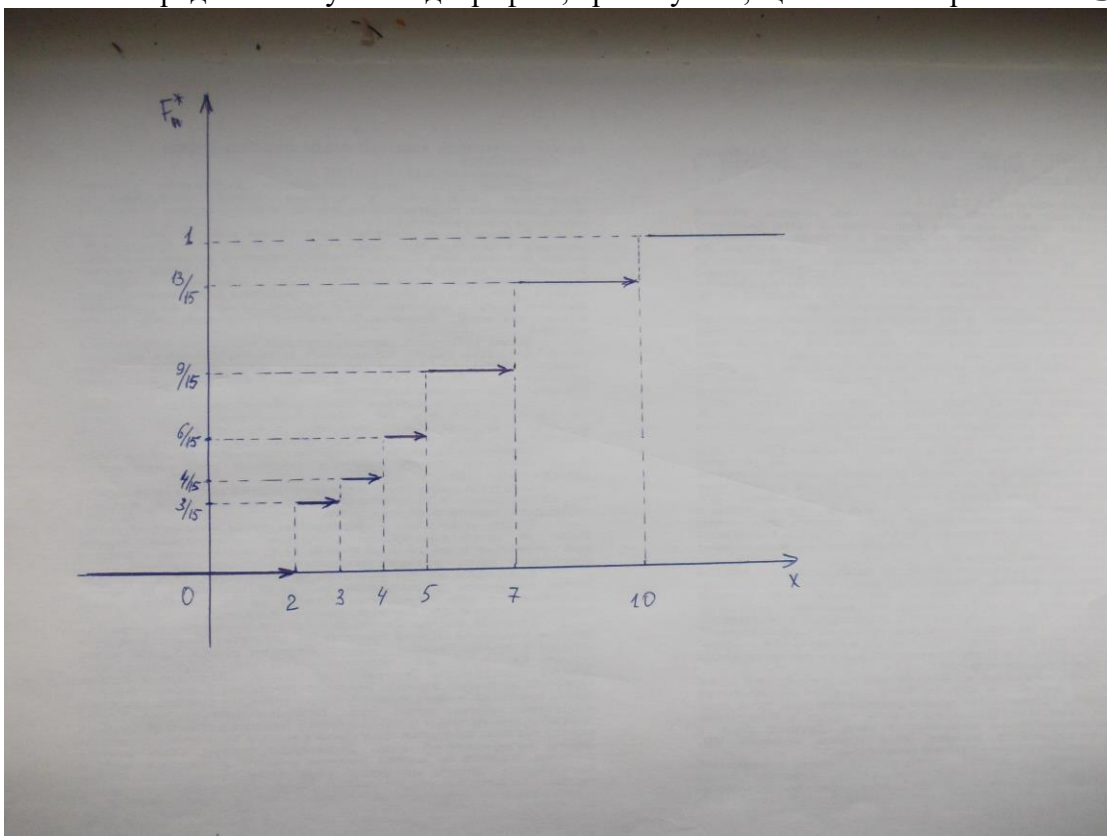
За вибіркою представленою у вигляді статистичного ряду

y_i	2	3	4	5	7	10
m_i	3	1	2	3	4	2

побудувати емпіричну функцію розподілу і оцінити ймовірність того, що наступне спостереження потрапить до інтервалу $(3,5; 7,5)$.

Розв'язання.

Розв'язок представимо у вигляді графіка, враховуючи, що об'єм вибірки $n=15$



Щоб оцінити за допомогою емпіричної функції розподілу ймовірність того, що наступне спостереження потрапить до інтервалу $(3,5;\ 7,5)$, скористаємось властивістю функції розподілу

$$P(3,5 < x_{16} < 7,5) = F_n^*(7,5) - F_n^*(3,5) = \frac{13}{15} - \frac{4}{15} = \frac{3}{5}.$$

Зауваження. Як видно з прикладу, емпірична функція розподілу має вигляд східчастої лінії і у точках, де відбуваються стрибки функції є невизначеною похідна, а отже ми не зможемо оцінити щільність функції розподілу.

Гістограма та полігон частот

Гістограма та полігон частот - це статистичні аналоги щільності, які будуються тільки для групованих даних, що відповідають розподілам абсолютно неперервного типу.

1) Гістограма: на кожному інтервалі групування будується прямокутник з висотою

$$\frac{m_i}{nh_i}, \quad i = 1,2,...,s,$$

де n - об'єм вибірки, m_i - частота i-го інтервалу, h_i - довжина i-го інтервалу.

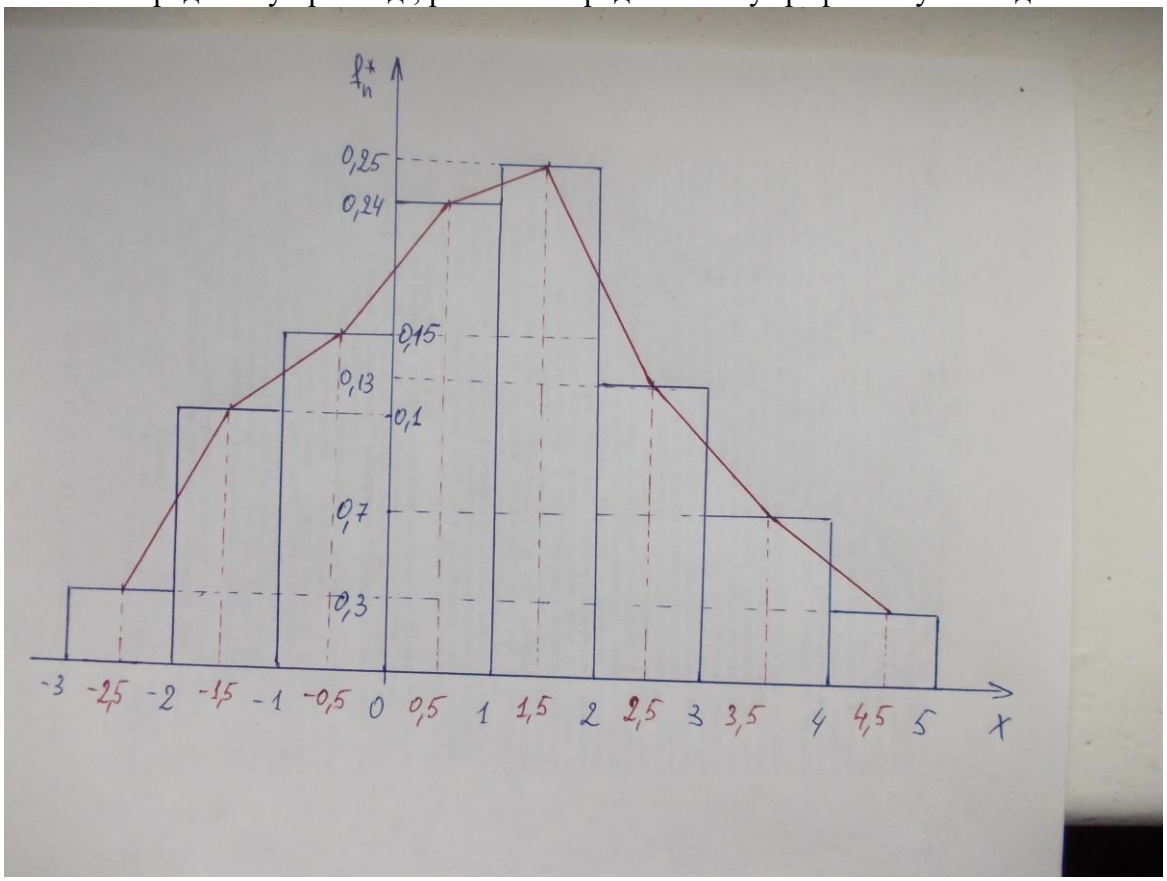
Твердження. Якщо об'єм вибірки n є великим, тоді гістограма $f_n^*(x)$ прямує до щільності теоретичної функції розподілу.

2) Полігон частот – це ламана лінія, яка з’єднує відрізками прямих наступні точки $(y_i^*, \frac{m_i}{nh_i}), i = 1,2,...,s$, де y_i^* - середина i-го інтервалу.

Приклад 4.
За наступною вибіркою побудувати гістограму та полігон частот

Інтервал	[-3, -2)	[-2, -1)	[-1, 0)	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 5)
m_i	3	10	15	24	25	13	7	3

Розв’язання.
Як і в попередньому прикладі, розв’язок представимо у графічному вигляді



На цьому малюнку синім кольором представлена гістограма, а червоним – полігон частот.

Порядкові статистики.

Задача 1. Нехай задана вибірка з генеральної сукупності $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти $M\xi_{(n)}$.

$$F_{\xi_0}(x) = \begin{cases} 1, x > 1, \\ x, x \in [0;1], \\ 0, x < 0. \end{cases}$$

$$F_{\xi_{(n)}}(x) = P\left(\xi_{(n)} \le x\right) = P\left(\max\left\{\xi_1, \xi_2, ..., \xi_n\right\} \le x\right) =$$

$$= P\left(\xi_1 \le x, \xi_2 \le x, ..., \xi_n \le x\right) = \prod_{i=1}^{n} P\left(\xi_i \le x\right) = \left(F_{\xi_0}(x)\right)^n =$$

$$= \begin{cases} 1, x > 1, \\ x^n, x \in [0;1], \\ 0, x < 0. \end{cases} \Rightarrow f_{\xi_{(n)}}(x) = \begin{cases} nx^{n-1}, x \in [0;1], \\ 0, x \notin [0;1]. \end{cases}$$

$$M\xi_{(n)} = \int_0^1 x f_{\xi_{(n)}}(x) dx = \int_0^1 x \cdot nx^{n-1} dx = \frac{n}{n+1} x^{n+1} \Big|_0^1 = \frac{n}{n+1}.$$

Задача 2. Нехай задана вибірка з генеральної сукупності $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти сумісну функцію розподілу та сумісну щільність для $\xi_{(1)}$, $\xi_{(n)}$.

$$P(\bar{A} \cap B) = P(B) - P(A \cap B);$$
$$P(B) - P(A \cap B) = P(B \setminus (A \cap B)) = P(B \cap (\overline{A \cap B})) =$$
$$= P(B \cap (\bar{A} \cup \bar{B})) = P((\bar{A} \cap B) \cup (\bar{B} \cap B)) = P(\bar{A} \cap B);$$

$$F_{\xi_{(1)}, \xi_{(n)}}(x, y) = P\left(\xi_{(1)} \le x, \xi_{(n)} \le y\right) = P(\xi_{(n)} \le y) - P\left(\xi_{(1)} > x, \xi_{(n)} \le y\right) =$$
$$= P(\xi_{(n)} \le y) - P\left(x < \xi_1 \le y, x < \xi_2 \le y, ..., x < \xi_n \le y\right) =$$

$$= F_{\xi_{(n)}}(x) - \prod_{i=1}^{n} P\left(x < \xi_i \le y\right) = \left(F_{\xi_0}(x)\right)^n - \left(F_{\xi_0}(y) - F_{\xi_0}(x)\right)^n =$$

$$= \begin{cases} y^n - (y-x)^n, 0 \le x < y \le 1; \\ y^n, x > y > 0. \end{cases}$$

$$f_{\xi_{(1)}, \xi_{(n)}}(x, y) = \begin{cases} n(n-1)(y-x)^{n-2}, 0 \le x < y \le 1; \\ 0, else. \end{cases}$$

Задача 3. Нехай задана вибірка з генеральної сукупності $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти слабку границю послідовності $n(1 - \xi_{(n)})$.

$$F_{n(1-\xi_{(n)})}(x)=P\left(n(1-\xi_{(n)})\le x\right)=P\left(1-\xi_{(n)}\le\frac{x}{n}\right)=P\left(\xi_{(n)}\ge 1-\frac{x}{n}\right)=$$

$$=1-F_{\xi_{(n)}}\left(1-\frac{x}{n}\right)=1-\left(1-\frac{x}{n}\right)^n\to 1-e^{-x},\ n\to\infty.$$

Задача 4. Нехай задана вибірка з генеральної сукупності $\xi'=(\xi_1,\xi_2,...,\xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти щільності для $\xi_{(2)}$.

Відомо, що сумісна щільність усіх порядкових статистик для заданої вибірки має вигляд:

$$f_{\xi_{(1)},\xi_{(2)},...,\xi_{(n)}}(x_1,x_2,...,x_n)=\begin{cases}n!,\ 0\le x_1<x_2<...<x_n\le 1;\\ 0,\ else.\end{cases}$$

Теорема. Нехай функція розподілу $F_{\xi,\eta}(x,y)$ має щільність $f_{\xi,\eta}(x,y)$. Тоді функції розподілу випадкових величин ξ та η теж мають щільності:

$$f_\xi(x)=\int_{-\infty}^{\infty}f_{\xi,\eta}(x,y)dy,\quad f_\eta(y)=\int_{-\infty}^{\infty}f_{\xi,\eta}(x,y)dx,$$

$$f_{\xi_{(2)}}(x_2)=n!\int_0^{x_2}dx_1\left[\int_{x_2}^{1}...\left(\int_{x_{n-2}}^{1}\left(\int_{x_{n-1}}^{1}dx_n\right)dx_{n-1}\right)...dx_3\right]=$$

$$=n(n-1)x_2(1-x_2)^{n-2},\quad 0\le x_2\le 1.$$

Порядкові статистики (продовження)

Нехай задано вибірку, яка складається з незалежних у сукупності, однаково розподілених елементів, які мають неперервну функцію розподілу $F_{\xi_0}(x)$. Тоді функцію розподілу k-тої порядкової статистики $\xi_{(k)}$ можна подати у такому вигляді

$$F_{\xi_{(k)}}(x)=P\left(\xi_{(k)}\le x\right)=\sum_{i=k}^{n}C_n^iF_{\xi_0}^i(x)\left(1-F_{\xi_{(0)}}(x)\right)^{n-i},$$

А якщо існує щільність розподілу елементів вибірки $f_{\xi_0}(x)$, то для щільність розподілу k-тої порядкової статистики $\xi_{(k)}$ справедливе наступне співвідношення

$$f_{\xi_{(k)}}(x)=nC_{n-1}^{k-1}F_{\xi_0}^{k-1}(x)\left(1-F_{\xi_{(0)}}(x)\right)^{n-k}f_{\xi_0}(x).$$

ЗАДАЧІ

1) Нехай дані представлені у вигляді групованої вибірки

Інтервал	[-6; -4)	[-4; -2)	[-2; 0)	[0; 2)	[2; 4)	[4; 6)
Частота	7	4	9	5	5	10

Потрібно побудувати емпіричну функцію розподілу, гістограму та полігон частот.

2) Нехай задана вибірка з генеральної сукупності $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти слабку границю послідовності $n\xi_{(1)}$.

3) Нехай задана вибірка з генеральної сукупності $\xi' = (\xi_1, \xi_2, ..., \xi_n)$ з рівномірним розподілом на відрізку $[0;1]$. Знайти щільності для $\xi_{(n-1)}$.