

Раздел. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Тема 2. Элементы теории корреляции



План лекции

- 1. Функциональная, статистическая и корреляционная зависимости**
- 2. Корреляционная таблица**
- 3. Уравнение прямой линии регрессии**
- 4. Выборочный коэффициент корреляции**
- 5. Решение типовых задач**

&1. Функциональная, статистическая и корреляционная зависимости

Основные понятия

Зависимость между переменными X и Y называется **функциональной**, если существует функция $y=f(x)$, по которой каждому значению $x \in X$ ставится в соответствии единственное значение $y \in Y$.

Однако не всякую зависимость между X и Y можно представить в виде функции. Иногда одному фиксированному значению X соответствует множество значений Y .

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют **корреляционной**.

Условным средним \bar{x}_y называется среднее арифметическое наблюдавшихся значений X , соответствующих $Y = y$.

Уравнение $\bar{y}_x = f(x)$, называется **выборочным уравнением регрессии** Y на X ; функцию $f(x)$ называют **выборочной регрессией** Y на X , а ее график – **выборочной линией регрессии** Y на X . Аналогично уравнение $\bar{x}_y = \varphi(y)$ называется **выборочным уравнением регрессии** X на Y ; функцию $\varphi(y)$ называют **выборочной регрессией** X на Y , а ее график – **выборочной линией регрессии** X на Y .

&2. Корреляционная таблица

По опытным данным, приведенным в корреляционной таблице, можно судить о форме корреляционной связи между признаками X и Y . С этой целью находятся условные средние \bar{y}_{xi} , соответствующие значениям $x_i = (i = \overline{1, k})$, и \bar{x}_{yj} , соответствующие значениям $y_j = (j = \overline{1, l})$ по формулам

$$\bar{y}_{xi} = \frac{\sum_{j=1}^l y_j n_{ij}}{n_{xi}}; \quad \bar{x}_{yj} = \frac{\sum_{i=1}^k x_i n_{ij}}{n_{yj}}.$$

Эмпирической линией регрессии Y на X (X на Y) называют ломаную линию, соединяющую отрезками точки с координатами $M_i^*(x_i; \bar{y}_{xi}) \left(M_j^*(\bar{x}_{yj}; y_j) \right)$.

Форма полученной таким образом эмпирической ломаной является прообразом формы теоретической зависимости.

&3. Уравнение прямой линии регрессии

Допустим, что количественные признаки X и Y связаны линейной корреляционной зависимостью. В этом случае обе линии регрессии будут прямыми. Тогда $f(x, a_1, \dots, a_n) = a_1x + a_2$, а теоретической кривой Y на X будет прямая

$$\bar{Y}_x = a_1x + a_2 \quad . \quad (*)$$

Угловой коэффициент a_1 прямой линии регрессии Y на X называется *коэффициентом регрессии* Y на X и обычно обозначается $\rho_{yx} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x^2}$. Тогда уравнение (*) можно записать следующим образом:

$$\bar{Y}_x - \bar{y} = \rho_{yx}(x - \bar{x}).$$

Аналогично теоретическое уравнение $\bar{X}_y = b_1 y + b_2$ линейной регрессии X на Y с помощью коэффициента $\rho_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_y^2}$ приводится к виду

$$\bar{X}_y - \bar{x} = \rho_{xy} (y - \bar{y}).$$

Сравнивая коэффициенты регрессии Y на X и X на Y , можно отметить, что они имеют одинаковые знаки (в силу совпадения числителей и положительности знаменателей).

&4. Выборочный коэффициент корреляции

Выборочным коэффициентом корреляции r_B признаков X и Y называется число, равное среднему геометрическому коэффициентов регрессии и имеющее их знак:

$$r_B = \pm \sqrt{\rho_{xy} \rho_{yx}} = \frac{\overline{xy} - \bar{x} \bar{y}}{\sigma_x \sigma_y}.$$

Уравнения регрессии с помощью коэффициента корреляции примут вид:

$$\bar{Y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x} (x - \bar{x});$$

$$\bar{X}_y - \bar{x} = r_B \frac{\sigma_x}{\sigma_y} (y - \bar{y}).$$

Прямые регрессии пересекаются в точке $(\bar{x}; \bar{y})$, которая называется средней точкой корреляционного графика.

Коэффициент корреляции имеет важное самостоятельное значение. С его помощью оценивается теснота (сила) корреляционной связи между признаками. Коэффициент корреляции r_B обладает следующими свойствами:

1. $|r_B| \leq 1$ или $-1 \leq r_B \leq 1$.
2. Условие $|r_B| = 1$ ($r_B = \pm 1$) является необходимым и достаточным условием существования линейной функциональной зависимости.
3. При $r_B = 0$ линейной корреляционной связи между признаками не существует (при этом может быть нелинейная корреляционная связь и даже нелинейная функциональная зависимость).

Таким образом, чем ближе по модулю коэффициент линейной корреляции к единице, чем теснее линейная зависимость между X и Y , чем ближе коэффициент корреляции к нулю, тем слабее линейная зависимость.

&5. Решение типовых задач

Задача 1. Выборочно обследовано 100 заводов по величине основных производственных фондов X (млн. руб.) и объему готовой продукции Y (млн. руб.). Результаты представлены в корреляционной таблице (табл. 1).

Таблица 1

| Y | X | | | | | |
|----|----|----|----|----|----|---------------|
| | 5 | 15 | 25 | 35 | 45 | |
| 30 | 7 | 1 | | | | 8 |
| 32 | 2 | 7 | 1 | | | 10 |
| 34 | 1 | 5 | 4 | 1 | | 11 |
| 36 | | 1 | 15 | 10 | 8 | 34 |
| 38 | | | 3 | 12 | 15 | 30 |
| 40 | | | | 1 | 6 | 7 |
| | 10 | 14 | 23 | 24 | 29 | <i>n</i> =100 |

По данным исследования требуется:

1) в прямоугольной системе координат построить эмпирические ломаные регрессии Y на X и X на Y ;

2) оценить тесноту линейной корреляционной связи;

3) составить линейные уравнения регрессии Y на X и X на Y и построить их графики в одной системе координат.

Решение_1

1. Так как при $x = 5$ признак Y имеет распределение

| | | | |
|-------|----|----|----|
| Y | 30 | 32 | 34 |
| n_i | 7 | 2 | 1 |

то условное среднее $\bar{y}_{x=5} = \frac{30 \cdot 7 + 32 \cdot 2 + 34 \cdot 1}{10} = 30,8.$

При $x=15$ признак Y имеет распределение

| | | | | |
|-------|----|----|----|----|
| Y | 30 | 32 | 34 | 36 |
| n_i | 1 | 7 | 5 | 1 |

Следовательно $\bar{y}_{x=15} = \frac{30 \cdot 1 + 32 \cdot 7 + 34 \cdot 5 + 36 \cdot 1}{14} = 32,86.$

Аналогично вычисляются все условные средние \bar{y}_x . В результате получим таблицу, выражающую корреляционную зависимость \bar{y} от X (табл. 2).

Таблица 2

| | | | | | |
|-------------|------|-------|-------|-------|-------|
| X | 5 | 15 | 25 | 35 | 45 |
| \bar{y}_x | 30,8 | 32,86 | 35,74 | 37,08 | 37,86 |

Так как при $y=30$ признак X имеет распределение

| | | |
|-------|---|----|
| X | 5 | 15 |
| n_j | 7 | 1 |

то условное среднее $\bar{x}_{y=30} = \frac{5 \cdot 7 + 15 \cdot 1}{8} = 6,25$.

При $y = 32$ признак X имеет распределение

| | | | |
|-------|---|----|----|
| X | 5 | 15 | 25 |
| n_j | 2 | 7 | 1 |

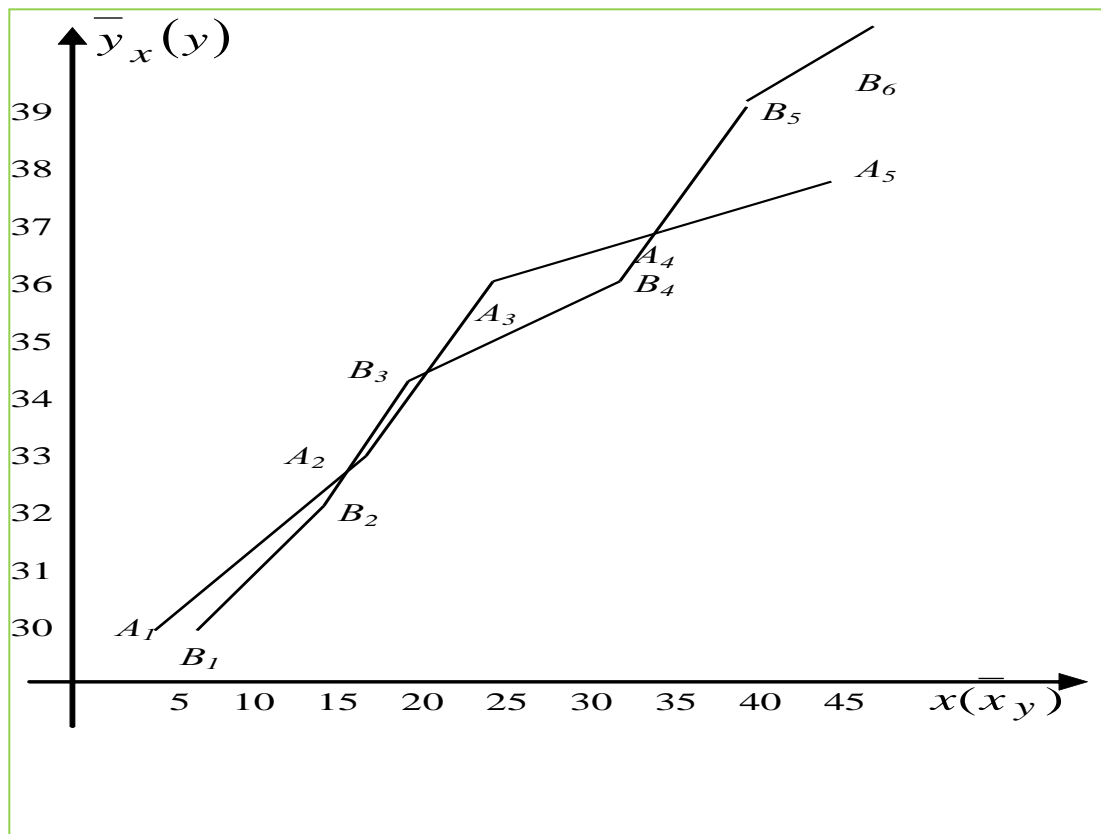
Следовательно $\bar{x}_{y=32} = \frac{5 \cdot 2 + 15 \cdot 7 + 25 \cdot 1}{10} = 14$.

Аналогично вычисляются все \bar{x}_y . В результате получим табл. 3.

Таблица 3

| | | | | | | |
|-------------|------|----|-------|-------|----|-------|
| Y | 30 | 32 | 34 | 36 | 38 | 40 |
| \bar{x}_y | 6,25 | 14 | 19,54 | 32,35 | 39 | 43,57 |

В прямоугольной системе координат построим точки $A_i(x_i; \bar{y}_{x_i})$, соединим их отрезками прямых, получим эмпирическую линию регрессии Y на X . Аналогично строятся точки $B_j(\bar{x}_y; y_j)$ и эмпирическая линия регрессии X на Y .



Решение_2

2. Выдвинув гипотезу о линейной корреляционной зависимости, оценим тесноту связи. Вычислим выборочный коэффициент корреляции

$$r_B = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}$$

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

$$\bar{y} = \frac{\sum y_j \cdot n_j}{n}$$

$$\overline{x^2} = \frac{\sum x_i^2 \cdot n_i}{n}$$

$$\overline{y^2} = \frac{\sum y_j^2 \cdot n_j}{n}$$

$$\overline{xy} = \frac{\sum x_i y_j \cdot n_{ij}}{n}$$

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$$

$$\bar{x} = \frac{5 \cdot 10 + 15 \cdot 14 + 25 \cdot 23 + 35 \cdot 24 + 45 \cdot 29}{100} = 29,8;$$

$$\bar{y} = \frac{30 \cdot 8 + 32 \cdot 10 + 34 \cdot 11 + 36 \cdot 34 + 38 \cdot 30 + 40 \cdot 7}{100} = 35,78;$$

$$\overline{x^2} = \frac{5^2 \cdot 10 + 15^2 \cdot 14 + 25^2 \cdot 23 + 35^2 \cdot 24 + 45^2 \cdot 29}{100} = 1059;$$

$$\overline{y^2} = \frac{30^2 \cdot 8 + 32^2 \cdot 10 + 34^2 \cdot 11 + 36^2 \cdot 34 + 45^2 \cdot 30 + 40^2 \cdot 7}{100} = 1287,4$$

$$\begin{aligned} \overline{xy} = & \frac{30 \cdot 5 \cdot 7 + 30 \cdot 15 \cdot 1 + 32 \cdot 5 \cdot 2 + 32 \cdot 15 \cdot 7 + 32 \cdot 25 \cdot 1 + 34 \cdot 5 \cdot 1 + 34 \cdot 15 \cdot 5}{100} + \\ & + \frac{34 \cdot 25 \cdot 4 + 34 \cdot 35 \cdot 1 + 36 \cdot 15 \cdot 1 + 36 \cdot 25 \cdot 15 + 36 \cdot 35 \cdot 10 + 36 \cdot 45 \cdot 8 + 38 \cdot 25 \cdot 3}{100} + \\ & + \frac{38 \cdot 35 \cdot 12 + 38 \cdot 45 \cdot 15 + 40 \cdot 35 + 40 \cdot 45 \cdot 6}{100} = 1095,5 \end{aligned}$$

$$\sigma_x = \sqrt{1059 - (29,8)^2} = 13,08; \sigma_y = \sqrt{1287,4 - (35,78)^2} = 2,68;$$

$$r_B = \frac{1095,5 - 29,8 \cdot 35,78}{13,08 \cdot 2,68} = 0,83.$$

Так как r_B близок к единице, то между Y и X имеется достаточно тесная корреляционная связь.

Решение_3

3. Подставляя найденные величины в уравнения

$$\bar{Y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x} (x - \bar{x}), \quad \bar{X}_y - \bar{x} = r_B \frac{\sigma_x}{\sigma_y} (y - \bar{y}),$$

получаем искомые уравнения регрессии:

1) уравнение регрессии Y на X

$$\bar{Y}_x - 35,78 = 0,83 \frac{2,68}{13,08} (x - 29,8), \quad \bar{Y}_x = 0,17x + 30,71.$$

2) уравнение регрессии X на Y

$$\bar{X}_y - 29,8 = 0,83 \frac{13,08}{2,68} (y - 35,78), \quad \bar{X}_y = 4,05y - 115,14.$$

Рефлексия деятельности

- 1. Указать Ф.И.О., группу*
- 2. Составить ключевые слова лекции*