

МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ

Харківський національний університет внутрішніх справ

**Навчально-науковий інститут підготовки фахівців для підрозділів
кримінальної міліції**

Кафедра інформаційної та економічної безпеки

ТЕКСТИ ЛЕКЦІЙ

**з дисципліни
„Математична статистика”**

Галузь знань	0301 - „Соціально-політичні науки”
Напрямок підготовки	6.030102 - „ Психологія ”
Спеціалізація	кримінальна міліція у справах дітей
Ступень вищої освіти	бакалавр

Харків 2015

СХВАЛЕНО

Науково-методичною радою
Харківського національного
університету внутрішніх справ
_25.08.2015 Протокол № 8
(дата, місяць, рік)

ЗАТВЕРДЖЕНО

Вченою радою навчально-наукового
інституту кримінальної міліції
21.08.2015 Протокол № 1/15-16
(дата, місяць, рік)
_____ Корнієнко Д.М.
(підпис) (П.І.Б.)

ПОГОДЖЕНО

Секцією Науково-методичної ради
ХНУВС з технічних дисциплін
_25.08.2015 Протокол № 8
(дата, місяць, рік)
_____ Сезонова І.К.
(підпис) (П.І.Б.)

ЗАТВЕРДЖЕНО

На засіданні кафедри інформаційної та
економічної безпеки
20.08.2015 Протокол № _12_
(дата, місяць, рік)
_____ Сезонова І.К.
(підпис) (П.І.Б.)

Рецензенти:

Гнусов Ю.В., кандидат технічних наук, доцент кафедри захисту інформації
ФПФПБКТЛ ХНУВС

Розробник: Соколовська О.Г. - Харків, Харківський національний
університет внутрішніх справ, Навчально-науковий інститут підготовки
фахівців для підрозділів кримінальної міліції ХНУВС, 2015

ВСТУП

Основним змістом математичної статистики є систематизація, обробка і використання статистичної інформації для виявлення статистичних закономірностей ознаки або ознак певної сукупності елементів.

Метою математичної статистики є розробка методів обробки статистичних даних з метою отримання наукових і практичних висновків. Теоретичною основою математичної статистики є теорія ймовірностей.

Найбільш природним шляхом, яким математика проникає в психологію, є математична статистика. Сучасна статистика є розділом математики. При цьому багато статистичних процедур досить прості й легко здійсненні.

Правильне застосування статистики дозволяє психологові:

- 1) доводити правильність і обґрунтованість використовуваних методичних прийомів і методів;
- 2) строго обґрунтовувати експериментальні плани;
- 3) узагальнювати дані експерименту;
- 4) знаходити залежності між експериментальними даними;
- 5) виявляти наявність істотних розходжень між групами випробуваних (наприклад, експериментальними й контрольними);
- 6) будувати статистичні прогнозування;
- 7) уникати логічних і змістовних помилок і багато чого іншого.

Не можна забувати, однак, що сама по собі статистика - це тільки інструментарій, що допомагає психологові ефективно розбиратися в складному експериментальному матеріалі. Найбільш важливим у будь-якому експерименті є чітка постановка завдання, ретельне планування експерименту, побудова несуперечливих гіпотез.

Математична статистика в руках психолога може й повинна бути потужним інструментом, що дозволяє не тільки успішно лавірувати в морі експериментальних даних, але й, насамперед, сприяти становленню його об'єктивного мислення.

Оскільки суцільна обробка всіх елементів сукупності практично неможлива, то, як правило, застосовується вибіркового метод. Отже, розрізняють генеральну і вибіркoву сукупності.

Групу об'єктів, поєднаних за якоюсь якісною чи кількісною ознакою, називають статистичною сукупністю. Розрізняють генеральну і вибіркoву сукупності. Вибірковою сукупністю або вибіркою називають сукупність випадково відібраних об'єктів. Генеральною сукупністю називають сукупність об'єктів, з якої виконується вибірка. Об'ємом сукупності називають кількість об'єктів, що до неї входить.

Множина Ω однотипних елементів, яким притаманні певні кількісні ознаки (розміри, вага, маса тощо), утворює генеральну сукупність. Кількість усіх елементів генеральної сукупності називають її обсягом і позначають символом N , значення якого здебільшого невідоме.

Кожна непорожня підмножина A множини Ω ($A \subset \Omega$) випадково вибраних елементів із генеральної сукупності називається вибіркою. Кількість усіх елементів вибірки називають її обсягом і позначають символом n . Його значення відоме, причому воно набагато менше за обсяг генеральної сукупності ($n \ll N$).

Математична статистика розв'язує дві категорії задач:

- 1) статистичне оцінювання (точкове, інтервальне) параметрів генеральної сукупності;
- 2) перевірка правдивості статистичних гіпотез про значення параметрів генеральної

сукупності або про закон розподілу ознаки генеральної сукупності на підставі обробки результатів вибірки.

Тема № 1: Основні поняття в математичній статистиці.

Лекція 1 за темою № 1 *Статистичні сукупності. Дискретний статистичний розподіл вибірки та її числові характеристики.*

Кількісні ознаки елементів генеральної сукупності можуть бути одновимірними і багатовимірними, дискретними і неперервними.

Коли реалізується вибірка, кількісна ознака, наприклад X , набуває конкретних числових значень ($X = x_i$), які називають *варіантою*.

Зростаючий числовий ряд варіант називають *варіаційним*.

Кожна варіанта вибірки може бути спостереженою n_i раз ($n_i \geq 1$), число n_i називають *частотою варіанти* x_i .

При цьому

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

де k — кількість варіант, що різняться числовим значенням;

n — обсяг вибірки.

Відношення частоти n_i варіанти x_i до обсягу вибірки n називають її *відносною частотою* і позначають через W_i , тобто

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

Для кожної вибірки виконується рівність

$$\sum_{i=1}^k W_i = 1.$$

Якщо досліджується ознака генеральної сукупності X , яка є неперервною, то варіант буде багато. У цьому разі варіаційний ряд — це певна кількість рівних або нерівних частинних інтервалів чи груп варіант зі своїми частотами.

Такі частинні інтервали варіант, які розміщені у зростаючій послідовності, утворюють *інтервальний варіаційний ряд*.

На практиці для зручності, як правило, розглядають інтервальні варіаційні ряди, у котрих інтервали є рівними між собою.

Перелік варіант варіаційного ряду і відповідних їм частот, або відносних частот, називають *дискретним статистичним розподілом вибірки*.

У табличній формі він має такий вигляд:

$X = x_i$	x_1	x_2	x_3	...	x_k
n_i	n_1	n_2	n_3	...	n_k
W_i	W_1	W_2	W_3	...	W_k

Дискретний статистичний розподіл вибірки можна подати емпіричною функцією $F^*(x)$. *Емпірична функція $F^*(x)$ та її властивості.* Функція аргументу x , що визначає відносну частоту події $X < x$, тобто

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

називається *емпіричною*, або *кумулятою*.

Тут n — обсяг вибірки;

n_x — кількість варіант статистичного розподілу вибірки, значення яких менше за фіксовану варіанту x ;

$F^*(x)$ — називають ще *функцією нагромадження відносних частот*.

Властивості $F^(x)$:*

- 1) $0 \leq F^*(x) \leq 1$;
- 2) $F(x_{\min}) = 0$, де x_{\min} є найменшою варіантою варіаційного ряду;
- 3) $F(x)|_{x > x_{\max}} = 1$, де x_{\max} є найбільшою варіантою варіаційного ряду;
- 4) $F(x)$ є неспадною функцією аргументу x , а саме: $F(x_2) \geq F(x_1)$ при $x_2 \geq x_1$.

Полігон частот і відносних частот. Дискретний статистичний розподіл вибірки можна зобразити графічно у вигляді ламаної лінії, відрізки якої сполучають координати точок $(x_i; n_i)$, або $(x_i; W_i)$.

У першому випадку ламану лінію називають *полігоном частот*, у другому — *полігоном відносних частот*.

Приклад. За заданим дискретним статистичним розподілом вибірки

$X = x_i$	-6	-4	-2	2	4	6
n_i	5	10	15	20	40	10
W_i	0,05	0,1	0,15	0,2	0,4	0,1

потрібно:

1. Побудувати $F^*(x)$ і зобразити її графічно;
2. Накреслити полігони частот і відносних частот.

Розв'язання. Згідно з означенням та властивостями $F^*(x)$ має такий вигляд:

$$F^*(x) = W(X < x) = \frac{n_x}{n} = \begin{cases} 0 & x \leq -6, \\ 0,05 & -6 < x \leq -4, \\ 0,15 & -4 < x \leq -2, \\ 0,3 & -2 < x \leq 2, \\ 0,5 & 2 < x \leq 4, \\ 0,9 & 4 < x \leq 6, \\ 1, & x > 6. \end{cases}$$

Графічне зображення $F^*(x)$ подано на рис. 106.

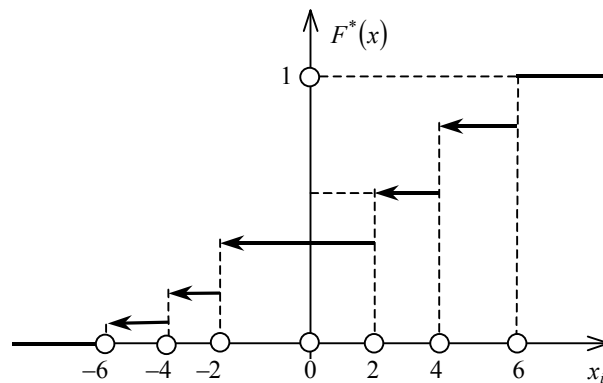


Рис. 1.1

Полігони частот та відносних частот зображено на рис.1.1, 1.3.

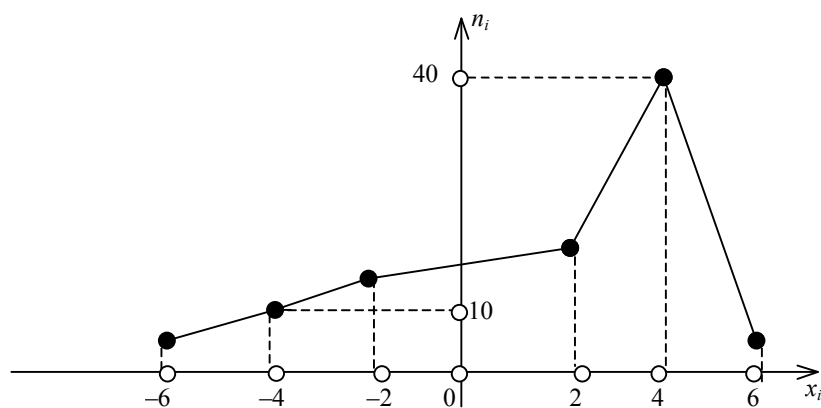


Рис. 1.2

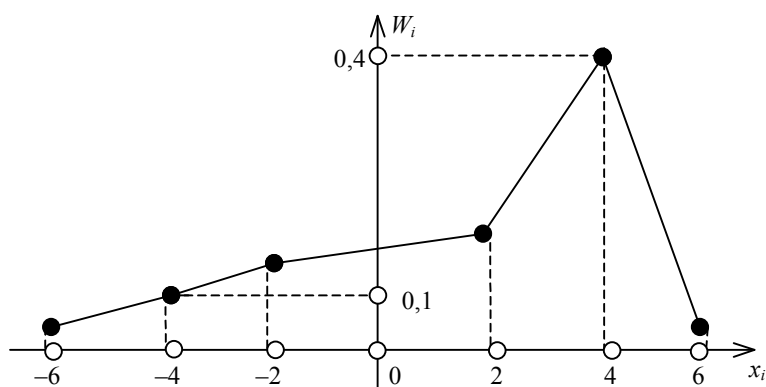


Рис. 1.3

Числові характеристики:

1) *вибіркова середня величина* \bar{x}_B . Величину, яка визначається формулою

$$\bar{x}_B = \frac{\sum x_i n_i}{n},$$

називають *вибірковою середньою величиною дискретного статистичного розподілу вибірки*.

Тут x_i — варіанта варіаційного ряду вибірки;

n_i — частота цієї варіанти;

n — обсяг вибірки ($n = \sum n_i$).

Якщо всі варіанти з'являються у вибірці лише по одному разу, тобто $n_i = 1$, то

$$\bar{x}_B = \frac{\sum x_i}{n};$$

2) *відхилення варіант*. Різницю $(x_i - \bar{x}_B)n_i$ називають відхиленням варіант.

При цьому

$$\sum (x_i - \bar{x}_B)n_i = \sum x_i n_i - \sum \bar{x}_B n_i = n \cdot \bar{x}_B - n \cdot \bar{x}_B = 0.$$

Отже, сума відхилень усіх варіант варіаційного ряду вибірки завжди дорівнює нулеві;

3) *мода* (Mo^*). *Модой дискретного статистичного розподілу вибірки* називають варіанту, що має найбільшу частоту появи.

Мод може бути кілька. Коли дискретний статистичний розподіл має одну моду, то він називається *одномодальним*, коли має дві моди — *двомодальним* і т. д.;

4) *медіана* (Me^*). *Медіаною дискретного статистичного розподілу вибірки* називають варіанту, яка поділяє варіаційний ряд на дві частини, рівні за кількістю варіант;

5) *дисперсія*. Для вимірювання розсіювання варіант вибірки відносно \bar{x}_B вибирається дисперсія.

Дисперсія вибірки — це середнє арифметичне квадратів відхилень варіант відносно \bar{x}_B , яке обчислюється за формулою

$$D_B = \frac{\sum (x_i - \bar{x}_B)^2 n_i}{n}$$

або

$$D_B = \frac{\sum x_i^2 n_i}{n} - (\bar{x}_B)^2;$$

6) *середнє квадратичне відхилення вибірки* σ_B . При обчисленні D_B відхилення підноситься до квадрата, а отже, змінюється одиниця виміру ознаки X , тому на основі дисперсії вводиться середнє квадратичне відхилення

$$\sigma_B = \sqrt{D_B},$$

яке вимірює розсіювання варіант вибірки відносно \bar{x}_B , але в тих самих одиницях, в яких вимірюється ознака X ;

7) *розмах (R)*. Для грубого оцінювання розсіювання варіант відносно \bar{x}_B застосовується величина, яка дорівнює різниці між найбільшою x_{\max} і найменшою x_{\min} варіантами варіаційного ряду. Ця величина називається *розмахом*

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

8) *коефіцієнт варіації V*. Для порівняння оцінок варіацій статистичних рядів із різними значеннями \bar{x}_B , які не дорівнюють нулеві, вводиться коефіцієнт варіації, який обчислюється за формулою

$$V = \frac{\sigma_B}{\bar{x}_B} 100\%.$$

Приклад. За заданим статистичним розподілом вибірки

$X = x_i$	2,5	4,5	6,5	8,5	10,5
n_i	10	20	30	30	10

потрібно:

- 1) обчислити \bar{x}_B , D_B , σ_B ;
- 2) знайти Mo^* , Me^* ;
- 3) обчислити R , V .

Розв'язання. Оскільки $n = \sum n_i = 100$, то згідно з формулами (354), (357), (358) дістанемо:

$$\bar{x}_B = \frac{\sum x_i n_i}{n} = \frac{2,5 \cdot 10 + 4,5 \cdot 20 + 6,5 \cdot 30 + 8,5 \cdot 30 + 10,5 \cdot 10}{100} = 6,7;$$

$$\bar{x}_B = 6,7.$$

Для обчислення D_B визначається

$$\frac{\sum x_i^2 n_i}{n} = \frac{(2,5)^2 \cdot 10 + (4,5)^2 \cdot 20 + (6,5)^2 \cdot 30 + (8,5)^2 \cdot 30 + (10,5)^2 \cdot 10}{100} = 50,05.$$

Тоді

$$D_B = \frac{\sum x_i^2 n_i}{n} - (\bar{x}_B)^2 = 50,05 - (6,7)^2 = 50,05 - 44,89 = 5,16.$$

$$D_B = 5,16.$$

$$\sigma_B = \sqrt{D_B} = \sqrt{5,16} \approx 2,27.$$

$$\sigma_B = 2,27.$$

$$Mo^* = 6,5; 8,5.$$

Отже, наведений статистичний розподіл вибірки буде двомодальним. $Me^* = 6,5$, оскільки варіанта $x = 6,5$ поділяє варіаційний ряд 2,5; 4,5; **6,5**; 8,5; 10,5 на дві частини: 2,5; 4,5 і 8,5; 10,5, які мають однакову кількість варіант.

$$R = x_{\max} - x_{\min} = 10,5 - 2,5 = 8.$$

$$V = \frac{\sigma_B}{\bar{x}_B} 100\% = \frac{2,27}{6,7} 100\% = 33,88\%.$$

Лекція 2 за темою № 1. Інтервальний статистичний розподіл вибірки та його числові характеристики.

Перелік часткових інтервалів і відповідних їм частот, або відносних частот, називають *інтервальним статистичним розподілом вибірки*.

У табличній формі цей розподіл має такий вигляд:

h	$x_1 - x_2$	$x_2 - x_3$	$x_3 - x_4$...	$x_{k-1} - x_k$
n_i	n_1	n_2	n_3	...	N_k
W_i	W_1	W_2	W_3	...	W_k

Тут $h = x_i - x_{i-1}$ є довжиною часткового i -го інтервалу. Як правило, цей інтервал береться однаковим.

Інтервальний статистичний розподіл вибірки можна подати графічно у вигляді гістограми частот або відносних частот, а також, як і для дискретного статистичного розподілу, емпіричною функцією $F^*(x)$ (кумулятою).

Гістограма частот та відносних частот. Гістограма частот являє собою фігуру, яка складається з прямокутників, кожен з яких має основу h і висоту $n_i \frac{1}{h}$.

Гістограма відносних частот є фігурою, що складається з прямокутників, кожен з яких має основу завдовжки h і висоту, що дорівнює $W_i \frac{1}{h}$.

Приклад. За заданим інтервальним статистичним розподілом вибірки

$h = 8$	0—8	8—16	16—24	24—32	32—40	40—48
n_i	10	15	20	25	20	10
W_i	0,1	0,15	0,2	0,25	0,2	0,1

потрібно побудувати гістограму частот і відносних частот.

Розв'язання. Гістограми частот і відносних частот наведені на рис. 1.4, 1.5.

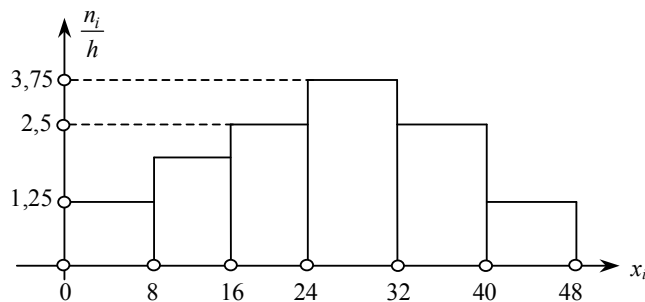


Рис. 1.4

Площа гістограми частот $S = \sum h \frac{n_i}{h} = \sum n_i = n = 100$.

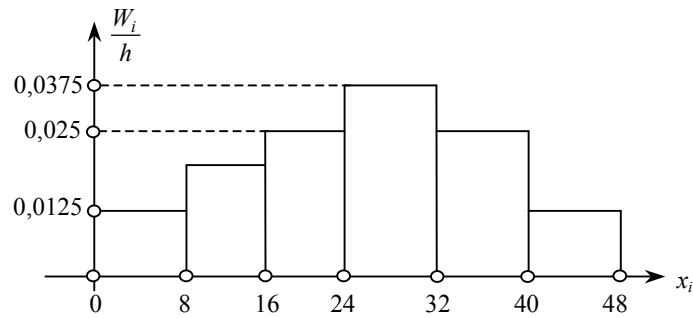


Рис. 1.5

Площа гістограми відносних частот

$$S = \sum h \frac{W_i}{h} = \sum W_i = 1.$$

Емпірична функція $F^*(x)$ (комулята). При побудові комуляти $F^*(x)$ для інтервального статистичного розподілу вибірки за основу береться припущення, що ознака на кожному частинному інтервалі має рівномірну щільність імовірностей. Тому комулята матиме вигляд ламаної лінії, яка зростає на кожному частковому інтервалі і наближається до одиниці.

Приклад. Для заданого інтервального статистичного розподілу вибірки

$h = 10$	$0 - 10$	$10 - 20$	$20 - 30$	$30 - 40$	$40 - 50$	$50 - 60$
n_i	5	15	20	25	30	5

побудувати $F^*(x)$ і подати її графічно.

Розв'язання.

$$F^*(x) = W(X < x) = \frac{n_x}{n} = \begin{cases} 0, & x \leq 0, \\ 0,05 & 0 < x \leq 10, \\ 0,2 & 10 < x \leq 20, \\ 0,4 & 20 < x \leq 30, \\ 0,65 & 30 < x \leq 40, \\ 0,95 & 40 < x \leq 50, \\ 1 & 50 < x \leq 60. \end{cases}$$

Графік $F^*(x)$ зображено на рис. 1.6.

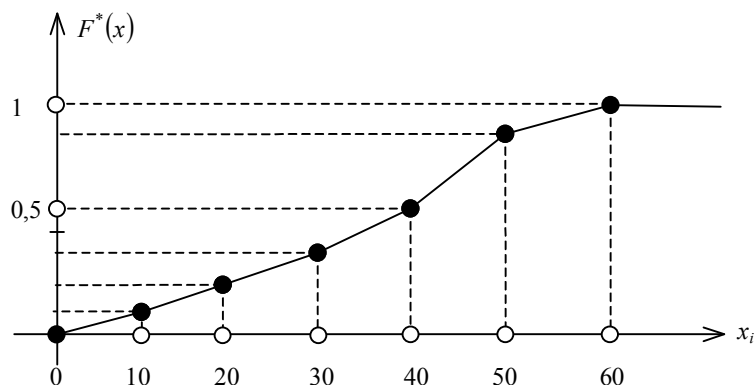


Рис. 1.6

Аналогом емпіричної функції $F^*(x)$ у теорії ймовірностей є інтегральна функція $F(x) = P(X < x)$.

Медіана. Для визначення медіани інтервального статистичного розподілу вибірки необхідно визначити інтервал, в якому знаходиться медіана.

існує таке значення $X = Me$, де $F^*(Me) = 0,5$.

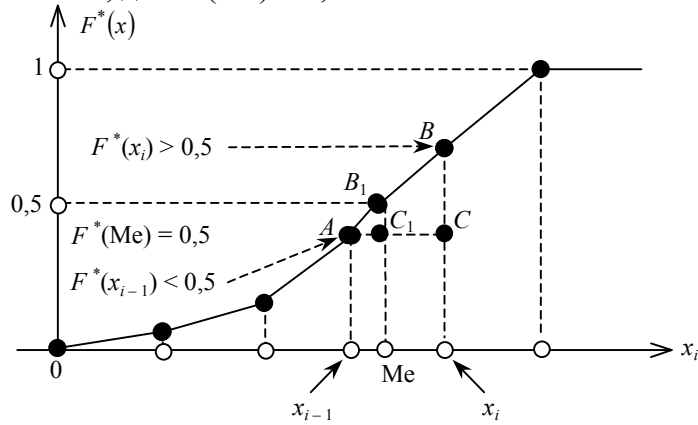


Рис. 1.7

З подібності трикутників $\triangle ABC$ і $\triangle AB_1C_1$, зображених на рис. 1.7, маємо:

$$\frac{x_i - x_{i-1}}{Me^* - x_{i-1}} = \frac{F^*(x_i) - F^*(x_{i-1})}{0,5 - F^*(x_{i-1})} \rightarrow Me^* = x_{i-1} + \frac{0,5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} h,$$

де $h = x_i - x_{i-1}$ називають *кроком*.

Мода. Для визначення моди інтервального статистичного розподілу необхідно знайти модальний інтервал, тобто такий частинний інтервал, що має найбільшу частоту появи.

Використовуючи лінійну інтерполяцію, моду обчислимо за формулою

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} h,$$

де x_{i-1} — початок модального інтервалу;

h — довжина, або крок, часткового інтервалу;

n_{Mo} — частота модального інтервалу;

n_{Mo-1} — частота домодального інтервалу;

n_{Mo+1} — частота післямодального інтервалу.

Приклад. За заданим інтервальним статистичним розподілом вибірки

$h = 4$	0—4	4—8	8—12	12—16	16—20	20—24
n_i	6	14	20	25	30	5

побудувати гістограму частот і $F^*(x)$.

Визначити Mo^* , Me^* .

Розв'язання. Гістограма частот зображена на рис. 1.8.

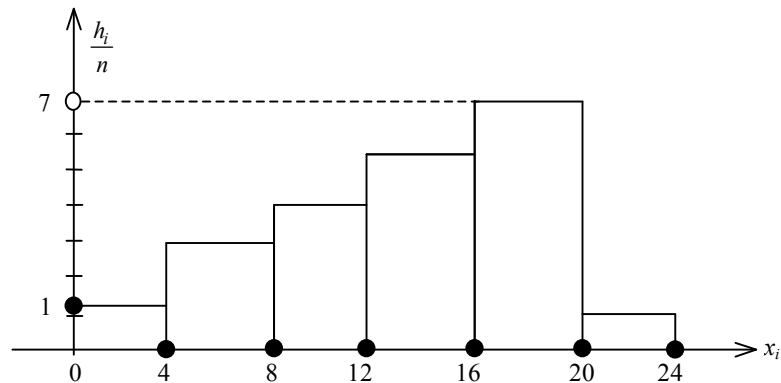


Рис. 1.8

Графік $F^*(x)$ зображено на рис. 1.9.

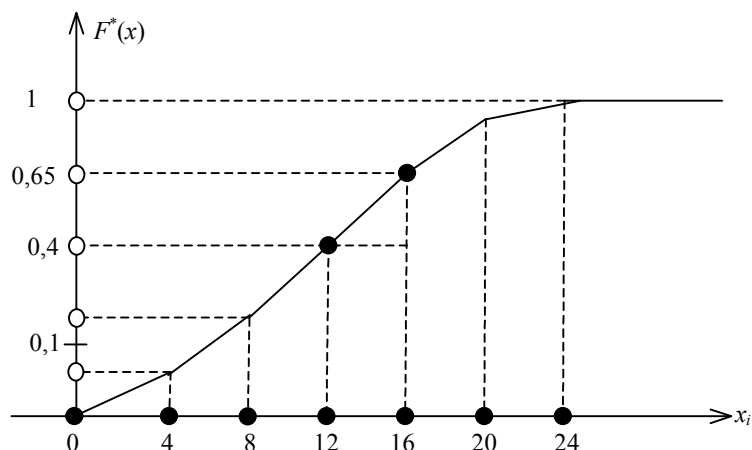


Рис. 1.9

З рис. 1.9 визначається модальний інтервал, який дорівнює 16—20.

Застосовуючи (362) і беручи до уваги, що $n_{Mo} = 30$, $n_{Mo-1} = 25$, $n_{Mo+1} = 5$, $h = 4$, $x_{i-1} = 16$, дістанемо

$$Mo^* = x_{i-1} + \frac{n_{Mo} - n_{Mo-1}}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}} h;$$

$$Mo^* = 16 + \frac{30 - 25}{60 - 25 - 5} 4 = 16 + \frac{5}{30} = 16,17.$$

Отже, $Mo^* = 16,17$.

З графіка $F^*(x)$ визначається медіанний інтервал, який дорівнює 12—16.

Беручи до уваги, що $F(12) = 0,4$, $F(16) = 0,65$, $h = 4$ і застосовуючи (361), дістанемо:

$$Me^* = x_{i-1} + \frac{0,5 - F^*(x_{i-1})}{F^*(x_i) - F^*(x_{i-1})} h = 12 + \frac{0,5 - 0,4}{0,65 - 0,4} 4 = 12 + \frac{0,1}{0,25} 4 = 13,6.$$

Отже, $Me^* = 13,6$.

\bar{x}_B, D_B, σ_B для інтервального статистичного розподілу вибірки. Для визначення \bar{x}_B, D_B, σ_B перейдемо від інтервального розподілу до дискретного, варіантами якого є середина часткових інтервалів $x_i^* = x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}$ і який має такий вигляд:

$x_i^* = x_{i-1} - \frac{h}{2} = x_i - \frac{h}{2}$	x_1^*	x_2^*	x_3^*	...	x_k^*
h_i	h_1	h_2	h_3	...	h_k

Тоді \bar{x}_B, D_B, σ_B обчислюються за формулами:

$$\bar{x}_B = \frac{\sum x_i^* n_i}{h};$$

$$D_B = \frac{\sum (x_i^*)^2 n_i}{h} - (\bar{x}_B)^2;$$

$$\sigma_B = \sqrt{D_B}.$$

Приклад. За заданим інтервальним статистичним розподілом вибірки, в якому наведено розподіл маси новонароджених x_i ,

$X = x_i$, КГ	1—1,2	1,2—	1,4—	1,6—	1,8—	1,8—2	1,8—2	2—2,2	2,4—	2,6—	2,8—	2,8—3	3—3,2
----------------	-------	------	------	------	------	-------	-------	-------	------	------	------	-------	-------

n_i	5	12	18	22	36	24	19	15	11	9	2
-------	---	----	----	----	----	----	----	----	----	---	---

обчислити \bar{x}_B, D_B, σ_B .

Розв'язання. Побудуємо дискретний статистичний розподіл за заданим інтервальним. Оскільки $h = 0,2$, то дістанемо:

$x_i^* = x_i - \frac{h}{2} = x_{i-1} + \frac{h}{2}$	1,1	1,3	1,5	1,7	1,9	2,1	2,3	2,5	2,7	2,9	3,1
h_i	5	12	18	22	36	24	19	15	11	9	2

Беручи до уваги (363), (364), (365) і те, що $n = 173$, дістанемо:

$$\bar{x}_B = \frac{\sum x_i^* n_i}{n} = \frac{5,5 + 15,6 + 27 + 37,4 + 68,4 + 50,4 + 43,7}{173} + \frac{37,5 + 29,7 + 26,1 + 6,2}{173} = \frac{347,5}{173} \approx 2,008671 \text{ кг.}$$

Отже, $\bar{x}_B = 2,008671 \text{ кг.}$

$$\frac{\sum (x_i^*)^2 n_i}{n} = \frac{6,05 + 20,29 + 40,5 + 63,58 + 129,96 + 105,84 + 100,51}{173} + \frac{93,75 + 80,19 + 75,69 + 19,22}{173} = \frac{735,58}{173} = 4,251908.$$

$$D_B = \frac{\sum (x_i^*)^2 n_i}{n} - (\bar{x}_B)^2 = 4,251908 - (2,008671)^2 = 4,251908 - 4,034759 = 0,217149.$$

$$D_B = 0,217149.$$

$$\sigma_B = \sqrt{D_B} = \sqrt{0,217149} \approx 0,466.$$

Отже, $\sigma_B = 0,466 \text{ кг.}$

Тема № 2: Варіаційні ряди та їх характеристики. Перевірка статистичних гіпотез.

Лекція 1 за темою №2. Варіаційні ряди та їх характеристики. Дискретні варіаційні ряди.

Більшість величин у статистичних дослідженнях приймають неоднакові значення в різних елементів досліджуваної статистичної сукупності, тобто варіюють. З метою вивчення варіювання й установлення закономірностей, яким підкоряється досліджуване явище, проводять спостереження й одержують значення ознаки в кожного елемента сукупності.

Позначимо через X - досліджувану ознаку, через x_i - спостережувані значення ознаки, $i = 1, 2, \dots, n$ де n - об'єм досліджуваної сукупності.

Різні значення ознаки, що спостерігаються в елементів вибіркової сукупності називаються *варіантами* v_1, v_2, \dots, v_k , а числа, що показує, скільки разів зустрічається кожний варіант називаються *частотами* варіантів f_1, f_2, \dots, f_k . Ясно, що завжди виконується умова $k \leq n$.

Невпорядкованість інформації, що втримується в наведених статистичних даних, утрудняє їхнє використання для подальшого аналізу. Тому дані спостережень піддають первинній обробці, що складається в угрупованні сукупності по варіантах.

Статистичний ряд розподілу - це впорядкований розподіл одиниць сукупності на групи по певній ознаці, що варіює.

Розташуємо значення, що спостерігалися, в порядку їх зростання. Ця операція називається ранжируванням даних спостережень.

Дискретним варіаційним рядом або **рядом розподілу частот** називається ранжируваний ряд варіантів із відповідними їм частотами.

Варіаційним рядом або **рядом розподілу** називається, ранжируваний у порядку зростання або убуття ряд варіантів з відповідними їм частотами.

Варіаційний ряд називається **дискретним**, якщо будь-які його варіанти відрізняються на деяку величину, і – **інтервальним (безперервним)**, якщо варіанти можуть відрізнятися один від іншого на як завгодно малу величину.

Кожному варіанту можна поставити у відповідність не частоту, а одну з наступних величин: відносну частоту, накопичену частоту, накопичену відносну частоту.

Відотною частотою (частотістю) j -го варіанта називається відношення його частоти до об'єму сукупності, тобто величина $w_j = \frac{f_j}{n}$, $j = 1, 2, \dots, k$.

Відносна частота j -го варіанта визначає частку (питома вага) елементів у сукупності, значення ознаки в яких дорівнюють значенню v_j .

Для частот і відносних частот виконуються рівності:

$$\sum_{i=1}^k f_i = n, \quad \sum_{i=1}^k w_i = 1.$$

Накопиченою частотою j -го варіанта називається сума частоти даного варіанта й частот всіх попередніх йому варіантів: $f_j^c = f_1 + f_2 + \dots + f_j$.

Накопичена частота j -го варіанта показує, скільки елементів вибіркової сукупності мають значення ознаки, менше або рівне значенню цього варіанта.

Накопиченою відотною частотою j -го варіанта називається сума відносних частот всіх попередніх йому варіантів і відносною частоти даного варіанта: $w_j^c = w_1 + w_2 + \dots + w_j$.

Накопичена відносна частота j -го варіанта показує частку тих елементів сукупності, у яких значення досліджуваної ознаки менше або дорівнює значенню цього варіанта.

Накопичена відносна частота може також визначатися як відношення накопиченої частоти до загального числа спостережень, тобто $w_j^c = \frac{f_j^c}{n}$.

Результати побудови дискретних варіаційних рядів можна представити у вигляді таблиці 1.

Таблиця 1
Дискретні варіаційні ряди

№	Варіант v_i	Частота f_i	Відносна частота w_i	Накопичена частота f_i^c	Накопичена відносна частота w_i^c
1	v_1	f_1	$w_1 = \frac{f_1}{n}$	f_1	w_1
2	v_2	f_2	$w_2 = \frac{f_2}{n}$	$f_1 + f_2$	$w_1 + w_2$
...

характеристик розподілу, обчислених з використанням варіантів і їхніх частот. Ці характеристики повинні відбивати властиві досліджуваної сукупності закономірності й тенденції.

Для опису статистичних розподілів звичайно використовуються 4 види характеристик: 1) середні величини або характеристики центральної тенденції; 2) характеристики мінливості (варіації) ознаки; 3) характеристики, що відбивають додаткові особливості розподілів, зокрема їхню форму, 4) характеристики положення окремого спостереження в ряді розподілу.

Середні величини

Середня характеризує типовий для сукупності розмір ознаки, тобто центральну тенденцію в розподілі. Практичне використання такої характеристики доцільно в тому випадку, коли окремі варіанти ряду розподіли концентруються поблизу деякого значення. Якщо ж сукупність неоднорідна, результати спостережень значно відрізняються друг від друга й не виявляють загальної тенденції, то використання середньої стає формальним.

Середні величини мають ту ж розмірність, що й досліджувана ознака. Існують різні форми середніх величин: середня арифметична, середня геометрична, середня гармонійна, середня квадратична й т.д.

Середня арифметична. Найпоширенішим видом середньої є *середня арифметична*, котра обчислюється по формулі

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

де x_i , $i = 1, 2, \dots, n$ – елемент сукупності, n – її об'єм.

Якщо за спостереженнями складений варіаційний ряд, то варто використовувати наступну формулу:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k v_i f_i$$

де v_i – варіанти ознаки в дискретному або середини інтервалів в інтервальному ряді, f_i – відповідні частоти, k – кількість варіантів або інтервалів, n – об'єм сукупності. Величини f_i у цій формулі часто називають *вагами*, а саму характеристику, обчислену по формулі (2.2) – *зваженою*.

Якщо за даними спостережень побудований дискретний варіаційний ряд, то формули (2.1) і (2.2) дають однакові значення середньої арифметичної. Якщо ж побудований інтервальный ряд, то середні арифметичні, обчислені по формулах (2.1) і (2.2), як правило, не збігаються, тому що у формулі (2.2) значення ознаки усередині кожного інтервалу приймаються рівними серединам інтервалів. Однак помилка, що виникає в результаті такої заміни, буде мала, якщо спостереження розподілені рівномірно усередині кожного інтервалу й не скапливаються до однойменних границь інтервалів (тобто або все до нижніх границь, або все до верхніх границь).

Середня арифметична застосовується в тому випадку, коли сума спостережень повинна залишитися незмінною, якщо кожне з них замінити середньою арифметичною.

Розглянемо деякі властивості середньої арифметичної.

1. *Сума добутків відхилень варіантів від середньої арифметичної на відповідні частоти дорівнює нулю:*

$$\sum_{i=1}^k (x_i - \bar{x}) f_i = 0.$$

Дійсно,

$$\sum_{i=1}^k (x_i - \bar{x}) f_i = \sum_{i=1}^k x_i f_i - \sum_{i=1}^k \bar{x} f_i = n \cdot \bar{x} - n \cdot \bar{x} = 0,$$

що й було потрібно довести.

2. Середня арифметична постійної величини дорівнює цієї постійної.

3. Якщо всі варіанти збільшити (зменшити) на одне й теж число, то середня арифметична збільшиться (зменшиться) на те ж число:

$$\overline{(x - c)} = \frac{1}{n} \sum_{i=1}^k (x_i - c) f_i = \frac{1}{n} \sum_{i=1}^k x_i f_i - \frac{1}{n} \sum_{i=1}^k c f_i = \frac{1}{n} \sum_{i=1}^k x_i f_i - c \frac{n}{n} = \bar{x} - c.$$

4. Якщо всі варіанти збільшити (зменшити) в одне й теж число раз, то середня арифметична збільшиться (зменшиться) у стільки ж раз:

$$\overline{(xc)} = \frac{1}{n} \sum_{i=1}^k (x_i c) f_i = c \frac{1}{n} \sum_{i=1}^k x_i f_i = c \bar{x}; \quad \overline{\left(\frac{x}{c}\right)} = \frac{1}{n} \sum_{i=1}^k \left(\frac{x_i}{c}\right) f_i = \frac{\frac{1}{n} \sum_{i=1}^k x_i f_i}{c} = \frac{\bar{x}}{c}.$$

5. Якщо ряд спостережень складається із двох груп, то середня арифметична всього ряду дорівнює зваженої середньої арифметичної групових середніх, причому вагами є об'єми груп:

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2},$$

де n_1, n_2 – об'єми груп; \bar{x}_1, \bar{x}_2 – середні арифметичні 1-й і 2-й груп спостережень.

Аналогічна властивість виконується також у тому випадку, коли ряд спостережень складається з m груп спостережень.

Середня геометрична. Середньої геометричної \bar{x}_g називається корінь n -й ступеня з добутку значень x_1, x_2, \dots, x_n :

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i} \dots$$

Якщо за спостереженнями побудований варіаційний ряд, то застосовують формулу:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^k (v_i)^{f_i}},$$

Середню геометричну варто застосовувати в тих випадках, коли при заміні нею кожного спостереження необхідно зберегти незмінним добуток спостережень.

Поряд із середніми як характеристики центра розподілу застосовуються так звані структурні середні - мода й медіана.

Мода. Модой називається значення ознаки, якому відповідає найбільша частота, тобто яке спостерігалось найбільше число раз.

Для дискретного ряду мода перебуває безпосередньо з розподілу частот.

Якщо складено інтервальный варіаційний ряд, то мода обчислюється по наступній наближеній формулі:

$$\bar{x}_{Mo} = x_0 + h \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})},$$

де x_0 – початок модального інтервалу, тобто інтервалу, що має максимальну частоту, h – довжина модального інтервалу, f_i – частота модального інтервалу, f_{i-1} і f_{i+1} – частоти відповідно попереднього й наступного за модальним інтервалів.

Розподіл, що має одну моду називається *унимодальним*; дві моди – *бімодальним*; три й більше моди – *мультимодальним*.

Медіана. Медіаною називається значення ознаки, що доводиться на середину ранжируваного ряду спостережень.

Якщо проведено непарне число спостережень, тобто $n = 2m + 1$, $m \in Z$, і результати спостережень розміщені по зростанню, то медіана дорівнює:

$$\bar{x}_{Me} = x_{m+1}.$$

Якщо проведено парне число спостережень, тобто $n = 2m$, $m \in Z$, то на середину ранжируваного ряду доводяться значення x_m й x_{m+1} . У цьому випадку як медіана приймають середнє арифметичне двох серединних елементів, тобто

$$\bar{x}_{Me} = \frac{x_m + x_{m+1}}{2}.$$

Якщо складено інтервальний варіаційний ряд, то медіана обчислюється по наступній формулі:

$$\bar{x}_{Me} = x_0 + h \frac{n/2 - T_{i-1}}{f_i},$$

де x_0 – початок медіанного інтервалу, тобто інтервалу, у якому накопичена відносна частота вперше перевищує значення 0.5, h – довжина медіанного інтервалу, n – об'єм вибірки, T_{i-1} – сума частот інтервалів, що передують медіанному інтервалу; f_i – частота медіанного інтервалу.

Очевидно, що середня реагує на кожну зміну ряду, у той час як медіана й мода – тільки на деякі. Із цієї причини про середню говорять як про показник, у якому знаходить своє відбиття весь ряд розподілу й показують наскільки великі у своїй масі елементи сукупності. Однак чутливість середньої може позначитися негативно на показності цього показника, якщо серед спостережень є трохи таких, які істотно відрізняються від інших в одному напрямку (або набагато більше, або набагато менше). У цьому випадку більше змістовним і правильно відбиває центр розподілу показником буде медіана, тому що на неї не роблять впливу значення ознаки на "хвостах" розподілу.

Слід зазначити, що середня арифметична й медіана можуть не бути елементами розподілу, у той час як модою обов'язково є одне або кілька значень досліджуваної ознаки.

Лекція 2 за темою №2. Інтервальні варіаційні ряди. Розрахунок числових характеристик.

Якщо кількість варіантів k занадто велике або близько до об'єму вибірки, то дискретний варіаційний ряд не використовується для проведення аналізу варіації ознаки. У цьому випадку доцільно скласти варіаційний ряд по інтервалах значень досліджуваної ознаки, тобто інтервальний варіаційний ряд.

Для побудови інтервального варіаційного ряду весь діапазон зміни ознаки, від мінімального x_{\min} до максимального x_{\max} , розбивають на певне число рівних або нерівних інтервалів. Потім підраховують число елементів сукупності, значення ознаки яких попадає в той або інший інтервал, тобто обчислюють частоти влучень значень ознаки в інтервал. Число інтервалів, як правило, вибирають від 7 до 16, так, щоб у кожний інтервал попадало не менш 5 % всіх спостережень.

Якщо число інтервалів важко визначити заздалегідь, то для розрахунку величини рівних інтервалів при достатньому об'ємі сукупності може бути використана формула Стерджесса :

$$h = \frac{x_{\max} - x_{\min}}{1 + 3,322 \lg n}.$$

Якщо h виявляється дробовим числом, то за величину інтервалу варто взяти або

найближче ціле число, або найближчу "гарну" дріб.

За початок першого інтервалу рекомендується приймати величину, приблизно рівну $a_1 = x_{\min} - \frac{h}{2}$. Побудова інтервалів продовжують доти, поки початок наступного один по одному інтервалу не буде рівним або більшим x_{\max} .

Тому що деякі значення ознаки можуть збігатися із границями інтервалів, то в кожний інтервал включаються варіанти більші, ніж нижня границя інтервалу й менші або рівні верхній границі інтервалу. Інакше кажучи, граничне значення варто відносити до інтервалу, у якого дане значення є верхньою границею.

Як і для дискретного розподілу, інтервальний варіаційний ряд можна перетворити в інтервальні ряди відносних частот, накопичених частот і накопичених відносних частот. Крім того, інтервальний ряд може бути умовно перебудований у дискретний шляхом заміни кожного інтервалу його серединою.

У загальному виді інтервальний варіаційний ряд можна представити у вигляді таблиці 2.

Таблиця 2

Інтервальний варіаційний ряд

№	Нижня границя інтервалу	Верхня границя інтервалу	Середина інтервалу v_i	Частота f_i	Відносна частота w_i	Накопичена частота f_i^c	Накопичена відносна частота w_i^c
1	a_1	b_1	$v_1 = \frac{a_1 + b_1}{2}$	f_1	$w_1 = \frac{f_1}{n}$	f_1	w_1
2	a_2	b_2	$v_2 = \frac{a_2 + b_2}{2}$	f_2	$w_2 = \frac{f_2}{n}$	$f_1 + f_2$	$w_1 + w_2$
...
s	a_s	b_s	$v_s = \frac{a_s + b_s}{2}$	f_s	$w_s = \frac{f_s}{n}$	$f_1 + f_2 + \dots + f_s$	$w_1 + w_2 + \dots + w_s$
...
k	a_k	b_k	$v_k = \frac{a_k + b_k}{2}$	f_k	$w_k = \frac{f_k}{n}$	$\sum_{j=1}^k f_j$	$\sum_{j=1}^k w_j = 1$

Характеристики варіації

Розглянуті вище характеристики центра статистичної сукупності тим більше характерні для даного розподілу, чим ближче групуються спостереження навколо середньої арифметичної, тобто чим менш вони розсіяні. Тому середні характеристики повинні бути доповнені виміром варіації ознаки щодо середньої, тобто характеристиками мінливості значень, що спостерігалися, ознаки.

Мірою варіації повинне бути число, що обчислюється на основі елементів ряду розподілу й задовольняє наступним вимогам:

- значення даного показника повинне бути малим, якщо елементи ряду мало відрізняються друг від друга, і повинне бути більшим, якщо елементи ряду сильно розсіяні;

- значення показника не повинне залежати від числа елементів ряду, тобто воно не

повинне зростати тільки в результаті збільшення числа елементів ряду;

- показник не повинен залежати від середньої, а відбивати розкид елементів навколо її.

Найпростішим вимірником варіації є розмах варіювання R , тобто різниця між найбільшим і найменшим значенням ознаки: $R = x_{max} - x_{min}$. Розмах варіювання дає лише наближене подання про варіацію ознаки й, крім того, на крайні значення ряду розподілу можуть впливати різні випадкові фактори, що робить їх досить ненадійними.

Вище відзначалося, що найбільший інтерес представляє угруповання значень ознаки біля середньої арифметичної. Відхилення варіантів від середньої арифметичної визначають різниці $(x_i - \bar{x})$, а ваги варіантів – як часто мають місце ці різниці в даному розподілі. Однак сума добутоків відхилень на їхні ваги не може бути мірою розсіювання ознаки, тому що по доведеній вище теоремі ця сума завжди дорівнює нулю. Для усунення впливу знака відхилень переходять або до абсолютних величин відхилень, або до квадратів відхилень, одержуючи при цьому різні характеристики варіації ознаки.

Середнє лінійне відхилення. Середнім лінійним відхиленням варіаційного ряду називається середня арифметична абсолютних величин відхилень варіантів від них середньої арифметичної:

$$d = \frac{1}{n} \sum_{i=1}^k |v_i - \bar{x}| \cdot f_i,$$

де v_i , $i = 1, 2, \dots, k$ – варіанти ознаки в дискретному або середини інтервалів в інтервальному ряду, f_i – відповідні частоти, \bar{x} – середня арифметична, k – кількість варіантів або інтервалів, n – об'єм сукупності.

Дисперсія вибірки. Дисперсією \bar{S}^2 називається середня арифметична квадратів відхилень варіантів від них середньої арифметичної:

$$\bar{S}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2;$$

де x_i , $i = 1, 2, \dots, n$ – елемент сукупності, n – її об'єм.

Якщо за спостереженнями побудований варіаційний ряд, то застосовують формулу:

$$\bar{S}^2 = \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^2 f_i = \frac{1}{n} \sum_{i=1}^k v_i^2 f_i - (\bar{x})^2,$$

де v_i , $i = 1, 2, \dots, k$ – варіанти ознаки в дискретному або середини інтервалів в інтервальному ряду, f_i – відповідні частоти, k – кількість варіантів або інтервалів, n – об'єм сукупності.

Розглянемо деякі властивості дисперсії.

1. Дисперсія постійної величини дорівнює нулю.

2. Якщо всі варіанти збільшити (зменшити) на одне й теж число, то дисперсія не зміниться:

$$\bar{S}_{v+c}^2 = \frac{1}{n} \sum_{i=1}^k [(v_i + c) - (\bar{x} + c)]^2 f_i = \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^2 f_i = \bar{S}^2.$$

3. Якщо всі варіанти збільшити (зменшити) в одне й теж число c раз, то дисперсія збільшиться (зменшиться) в c^2 разів:

$$\bar{S}_{cv}^2 = \frac{1}{n} \sum_{i=1}^k (cv_i - c\bar{x})^2 f_i = \frac{1}{n} \sum_{i=1}^k c^2 (v_i - \bar{x})^2 f_i = c^2 \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^2 f_i = c^2 \bar{S}^2;$$

$$\bar{S}_{v/c}^2 = \frac{1}{n} \sum_{i=1}^k \left(\frac{1}{c} v_i - \frac{1}{c} \bar{x} \right)^2 f_i = \frac{1}{n} \sum_{i=1}^k \frac{1}{c^2} (v_i - \bar{x})^2 f_i = \frac{1}{c^2} \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^2 f_i = \frac{\bar{S}^2}{c^2}.$$

4. Якщо ряд спостережень складається із двох груп, то дисперсія всього ряду дорівнює сумі зваженої середньої арифметичної групових дисперсій і зваженого середньої арифметичної квадратів відхилень групових середніх від середньої всього ряду, причому вагами є об'єми груп:

$$\bar{S}^2 = \frac{n_1 \bar{S}_1^2 + n_2 \bar{S}_2^2}{n_1 + n_2} + \frac{n_1 (\bar{x}_1 - \bar{x}) + n_2 (\bar{x}_2 - \bar{x})}{n_1 + n_2},$$

де n_1, n_2 – об'єми груп; $\bar{x}_1, \bar{x}_2, \bar{S}_1^2, \bar{S}_2^2$ – середні арифметичні й дисперсії 1- й і 2-й груп спостережень відповідно.

Середнє квадратическое (стандартне) відхилення. При обчисленні дисперсії підсумуються квадрати відхилень варіантів, у силу чого дисперсія вимірюється у квадратах тих одиниць, у яких вимірюється досліджувана ознака. Для того, щоб характеристика варіації виражалася в тих же одиницях, що й значення ознаки, використовують корінь квадратний з дисперсії.

Середнім квадратическим (стандартним) відхиленням називається арифметичне значення кореня квадратного з дисперсії:

$$\bar{S} = +\sqrt{\bar{S}^2}.$$

Коефіцієнт варіації. Для порівняльної оцінки варіації в розподілах з різними значеннями середньої використовується також коефіцієнт варіації, рівний вираженому у відсотках відношенню середнього квадратического відхилення до середній арифметичного:

$$V = \frac{\bar{S}}{\bar{x}} 100\%.$$

Коефіцієнт варіації дозволяє визначити, наскільки добре середня арифметична представляє всі елементи статистичної сукупності. Якщо $V < 33\%$, то статистична сукупність є однорідною й середня буде відбивати типовий розмір досліджуваної ознаки. Якщо коефіцієнт варіації $V > 33\%$, те, як правило, можна зробити висновок про неоднорідність статистичної сукупності й середня арифметична не є типовим значенням для всіх елементів. Однак коефіцієнт варіації втрачає зміст при $\bar{x} = 0$ і стає малонадійним при близьких до нуля значеннях середньої.

Моменти розподілу

Узагальненням понять середньої арифметичної й дисперсії варіаційного ряду є поняття моментів розподілу, уперше запропонованих П.Л.Чебышевым.

Початковим моментом порядку m варіаційного ряду називається середня арифметична m -х ступенів варіантів, тобто

$$\tilde{v}_m = \frac{1}{n} \sum_{i=1}^k v_i^m \cdot f_i,$$

де v_i – варіанти ознаки в дискретному або середини інтервалів в інтервальному ряді, f_i – відповідні частоти, k – кількість варіантів або інтервалів, n – об'єм сукупності.

Початковий момент нульового порядку \tilde{v}_0 дорівнює одиниці для будь-якого розподілу. Початковий момент першого порядку \tilde{v}_1 дорівнює середньої арифметичної варіаційного ряду

Центральним моментом порядку m називається середня арифметична m -х ступенів

відхилень варіантів від їх середньої арифметичної, тобто

$$\tilde{\mu}_m = \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^m f_i.$$

Центральний момент нульового порядку дорівнює одиниці для будь-якого розподілу. Центральний момент першого порядку дорівнює нулю для будь-якого розподілу (у силу теореми про суму відхилень)

Центральний момент другого порядку має вигляд

$$\tilde{\mu}_2 = \frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^2 f_i,$$

т.е. це дисперсія варіаційного ряду.

Моменти розподілу являють собою систему числових характеристик, за допомогою яких можна описати всі особливості варіації ознаки. Чим більше моментів для даного розподілу обчислено, тим точніше можна описати його властивості. Однак з ростом порядку моментів росте вплив випадкових помилок у статистичних даних, тому на практиці використовуються моменти до четвертого порядку.

Характеристики форми розподілу

Якщо полігон варіаційного ряду скошений у ту або іншу сторону від середньої арифметичної, то такий ряд називають асиметричним. Як міра асиметрії варіаційного ряду використовується коефіцієнт асиметрії.

Коефіцієнтом асиметрії варіаційного ряду називається відношення центрального моменту третього порядку до куба середнього квадратического відхилення:

$$As = \frac{\tilde{\mu}_3}{\bar{S}^3} = \frac{\frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^3 \cdot f_i}{\bar{S}^3}.$$

де v_i – варіанти ознаки в дискретному або середини інтервалів в інтервальному ряді, f_i – відповідні частоти, k – кількість варіантів або інтервалів, n – об'єм сукупності.

Якщо коефіцієнт асиметрії $As < 0$, то в цьому випадку має місце лівостороння асиметрія: у варіаційному ряді переважають варіанти менші, ніж середня. Якщо ж у варіаційному ряді переважають значення більші, ніж середня, то має місце правобічна асиметрія й $As > 0$. Для симетричного розподілу варіанти, равноудаленные від середньої \bar{x} , мають однакові частоти, і тому $\tilde{\mu}_3 = 0$, отже, $As = 0$.

Ще одна особливість форми розподілу вимірюється за допомогою коефіцієнта крутості або ексцесу розподілу, що характеризує крутість (загостреність) графіка розподілу.

Ексцесом або *коефіцієнтом островершинності* варіаційного ряду називається зменшене на три одиниці відношення центрального моменту четвертого порядку до четвертого ступеня середнього квадратического відхилення:

$$Ek = \frac{\tilde{\mu}_4}{\bar{S}^4} - 3 = \frac{\frac{1}{n} \sum_{i=1}^k (v_i - \bar{x})^4 \cdot f_i}{\bar{S}^4} - 3.$$

Ексцес характеризує відносну остроконечность або сглаженность розподілу в порівнянні з нормальним розподілом, у якого він дорівнює нулю. Позитивний ексцес позначає відносно гострий розподіл. Негативний ексцес позначає відносно згладжений розподіл.

Характеристики положення окремого спостереження в ряді розподілу

У різних дослідженнях виникає необхідність описати положення деякого спостереження в ряді розподілу, не вказуючи загального числа спостережень. Із цією метою окремому спостереженню x_i ставиться у відповідність величина, називана процентильним рангом.

Процентильний ранг спостереження x_i дорівнює відсотку $q\%$ тих спостережень, значення яких менше або рівні значення x_i .

Процентильний ранг дозволяє зрівняти положення спостережень у різних рядах розподілу, однак не дає відповідь на деякі питання. Наприклад, якщо два спостереження з різних сукупностей мають процентильний ранг, рівний 100%, то їхні значення найбільші у своїх сукупностях. Однак одне з них може набагато перевищувати інші значення, а інше мало відрізнятися від них.

Розглянемо спосіб визначення положення елемента в ряді розподілу, що вказує, як далеко й по яку сторону даний елемент перебуває від середньої.

Відношення відхилення елемента розподілу від середньої до середнього квадратическому відхиленню називається *z-оцінкою* даного елемента, тобто

$$z = \frac{x_i - \bar{x}}{\bar{S}}$$

Таким чином, *z-оцінка* елемента розподілу показує, на скільки одиниць середнього квадратического відхилення даний елемент більше або менше середньої арифметичної. Знак *z-оцінки* показує, по яку сторону від середньої розташований даний елемент.

Побудова *z-оцінок* є одним зі способів *стандартизації (нормировки)* статистичних даних і дозволяє порівнювати елементи різних розподілів.

Властивості *z-оцінок*.

1. Середня ряду розподілу *z-оцінок* дорівнює нулю, тобто $\bar{z} = 0 = 0$.

Нехай $z_i = \frac{x_i - \bar{x}}{\bar{S}}$. Тоді

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{\bar{S}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})}{\bar{S}} = \frac{\frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} n \bar{x}}{\bar{S}} = \frac{\bar{x} - \bar{x}}{\bar{S}} = 0.$$

2. Середнє квадратическое відхилення ряду розподілу *z-оцінок* дорівнює одиниці, тобто $\bar{S}_z^2 = 1$.

Знайдемо дисперсію ряду розподілу *z-оцінок*:

$$\bar{S}_z^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\bar{S}_x} \right)^2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\bar{S}_x^2} = \frac{\bar{S}_x^2}{\bar{S}_x^2} = 1.$$

Тоді

$$\bar{S}_z = +\sqrt{\bar{S}_z^2} = 1.$$

Лекція 3 за темою №2. Перевірка статистичних гіпотез.

Інформація, яку дістають на підставі вибірки, реалізованої із генеральної сукупності, може бути використана для формулювання певних суджень про всю генеральну сукупність. Наприклад, розпочавши виготовляти покришки нового типу для автомобілів, відбирають певну кількість цих покришок і піддають їх певним тестам.

За результатами тестів можна зробити висновок про те, чи кращі нові покришки від покришок старого типу, чи ні. А це, у свою чергу, дає підставу для прийняття рішення:

виготовляти їх чи ні.

Такі рішення називають *статистичними*.

Статистичні рішення мають ймовірнісний характер, тобто завжди існує ймовірність того, що прийняті рішення будуть помилковими.

Головна цінність прийняття статистичних рішень полягає в тому, що в межах ймовірнісних категорій можна об'єктивно виміряти ступінь ризику, що відповідає тому чи іншому рішення.

Будь-які статистичні висновки, здобуті на підставі обробки вибірки, називають *статистичними гіпотезами*.

Параметричні і непараметричні статистичні гіпотези

Статистичні гіпотези про значення параметрів ознак генеральної сукупності називають *параметричними*.

Наприклад, висувається статистична гіпотеза про числові значення генеральної середньої $\bar{X}_Г$, генеральної дисперсії $D_Г$, генерального середнього квадратичного відхилення $\sigma_Г$ та ін.

Статистичні гіпотези, що висуваються на підставі обробки вибірки про закон розподілу ознаки генеральної сукупності, називаються *непараметричними*. Так, наприклад, на підставі обробки вибірки може бути висунута гіпотеза, що ознака генеральної сукупності має нормальний закон розподілу, експоненціальний закон та ін.

Нульова й альтернативна гіпотези

Гіпотезу, що підлягає перевірці, називають *основною*. Оскільки ця гіпотеза припускає відсутність систематичних розбіжностей (нульові розбіжності) між невідомим параметром генеральної сукупності і величиною, що одержана внаслідок обробки вибірки, то її називають *нульовою гіпотезою* і позначають H_0 .

Зміст нульової гіпотези записується так:

$$H_0 : \bar{x}_Г = a ;$$

$$H_0 : \sigma_Г = 2 ;$$

$$H_0 : r_{xy} = 0,95 .$$

Кожній нульовій гіпотезі можна протиставити кілька альтернативних (конкуруючих) гіпотез, які позначають символом H_α , що заперечують твердження нульової. Так, наприклад, нульова гіпотеза стверджує: $H_0 : \bar{x}_Г = a$, а альтернативна гіпотеза — $H_\alpha : \bar{x}_Г > a$, тобто заперечує твердження нульової.

Прості і складні статистичні гіпотези

Проста гіпотеза, як правило, належить до параметра ознак генеральної сукупності і є однозначною.

Наприклад, згідно з простою гіпотезою параметр генеральної сукупності дорівнює конкретному числу, а саме:

$$H_0 : \bar{x}_Г = 4 ;$$

$$H_0 : \sigma_Г = 4 .$$

Складна статистична гіпотеза є неоднозначною. Вона може стверджувати, що значення параметра генеральної сукупності належить певній області ймовірних значень, яка може бути дискретною і неперервною.

Наприклад:

$$H_0 : \bar{x}_Г \in [2; 2,1; 2,2] \quad \text{або} \quad H_0 : \bar{x}_Г \in [5,2 \div 6,5] .$$

Нульова гіпотеза може стверджувати як про значення одного параметра генеральної сукупності, так і про значення кількох параметрів, а також про закон розподілу ознаки генеральної сукупності.

Для перевірки правильності висунутої статистичної гіпотези вибирають так званий статистичний критерій, керуючись яким відхиляють або не відхиляють нульову гіпотезу. Статистичний критерій, котрий умовно позначають через K , є випадковою величиною, закон розподілу ймовірностей якої нам заздалегідь відомий. Так, наприклад, для перевірки правильності $H_0: \bar{X}_T = a$ як статистичний критерій K можна взяти випадкову величину, яку позначають через $K = Z$, що дорівнює

$$Z = \frac{\bar{x}_B - a}{\sigma(\bar{x}_B)},$$

і яка має нормований нормальний закон розподілу ймовірностей. При великих обсягах вибірки ($n > 30$) закони розподілу статистичних критеріїв наближатимуться до нормального.

Спостережуване значення критерію, який позначають через K^* , обчислюють за результатом вибірки.

Область прийняття гіпотези. Критична область. Критична точка

Множину Ω всіх можливих значень статистичного критерію K можна поділити на дві підмножини A і \bar{A} , які не перетинаються.

$$(A \cup \bar{A} = \Omega, A \cap \bar{A} = \emptyset).$$

Сукупність значень статистичного критерію $K \in A$, за яких нульова гіпотеза не відхиляється, називають *областю прийняття нульової гіпотези*.

Сукупність значень статистичного критерію $K \in \bar{A}$, за яких нульова гіпотеза не приймається, називають *критичною областю*.

Отже, A — область прийняття H_0 ,

\bar{A} — критична область, де H_0 відхиляється.

Точку або кілька точок, що поділяють множину Ω на підмножини A і \bar{A} , називають *критичними* і позначають через $K_{кр}$.

Існують три види критичних областей:

Якщо при $K < K_{кр}$ нульова гіпотеза відхиляється, то в цьому разі ми маємо лівобічну критичну область, яку умовно можна зобразити (рис. 119).

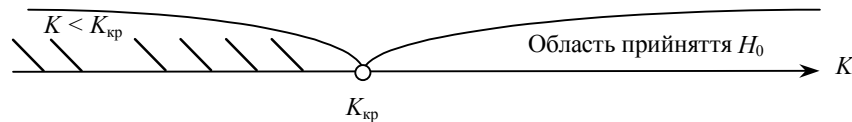


Рис. 119

Якщо при $K > K_{кр}$ нульова гіпотеза відхиляється, то в цьому разі маємо правобічну критичну область (рис. 120).

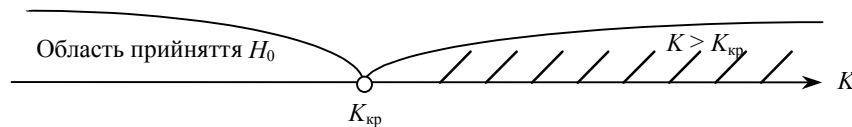


Рис. 120

Якщо ж при $K < K'_{кр}$ і при $K > K''_{кр}$ нульова гіпотеза відхиляється, то маємо двобічну критичну область (рис. 121).



Рис. 121

Лівобічна і правобічна області визначаються однією критичною точкою, двобічна критична область — двома критичними точками, симетричними відносно нуля.

Загальний алгоритм перевірки правильності нульової гіпотези

Для перевірки правильності H_0 задається так званий *рівень значущості* α .
 α — це мала ймовірність, якою наперед задаються. Вона може набувати значення $\alpha = 0,005; 0,01; 0,001$.

В основу перевірки H_0 покладено принцип $P(K \in \bar{A}) = \alpha$, тобто ймовірність того, що статистичний критерій потрапляє в критичну область \bar{A} , дорівнює малій ймовірності α . Якщо ж виявиться, що $K \in \bar{A}$, а ця подія малої ймовірності і все ж відбулася, то немає підстав приймати нульову гіпотезу.

Пропонується такий алгоритм перевірки правильності H_0 :

1. Сформулювати H_0 й одночасно альтернативну гіпотезу H_α .
 2. Вибрати статистичний критерій, який відповідав би сформульованій нульовій гіпотезі.

3. Залежно від змісту нульової та альтернативної гіпотез будується правобічна, лівобічна або двобічна критична область, а саме:

нехай $H_0 : \bar{x}_r = a$, тоді, якщо

$H_\alpha : \bar{x}_r > a$, то вибирається правобічна критична область, якщо

$H_\alpha : \bar{x}_r < a$, то вибирається лівобічна критична область і коли

$H_\alpha : \bar{x}_r \neq a$, то вибирається двобічна критична область.

4. Для побудови критичної області (лівобічної, правобічної чи двобічної) необхідно знайти критичні точки. За вибраним статистичним критерієм та рівнем значущості α знаходяться критичні точки.

5. За результатами вибірки обчислюється спостережуване значення критерію $K_{\text{сп}}^*$.

6. Відхиляють чи приймають нульову гіпотезу на підставі таких міркувань:

у разі, коли $K^* \in \bar{A}$, а це є малої ймовірною випадковою подією, $P(K^* \in \bar{A}) = \alpha$ і, незважаючи на це, вона відбулася, то в цьому разі H_0 відхиляється:
 для лівобічної критичної області

$$P(K_{\text{сп}}^* < K_{\text{кр}}) = \alpha;$$

для правобічної критичної області

$$P(K_{\text{сп}}^* > K_{\text{кр}}) = \alpha;$$

для двобічної критичної області

$$P(K_{\text{сп}}^* < K'_{\text{кр}}) + P(K_{\text{сп}}^* > K''_{\text{кр}}) = \alpha$$

або

$$P(K_{\text{сп}}^* < K'_{\text{кр}}) = P(K_{\text{сп}}^* > K''_{\text{кр}}) = \frac{\alpha}{2},$$

ураховуючи ту обставину, що критичні точки $K'_{\text{кр}}$ і $K''_{\text{кр}}$ симетрично розташовані відносно нуля.

Якою б не була малою величиною α , потрапляння спостережуваного значення $K_{\text{сп}}^*$ у критичну область ($K_{\text{сп}}^* \in \bar{A}$) ніколи не буде подією абсолютно неможливою. Тому не виключається той випадок, коли H_0 буде правильною, а $K_{\text{сп}}^* \in \bar{A}$, а тому нульову гіпотезу буде відхилено.

Отже, при перевірці правильності H_0 можуть бути допущені помилки. Розрізняють при цьому помилки першого і другого роду.

Якщо H_0 є правильною, але її відхиляють на основі її перевірки, то буде допущена помилка першого роду.

Якщо H_0 є неправильною, але її приймають, то в цьому разі буде допущена помилка другого роду.

Між помилками першого і другого роду існує тісний зв'язок.

Нехай, для прикладу, перевіряється $H_0: \bar{X}_\Gamma = a$. При великих обсягах вибірки n \bar{x}_B , як випадкова величина, закон розподілу ймовірностей якої асимптотично наблизитиметься до нормального з числовими характеристиками:

$$M(\bar{x}_B) = a = \bar{X}_\Gamma, \quad \sigma(\bar{x}_B) = \frac{\sigma_\Gamma}{\sqrt{n}}.$$

Тому, коли гіпотеза H_0 є правдивою, $M(\bar{x}_B) = a$. Цей розподіл має такий вигляд (рис. 122, крива $f(x; a)$).

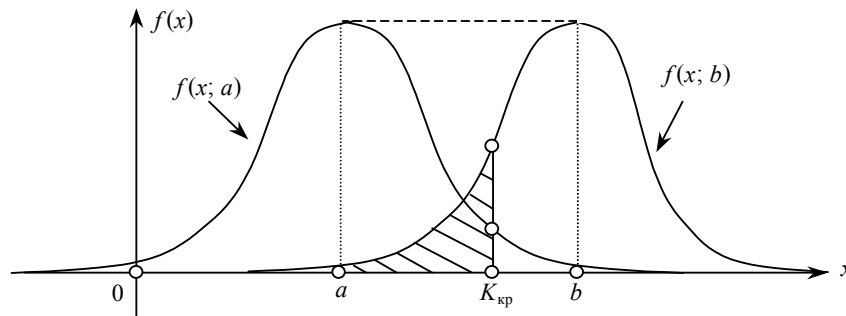


Рис. 122

Коли альтернативна гіпотеза заперечує H_0 і стверджує $H_a: \bar{X}_\Gamma = b > a$, то в цьому разі нормальна крива буде зміщена праворуч (на рис. 122, крива $f(x; b)$).

За вибраним рівнем значущості α визначається критична область (рис. 122).

Коли $\bar{x}_B > K_{кр}$, то H_0 відхиляється з ймовірністю помилки першого роду:

$$P(\bar{x}_B > K_{кр}) = \int_{K_{кр}}^{\infty} f(x; a) dx = \alpha.$$

Коли $\bar{x}_B < K_{кр}$, то H_0 не відхиляється, хоча може бути правильною альтернативна гіпотеза H_a .

Отже, в цьому разі припускаються помилки другого роду.

Ймовірність цієї помилки, яку позначають символом β , може бути визначена на кривій $f(x; b)$, а саме:

$$\beta = \int_{-\infty}^{K_{кр}} f(x; b) dx.$$

Ця ймовірність на рис. 122 показана штрихуванням площі під кривою $f(x; b)$, що міститься ліворуч $K_{кр}$.

Якщо з метою зменшення ризику відхилити правильну гіпотезу H_0 зменшуватимемо значення α , то в цьому разі критична точка $K_{кр}$ зміщуватиметься праворуч, що, у свою чергу, спричинює збільшення ймовірності помилки другого роду, тобто величини β .

Різницю $\pi = 1 - \beta$ називають *ймовірністю обґрунтованого відхилення H_0* , або *потужністю критерію*.

Під час розв'язування практичних завдань може виникнути потреба вибору статистичного критерію з їх певної множини. У цьому разі вибирають той критерій, якому притаманна найбільша потужність.

Перевірка правильності нульової гіпотези про значення генеральної середньої

Для перевірки правильності $H_0: \bar{X}_\Gamma = a$, ($M(x) = a$), де «a» є певним числом, при заданому рівні значущості α насамперед необхідно вибрати статистичний критерій K .

Найзручнішим критерієм для цього типу задач є випадкова величина $Z = Z$, що має нормований нормальний закон розподілу ймовірностей $N(0; 1)$, а саме:

$$Z = \frac{\bar{x}_B - a}{\sigma(\bar{x}_B)} = \frac{\bar{x}_B - a}{\frac{\sigma_\Gamma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{x}_B - a)}{\sigma_\Gamma}.$$

При розв'язуванні такого класу задач можливий один із трьох випадків:

1) при $H_\alpha : \bar{x}_r > a$ — будується правобічна критична область;

Лекція 1 за темою №3. Основи кореляційного та регресійного аналізів.

Однією з найважливіших задач дослідження злочинності є виявлення її зв'язку з факторами, що злочинність обумовлюють.

Яким же чином вирішується ця задача? Перш ніж відповісти на це питання необхідно дати поняття і визначити види взаємозв'язків явищ і процесів.

Усі явища, що мають місце в суспільстві, взаємозалежні між собою та обумовлюють одне інше. Ці взаємозв'язки зустрічаються на кожному кроці. Для того, щоб існувати, людині необхідно харчуватися, вдягатися, мати житло. А для того, щоб одержати продукти харчування, одяг, житло, їх необхідно виготовити, побудувати, а для цього у свою чергу необхідно мати необхідне устаткування, сировину, матеріали.

Взаємозв'язки, що мають місце в суспільстві і у природі реалізуються у формі *причинно-наслідкових відносин* між явищами.

Причинно-наслідкові відносини — це зв'язок явищ і процесів, якщо зміни одного з них — причини — веде до зміни іншого — наслідку.

Причина зветься факторна ознака, наслідок — результативна.

Факторними, чи просто *факторами*, називають ознаки, що обумовлюють зміну інших, пов'язаних з ними ознак.

Результативними, чи просто *результатами*, називають ознаки, що змінюються під впливом факторних ознак.

Зв'язки між факторними і результативними ознаками розрізняються за формою і напрямком дії.

За *формою* зв'язки між факторними і результативними ознаками поділяються на функціональні і стохастичні.

Функціональним називається такий зв'язок, при якому певному значенню факторної ознаки відповідає одне і тільки одне значення результативної ознаки.

Найчастіше функціональні зв'язки спостерігаються в явищах, що описуються математикою, фізикою та іншими точними науками.

Так, між температурою повітря і висотою ртутного стовпчика термометра існує функціональний зв'язок: при підвищенні температури повітря рівень ртутного стовпчика зростає, при зниженні — спадає. Величина струму в електричному ланцюзі прямо пропорційна напрузі і обернено пропорційна опорі ланцюга.

Стохастичним називається зв'язок, при якому кожному значенню факторної ознаки X відповідає *не одне, а декілька* значень результативної ознаки Y . Це обумовлено тим, що ознака Y змінюється не тільки під впливом відомої факторної ознаки X , а також під впливом не контролюємих (випадкових) факторних ознак.

Прикладом стохастичного зв'язку є залежність між величиною X — зростом людини і випадковою величиною Y — вагою людини. Кожному значенню величини X (наприклад, $X = 178$ см) відповідає кілька значень величини Y (наприклад, 55 кг, 60 кг, 78 кг, 80 кг, 86 кг, 92 кг і т.д.).

Якщо в стохастичному зв'язку замінити декілька значень результативної ознаки середньою ознакою \bar{Y} , то з'являється різновид стохастичного зв'язку — *кореляційний зв'язок*.

Кореляційним називається такий зв'язок, при якому певному значенню факторної ознаки відповідає *середнє* значення результативної.

У наведеному вище прикладі факторній ознаці (зросту $X = 178$ см) відповідає середнє значення результативної ознаки (вазі $Y = 78$ кг).

В області злочинності не існує функціональних зв'язків. Злочини — це результат

одночасного впливу великого числа різних факторів. Тому зв'язки, що мають місце між злочинами та факторами, що їх обумовлюють, є *кореляційними*.

Так, наприклад, при дослідженні злочинності має місце велика кількість факторів, що її обумовлюють. Їх нараховується до декількох сотень. Одним з таких факторів є стан алкогольного сп'яніння особи. Але це зовсім не означає, що будь-яка людина в стані алкогольного сп'яніння обов'язково вчинить злочин. Фактор “алкогольне сп'яніння” супроводжує велике число інших факторів, *криміногенних* і *антикриміногенних*, які або можуть підсилювати криміногенний характер фактора “алкогольне сп'яніння”, або – навпаки, послабляти, декриміналізувати його. Але при вивченні великої кількості злочинів виявляється кореляційний зв'язок між вчиненням злочину і станом алкогольного сп'яніння - значна кількість окремих видів злочинів вчиняється в стані алкогольного сп'яніння (убивств і зґвалтувань до 80 %, розбоїв – до 65 %, хуліганств – до 70 % і т.д.). Тобто, між злочинністю та факторами, що її обумовлюють, має місце кореляційний зв'язок.

За напрямком дії зв'язки поділяються на прямі і зворотні.

Прямим називається зв'язок, при якому із збільшенням або зменшенням факторної ознаки відбувається збільшення або зменшення результативної. Так, наприклад, при збільшенні споживання на душу населення алкогольних напоїв відзначається зростання рівня злочинності.

Зворотнім називається зв'язок, при якому результативна і факторна ознаки змінюються в протилежному напрямку, тобто із збільшенням факторної ознаки результативна зменшується та навпаки. Так, наприклад, рівень освіти населення і рівень злочинності завжди знаходяться в зворотній залежності. Така ж залежність має місце і між активністю правоохоронних органів і рівнем злочинності.

При дослідженні кореляційного зв'язку необхідно визначити *ступень його щільності*.

2. Визначення ступеня щільності кореляційного зв'язку.

Для визначення ступеня щільності кореляційного зв'язку між факторною і результативною ознаками використовуються такі показники:

- лінійний коефіцієнт кореляції
- коефіцієнт рангової кореляції Спірмена

Лінійний коефіцієнт кореляції характеризується наступною формулою:

$$r = \frac{\sum (dX \cdot dY)}{\sqrt{\sum dX^2 \cdot \sum dY^2}}$$

де: dX – відхилення факторної ознаки від її середнього значення

dY – відхилення результативної ознаки від її середнього значення

Як приклад визначимо тісноту зв'язку між рівнем незайнятості населення та інтенсивністю злочинності у п'яти районах.

Райони	Кількість осіб, які не працюють і не навчаються, на 1000 населення (X)	Кількість злочинів на 1000 населення (Y)	$dX = X - \bar{X}$	$dY = Y - \bar{Y}$	dX^2	dY^2
1	7	4	-5	-1,2	25	1,44
2	9	3	-3	-2,2	9	4,84

3	12	6	0	0,8	0	0,64
4	14	5	2	-0,2	4	0,04
5	18	8	6	2,8	36	7,84
	$\bar{X} = 12$	$\bar{Y} = 5,2$			$\Sigma dX^2 = 7$ 4	$\Sigma dY^2 = 1$ 4,8

$$r = \frac{(-5) \cdot (-1,2) + (-3) \cdot (-2,2) + 0 \cdot 0,8 + 2 \cdot (-0,2) + 6 \cdot 2,8}{\sqrt{74 \cdot 14,8}} = 0,88$$

Чим більше лінійний коефіцієнт кореляції, тим більше щільність зв'язку між двома ознаками.

Лінійний коефіцієнт кореляції змінюється від 0 до +1 і від 0 до -1. При $R=0$ зв'язок відсутній, а при $R=1$ зв'язок не кореляційний, а функціональний. Знак «+» або «-» свідчить про напрямок зв'язку («+» - прямий, «-» - зворотній).

Щільність зв'язку оцінюється за наступними значеннями лінійного коефіцієнта кореляції:

Значення лінійного коефіцієнта кореляції	Характер зв'язку
$0 \leq r < 0,3$	практично відсутній
$0,3 \leq r < 0,5$	слабкий
$0,5 \leq r < 0,7$	помірний
$0,7 \leq r $	сильний

Таким чином, значення лінійного коефіцієнта кореляції 0,88 свідчить про наявність сильного та прямого зв'язку між рівнем незайнятості та інтенсивністю злочинності неповнолітніх осіб.

Коефіцієнт рангової кореляції Спірмена. визначається за формулою:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)}$$

де: d – різниця рангів, n – кількість значень факторної та результативної ознак.

В нашому прикладі проранжуємо ряд X (визначимо ранги або номери місць від 1 до 5). Так як значення X вже розміщені в порядку зростання, то значення рангів збігаються з номерами районів. Тепер проранжуємо ряд Y (графа 5).

Райони	Кількість населення, яке не працює і не навчається, на 1000 н/л (X)	Кількість злочинів, на 1000 населення (Y)	Ранги X	Ранги Y	Різниця рангів (d)	d ²
1	7	4	1	2	1	1
2	9	3	2	1	1	1
3	12	6	3	4	1	1
4	14	5	4	3	1	1
5	18	8	5	5	0	0
						$\Sigma d^2=4$

Рангу 1 привласнюється значенню “3 злочина”, рангу 2 - “4 злочина”, рангу 3 - “5 злочинів”, рангу 4 - “6 злочинів”, рангу 5 - “8 злочинів”. Тепер визначимо різниці рангів (d), зведемо їх у квадрат (d²) і просумуємо ($\Sigma d^2 = 4$). Розрахуємо коефіцієнт рангової кореляції Спірмена:

$$\rho = 1 - \frac{6 \cdot \Sigma d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 4}{5 \cdot (25 - 1)} = 0,8$$

Цей коефіцієнт, як і лінійний коефіцієнт кореляції, змінюється від +1 до -1. Таким чином, значення коефіцієнта Спірмена 0,8 свідчить про наявність сильного та прямого зв'язку між рівнем незайнятостю населення та інтенсивністю злочинності.

3. Визначення аналітичного вираження кореляційного зв'язку

Дослідження кореляційних зв'язків здійснюється шляхом *кореляційно-регресійного аналізу*, що включає до себе встановлення аналітичного вираження (форми) зв'язку, визначення його щільності і напрямку.

Кореляційний аналіз визначає силу (щільність) статистичного зв'язку, а регресійний – досліджує його форму, тобто функціональну залежність.

У регресійному аналізі заздалегідь мається на увазі наявність причинно-наслідкових відносин між результативною (Y) і факторною (X) ознаками. Рівняння регресії, або статистична модель зв'язку соціально-правових явищ, виражається математичною функцією:

$$\bar{Y}_x = f(x) \text{ – парна регресія}$$

$$\bar{Y}_{x_1, x_2, \dots, x_n} = f(x_1, x_2, \dots, x_n) \text{ – множинна регресія}$$

При дослідженні парної регресії, що характеризує зв'язок між двома ознаками – факторною і результативною, аналітичне вираження зв'язку описується наступними рівняннями:

$$\text{прямої} \quad - \quad \bar{Y}_x = a + bx$$

$$\text{гіперболи} \quad - \quad \bar{Y}_x = a + b \frac{1}{x}$$

$$\text{параболи} \quad - \quad \bar{Y}_x = a + bx + cx^2 \text{ і т.д.}$$

У цих рівняннях a , b і c – параметри рівнянь регресії, параметр a показує усереднений вплив на результативну ознаку неврахованих факторів, параметр b (а в

рівнянні параболи і с) показує, на скількох змінюється в середньому значення результативної ознаки при збільшенні факторної на одиницю виміру.

Установити вид рівняння можна лише досліджуючи зв'язок графічно.

Оцінка параметрів регресії (a, b) здійснюється *методом найменших квадратів*.

Сутність цього методу покладається в визначені таких параметрів a і b , при яких *мінімізується* сума квадратів відхилень *емпіричних (фактичних)* значень результативної ознаки від *теоретичних*, отриманих за обраним рівнянням регресії:

$$S = \sum (Y - \bar{Y}_x)^2 = \min$$

Для лінійної залежності це виглядає так:

$$S = \sum (Y - a - bx)^2 = \min$$

Проводячи математичне перетворення (диференцювання) одержуємо систему нормальних рівнянь:

$$\begin{cases} n a + b \sum X = \sum Y \\ a \sum X + b \sum X^2 = \sum XY \end{cases}$$

де: n – обсяг досліджуємої сукупності.

Вирішуючи цю систему рівнянь можна визначити параметри лінійного рівняння:

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n \sum X^2 - \sum X \sum X} \quad b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - \sum X \sum X}$$

Побудуємо рівняння регресії для зв'язку між ступенем незайнятості та інтенсивністю злочинності неповнолітніх осіб у п'ятих районах (таблиця 1).

Таблиця 1

Район	Кількість н/л осіб, які не працюють і не навчаються на 1000 н/л (x)	Кількість злочинів, вчинених н/л на 10000 населення (y)	X^2	XY	\bar{Y}_x
1	7	4	49	28	3,25
2	9	3	81	27	4,02
3	12	6	144	72	5,2
4	14	5	196	70	5,98
5	18	8	324	144	7,54

$$\sum X = 60$$

$$\sum Y = 26$$

$$\sum X^2 = 794$$

$$\sum XY = 341$$

Система нормальних рівнянь для цього приклада має вигляд:

$$\begin{cases} 5a + b60 = 26 \\ a60 + b794 = 341 \end{cases}$$

Звідси: $a = 0,52$; $b = 0,39$; при цьому $\bar{Y}_x = 0,52 + 0,39X$

Параметр b показує, що підвищення на одиницю кількості неповнолітніх, які не працюють і не вчать, приводить до зростання рівня злочинності неповнолітніх у середньому на 0,39.

Підставляючи значення параметрів a і b у рівняння прямої, знаходимо теоретичні вирівняні значення \bar{Y}_x (табл.1, гр.6).

Параметр b свідчить про те, що підвищення на одиницю коефіцієнта злочинної активності приводить до зростання коефіцієнта злочинної інтенсивності в середньому на 1,58.

