Лекція 23.Статистичне вивчення кореляційного зв'язку випадкових величин.

1. Кореляційний момент, вибірковий коефіцієнт кореляції
2. Емпіричні моменти
3. Коефіцієнт асиметрії A_s^*. Ексцес

1. Кореляційний момент, вибірковий коефіцієнт кореляції

Під час дослідження двовимірного статистичного розподілу вибірки постає потреба з'ясувати наявність зв'язку між ознаками X і Y, який у статистиці називають кореляційним. Для цього обчислюється емпіричний кореляційний момент K_{xy}^* за формулою

$$K_{xy}^* = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} y_i\ x_i\ n_{ij}}{n} - \overline{x} \cdot \overline{y} \qquad (23.1)$$

Якщо $K_{xy}^* = 0$, то кореляційного зв'язку між ознаками X і Y немає. Якщо ж $K_{xy}^* \neq 0$, то цей зв'язок існує.

Отже, кореляційний момент дає лише відповідь на запитання: є зв'язок між ознаками X і Y, чи його немає.

Для вимірювання тісноти кореляційного зв'язку обчислюється вибірковий коефіцієнт кореляції $r_{\text{В}}$ за формулою

$$r_{\text{В}} = \frac{K_{xy}^*}{\sigma_x \sigma_y} \qquad (23.2)$$

Як і в теорії ймовірностей, $|r_{\text{В}}| \le 1, \quad -1 \le r_{\text{В}} \le 1.$

Приклад. За заданим двовимірним статистичним розподілом вибірки ознак X і Y

$Y = y_i$	$X = x_j$				
	10	20	30	40	n_{y_i}
2	—	2	4	4	10
4	10	8	6	6	30
6	5	10	5	—	20
8	15	—	15	10	40
n_{x_j}	30	20	30	20	

потрібно: обчислити K_{xy}^*, $r_{\text{В}}$.

Розв'язання. Щоб обчислити K_{xy}^*, $r_{\text{В}}$ визначимо $\overline{x}, \sigma_x, \overline{y}, \sigma_y$. Оскільки $n = \sum\sum n_{ij} = 100,$ то

$$\overline{x} = \frac{\sum x_j \, n_{x_j}}{n} = \frac{10\cdot 30 + 20\cdot 20 + 30\cdot 30 + 40\cdot 20}{100} =$$

$$= \frac{300+400+900+800}{100} = \frac{2400}{100} = 24.$$

$$\overline{x} = 24.$$

$$\frac{\sum x_j^2 \, n_{x_j}}{n} = \frac{(10)^2\cdot 30 + (20)^2\cdot 20 + (30)^2\cdot 30 + (40)^2\cdot 20}{100} =$$

$$= \frac{3000+8000+27000+32000}{100} = \frac{70000}{100} = 700.$$

$$D_x = \frac{\sum x_j^2 \, n_{x_j}}{n} - (\overline{x})^2 = 700 - (24)^2 = 700 - 576 = 124.$$

$$\sigma_x = \sqrt{D_x} = \sqrt{124} \approx 11{,}14.$$

Отже, $\sigma_x = 11{,}14.$

$$\overline{y} = \frac{\sum y_i \, n_{y_i}}{n} = \frac{2\cdot 10 + 4\cdot 30 + 6\cdot 20 + 8\cdot 40}{100} =$$

$$= \frac{20+120+120+320}{100} = 5{,}8.$$

Отже, $\overline{y} = 5{,}8$.

$$\frac{\sum y_i^2 \, n_{y_i}}{n} = \frac{(2)^2\cdot 10 + (4)^2\cdot 30 + (6)^2\cdot 20 + (8)^2\cdot 40}{100} =$$

$$= \frac{40+480+720+2560}{100} = \frac{3800}{100} = 38.$$

$$D_y = \frac{\sum y_i^2 n_{y_i}}{n} - (\overline{y})^2 = 38 - (5{,}8)^2 = 38 - 33{,}64 = 4{,}36,$$

$$\sigma_y = \sqrt{D_y} = \sqrt{4{,}36} \approx 2{,}1.$$

Для визначення K_{xy}^* обчислюють

$$\sum\sum y_i x_j n_{ij} = 2\cdot 10\cdot 0 + 2\cdot 20\cdot 2 + 2\cdot 30\cdot 4 + 2\cdot 40\cdot 4 + 4\cdot 10\cdot 10 + 4\cdot 20\cdot 8 +$$
$$+4\cdot 30\cdot 6 + 4\cdot 40\cdot 6 + 6\cdot 10\cdot 5 + 6\cdot 20\cdot 10 + 6\cdot 30\cdot 5 + 6\cdot 40\cdot 0 + 8\cdot 10\cdot 15 +$$
$$+8\cdot 20\cdot 0 + 8\cdot 30\cdot 15 + 8\cdot 40\cdot 10 = 0 + 80 + 240 + 320 + 400 + 640 + 720 +$$
$$+960 + 300 + 1200 + 900 + 0 + 1200 + 0 + 3600 + 3200 = 13760.$$

Тоді

$$K_{xy}^* = \frac{\sum\sum y_i x_j n_{ij}}{n} - \overline{x}\cdot\overline{y} = \frac{13760}{100} - 24\cdot 5{,}8 = 137{,}6 - 139{,}2 = -1{,}6.$$

Отже, $K_{xy}^* = -1{,}6,$ а це свідчить про те, що між ознаками X і Y існуватиме від'ємний кореляційний зв'язок.

Для вимірювання тісноти цього зв'язку обчислимо вибірковий коефіцієнт кореляції.

$$r_{\text{В}} = \frac{K_{xy}^{*}}{\sigma_x \sigma_y} = \frac{-1{,}6}{11{,}14 \cdot 2{,}1} = \frac{-1{,}6}{23{,}394} \approx -0{,}068.$$

Отже, $r_{\text{В}} = -0{,}068$, тобто тіснота кореляційного зв'язку між ознаками X та Y є слабкою.

2. Емпіричні моменти

Початкові емпіричні моменти. Середнє зважене значення варіант у степені k (k = 1, 2, 3,…) називають *початковим емпіричним моментом k-го порядку* ν_k^*, який обчислюється за формулою

$$\nu_k^* = \frac{\sum x_k\, n_i}{n} \qquad (23.3)$$

При k = 1 дістанемо початковий момент першого порядку:

$$\nu_1^* = \frac{\sum x_i\, n_i}{n} = \overline{x}_{\text{В}} \qquad (23.4)$$

При k = 2 обчислимо початковий момент другого порядку:

$$\nu_2^* = \frac{\sum x_i^2 n_i}{n} \qquad (23.5)$$

Отже, дисперсію вибірки можна подати через початкові моменти першого та другого порядків, а саме:

$$D_{\text{В}} = \nu_2^* - (\nu_1^*)^2 \qquad (23.6)$$

Центральний емпіричний момент *k*-го порядку. Середнє зважене відхилення варіант у степені k (k = 1, 2, 3,…) називають центральним емпіричним моментом k-го порядку

$$\mu_k^* = \frac{\sum (x_i - \overline{x}_{\text{В}})^k n_i}{n} \qquad (23.7)$$

При k = 1 дістанемо:

$$\mu_1^* = \frac{\sum (x_i - \overline{x}_{\text{В}}) n_i}{n} = \frac{\sum x_i\, n_i}{n} - \overline{x}_{\text{В}} \cdot \frac{\sum n_i}{n} = \overline{x}_{\text{В}} - \overline{x}_{\text{В}} = 0\,.$$

При k = 2 маємо:

$$\mu_2^* = \frac{\sum (x_i - \overline{x}_{\text{В}})^2 n_i}{n} = D_{\text{В}}.$$

На практиці найчастіше застосовуються центральні емпіричні моменти третього та четвертого порядків, що обчислюються за формулами:

$$\mu_3^* = \nu_3^* - 3\nu_2^* \cdot \nu_1^* + 2(\nu_1^*)^2, \qquad (23.8)$$

$$\mu_4^* = \nu_4^* - 4\nu_3^* \cdot \nu_1^* + 6\nu_2^*(\nu_1^*)^2 - 3(\nu_1^*)^4. \qquad (23.9)$$

3. Коефіцієнт асиметрії A_s^*. Ексцес.

Центральний емпіричний момент третього порядку застосовується для обчислення коефіцієнта асиметрії:

$$A_s^* = \frac{\mu_3^*}{\sigma_{\text{В}}^3} \qquad (23.10)$$

Якщо варіанти статистичного розподілу вибірки симетрично розміщені відносно $\overline{x}_{\text{В}}$, то в цьому разі $A_s = 0,$ оскільки $\mu_3^* = 0$.

При $A_s < 0$ варіанти статистичного розподілу $x_i < \overline{x}_{\text{В}}$ переважають варіанти $x_i > \overline{x}_{\text{В}}$. Таку асиметрію називають *від'ємною*. При $A_s > 0$ варіанти $x_j > \overline{x}_{\text{В}}$ переважають варіанти $x_i < \overline{x}_{\text{В}}$, і таку асиметрію називають *додатною*.

Ексцес. Центральний емпіричний момент четвертого порядку застосовується для обчислення ексцесу:

$$E_s^* = \frac{\mu_4^*}{\sigma_{\text{В}}^4} - 3. \tag{23.11}$$

E_s^* , як правило, використовується при дослідженні неперервних ознак генеральних сукупностей, оскільки він оцінює крутизну закону розподілу неперервної випадкової величини порівняно з нормальним. Для нормального закону розподілу, як відомо, $E_s^* = 0$.