ЛАБОРАТОРНА РОБОТА № 5

Статистичний аналіз одно- та двовимірної послідовності випадкових чисел.

Основні означення. Одним із основних понять математичної статистики є **статистичний ансамбль (або сукупність),** під яким розуміють сукупність певних об'єктів однієї природи, які володіють **змінними ознаками**. Статистичний ансамбль складається з окремих елементів (або одиниць), загальне число яких називають **об'ємом ансамблю**. При цьому елементи ансамблю можуть бути охарактеризованими **одним або декількома параметрами (ознаками)**.

Ознаки, які приймають різні значення, називають **варіаційними,** які, своєю чергою, можуть бути **якісними (або атрибутивними**) та **кількісними.** Якісні ознаки не можна виразити числом, тоді як кількісні мають числову міру.

Ознаки можуть бути **дискретними та неперервними**.

Залежно від повноти досліджуваних елементів розрізняють **генеральну та вибіркову сукупності**. Сукупність явищ, з яких здійснюють відбір частини елементів для вибіркового спостереження, називають **генеральною** сукупністю. Частина елементів, яку вибирають із генеральної сукупності, називають **вибірковим ансамблем** (або вибірковою сукупністю).

Основна задача формування вибіркової сукупності полягає у тому, щоб показники, які її характеризують, з найбільшою точністю відтворюють показники генеральної сукупності.

Залежно від числа ознак, за якими вибираються ансамблі, розрізняють одновимірні, двовимірні та багатовимірні сукупності.

Загальна схема аналізу статистичних ансамблів.

Експериментальний аналіз одновимірної випадкової величини. Опрацювання одновимірної вибірки випадкових величин x_1, x_2, ... x_N здійснюють у такому порядку.

1. Побудова варіаційного ряду. Варіаційний ряд z_1, z_2, ... z_N отримують із вихідних даних x_1, x_2, ... x_N шляхом їхнього розміщення у порядку зростання $x_{\min} = z_1 \le z_2 \le ... \le z_N = x_{\max}$.

Проілюструємо побудову ряду на прикладі із п’яти спостережень: $x_1 = 5, \quad x_2 = 2, \quad x_3 = 4, \quad x_4 = 5, \quad x_5 = 7$. Цим даним відповідає такий варіаційний ряд: $z_1 = 2, \quad z_2 = 4, \quad z_3 = 5, \quad z_4 = 5, \quad z_5 = 7$.

2. Побудова діаграми накопичених частот. Діаграма накопичених частот $\widehat{F}_N(x)$ служить аналогом функції інтегрального розподілу і будується згідно формули:

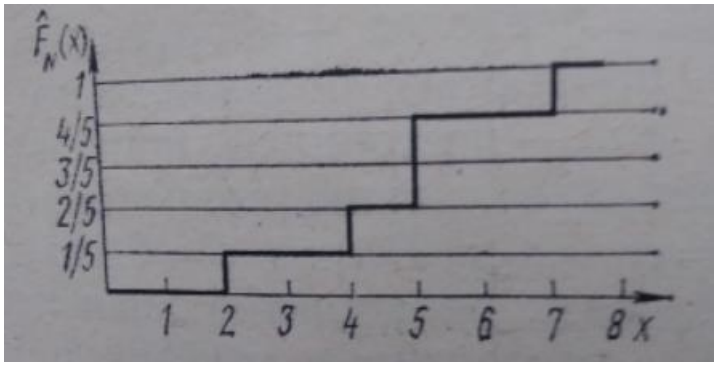
$$\widehat{F}_N(x) = \sum_{i=1}^{\mu_N(x)} \frac{1}{N},$$

де $\mu_N(x)$ - число елементів у вибірці, для яких значення $x_j < x$.

Для побудови по осі абсцис вказуються значення спостережень z_i (або $x_м$), а значення по осі ординат рівне нулеві для точок, менших від x_{min}. У інших точках $x_м$ діаграма має стрибок, який рівний $\frac{1}{N}$. Якщо існує декілька (λ) однакових значень $x_м$, то в цьому місці стрибок рівний $\frac{\lambda}{N}$. Для величин $x > x_{max}$ значення діаграми накопичених частот є рівним одиниці.

Відмітимо, що $\widehat{F}_N(x) \to F(x)$.

Для розглянутого прикладу діаграма накопичених частот має вигляд, показаний на рис.1.



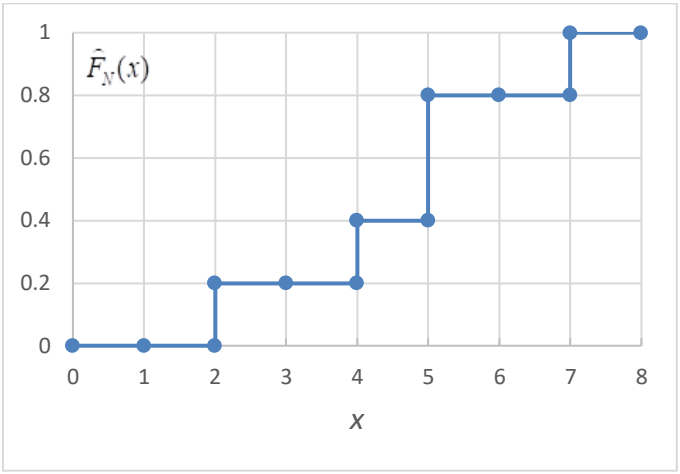


Рис.1

3. Побудова гістограми вибірки. Гістограма $\hat{w}_N(x)$ служить емпіричним аналогом густини розподілу ймовірностей $w(x)$. Процес побудови гістограми містить такі кроки:

1. Спочатку знаходять число інтервалів K, на які слід розділити вісь абсцис. Кількість інтервалів визначають за допомогою оціночної формули $K = 1 + 3{,}2 \lg N$. Зазвичай, отримане значення заокруглюють до найближчого цілого числа.

2. Визначають довжину інтервалу $\Delta x = \left(x_{max} - x_{min}\right) / K$, яку також заокруглюють для зручності розрахунків.

3. Середину ділянки зміни вибірки $\left(x_{max} + x_{min}\right) / 2$ приймають за центр деякого інтервалу, після чого знаходять границі та число інтервалів так, щоб вони повністю перекривали весь діапазон від x_{min} до x_{max}.

4. Підраховують число спостережень N_m, які попадають у кожний інтервал. N_m рівне числу членів варіаційного ряду, для яких виконується співвідношення $x_m \le z_l < x_m + \Delta x$, де x_m та $x_m + \Delta x$ - межі $m-$того інтервалу. При цьому значення, які потрапили на межу між $m-1$ та $m-$тим інтервалами відносять до $m-$того інтервалу.

5. Підраховують відносну частоту N_m / N для цього інтервалу.

6. Будують гістограму - сходинчасту криву, значення якої на $m-$тому інтервалі $\left(x_m, x_m + \Delta x\right)$, $m = 1, 2, \ldots K$, є постійним і рівним N_m / N.

Для ілюстрації вказаного алгоритму розглянемо вибірку із 40 спостережень, для якої варіаційний ряд має вигляд:

$x_{\min} = z_1 = 0,3, \quad z_2 = 0,4, \quad z_3 = 0,5, \quad ..., \quad z_{40} = x_{\max} = 7,1.$

Її аналіз здійснюємо у розглянутій вище послідовності.

1. Із формули $K = 1 + 3,2 \lg N$ знаходимо, що $K = 1 + 3,2 \lg 40 = 6,13$. Тому візьмемо значення K=7.

2. Довжина інтервалу $\Delta x = (7,1 - 0,3) / 7 = 0,971$. Приймаємо $\Delta x = 1$.

3. Знаходимо середину ділянки $(x_{\min} + x_{\max}) / 2 = (0,3 + 7,1) / 2 = 3,7$, після чого легко визначити межі інтервалів (рис.2).

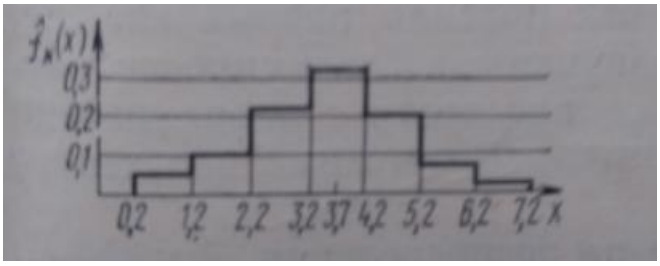


Рис.2

Будемо вважати, що після такого визначення інтервалів у них потрапили такі значення x_l:

$N_1 = 2, \; N_2 = 4, \quad N_3 = 9, \quad N_4 = 13, \quad N_5 = 8, \quad N_6 = 3, \quad N_7 = 1.$

Відповідна цим значенням гістограма показана на рис.2.

4. ***Оцінювання математичного сподівання, дисперсії та середньоквадратичного відхилення*** здійснюємо за формулами:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2,$$

$$\sigma_x = +\sqrt{\sigma_x^2}.$$

Експериментальний аналіз двовимірної сукупності.

Нехай у результаті спостережень отримано вибірку з двовимірної сукупності двох випадкових величин *x* та *y* (таблиця1).

j	1	2	3	…	j	…	N
X	x_1	x_2	x_3	…	x_j	…	x_N
Y	y_1	y_2	y_3	…	y_j	…	y_N

Опрацювання отриманих результатів спостереження будемо здійснювати за такою схемою.

1. Побудова поля розсіяння. Перший крок при аналізі результатів спостереження двовимірної сукупності випадкових величин *X* та *Y* полягає у побудові поля розсіяння. Для цього на площину з координатними осями *x* та *y* наносять експериментальні точки. Можливий вигляд такого поля показано на рис. 3.

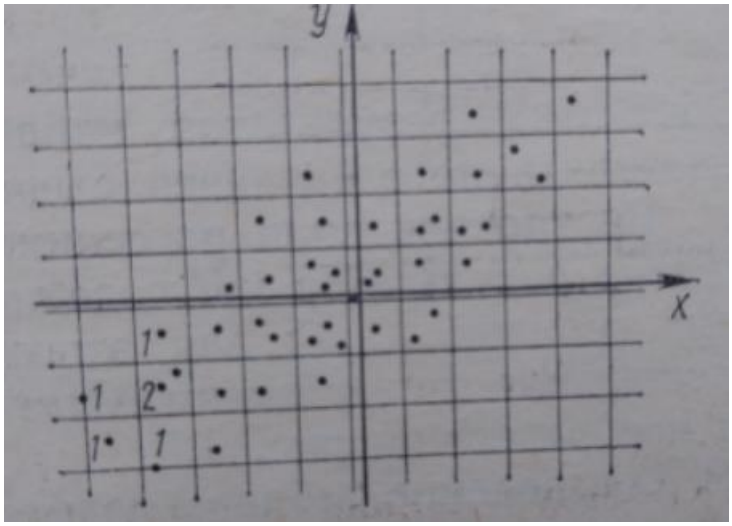


Рис.3.

2. Формування таблиці двовимірного розподілу. Ця таблиця формується так. Розбиваємо осі *0x* та *0y* на окремі інтервали Δx та Δy. Величини цих інтервалів, їхню кількість та розміщення визначають для кожної змінної за допомогою розглянутих правил. Оскільки число точок є спільним для величин *X* та *Y*, то, зазвичай, $K_x = K_y = K$.

Відповідні межі інтервалів наносять на діаграму розсіяння (рис.3), і підраховують число точок, які потрапили у кожен з утворених прямокутників. Якщо якась точка потрапила на межу прямокутника, то її відносять до правого та/або верхнього прямокутника.

У подальшому формуємо таблицю, у якій відмічаємо величини $N_{m_1 m_2}$ та $N_{m_1 m_2} / N$. Отриману таблицю можна використати як вихідну для побудови гістограм та діаграм накопичення частот у тривимірному просторі, які є експериментальними аналогами двовимірної функції розподілу та двовимірної густини ймовірності.

За допомогою таблиці двовимірного розподілу можна отримати вихідні дані для побудови гістограм, які відповідають кожній величині *X* та *У*. Для цього потрібно просумувати значення таблиць по стовпцю (для величини *У*) або стрічці (для величини X).

Для ілюстрації наведена таблиця, побудована по числових даних рис.3.

Інтервали для y та x	$y_1, y_1 + \Delta y$	$y_2, y_2 + \Delta y$	…	$y_{m_2}, y_{m_2} + \Delta y$	…	$y_{k_2}, y_{k_2} + \Delta y$
$x_1, x_1 + \Delta x$	0					1
	0					1/40
$x_2, x_2 + \Delta x$	0					
	0					
…						
$x_{m_1}, x_{m_1} + \Delta x$				$N_{m_1 m_2}$		
				$N_{m_1 m_2} / N$		
…						
$x_{k_1}, x_{k_1} + \Delta x$	1	1				
	1/40	1/40				

3. Розрахунок коефіцієнта кореляції. Обчислення коефіцієнта кореляції проводять за формулою

$$\hat{\rho}_{xy} = \frac{1}{(N-1)\sigma_x \sigma_y} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Середні значення та дисперсії, які фігурують у виразі для коефіцієнта кореляції, розраховують за наведеними вище формулами (п.4).

Зауважимо, що між значенням та знаком коефіцієнта кореляції та діаграмою розсіяння існує певний зв'язок. Зокрема, якщо початок координат сумістити з середніми значеннями $\overline{x}$ та $\overline{y}$, то:

при $\rho > 0$ точки на діаграмі розсіяння групуються переважно у I та III квадрантах, а при $\rho < 0$ - у II та IУ квадрантах;

якщо $\rho \approx 0$ точки хаотично розміщені по всіх квадрантах, а коли $\rho \approx 1$ - то точки групуються вздовж прямих, що знаходяться у I або III квадрантах. Відповідно, при $\rho \approx -1$ прямі розміщені у II або IУ квадрантах.

Завдання і порядок виконання роботи

1. Дослідити властивості одновимірної випадкової величини X_1 Для реалізації цього завдання:

- Згенерувати вибірку випадкової величини X_1, яка містить не менше 30 значень при заданому середньоквадратичному відхиленні $\sigma_x = 15$;
- Побудувати варіаційний ряд;
- З допомогою варіаційного ряду побудувати діаграму накопичених частот;
- Побудувати гістограму вибірки;
- Розрахувати за формулами оцінки математичного сподівання, дисперсії та середньо квадратичного відхилення випадкової величини X_1.

2. Дослідити властивості двовимірної сукупності випадкових величин X_1 та Y_1. Для реалізації цього завдання:

- Згенерувати різні вибірки випадкових величин X_1 та Y_1.
- Побудувати поля розсіяння;
- Скласти таблицю двовимірних розподілів;
- Порахувати коефіцієнти кореляції.