ЭЛЕМЕНТЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Выполнила студентка группы
С-1841
Направления 43.04.01
Кабанова Анастасия

Термин корреляция употребляется в науке с конца XYIII века. Его ввел французский палеонтолог Жорж Кювье.

Это систематическая и обусловленная связь между двумя рядами данных. Или связь переменных, при которой одному значению признака соответствует несколько значений другого признака.

Корреляционный анализ – это статистический метод, изучающий связь между явлениями, если одно из них входит в число причин, определяющих другое или, если имеются общие причины, воздействующие на эти явления.

Основная задача – выявление связи между случайными величинами.

Функциональная зависимость –

это зависимость вида

$$y = f(x)$$

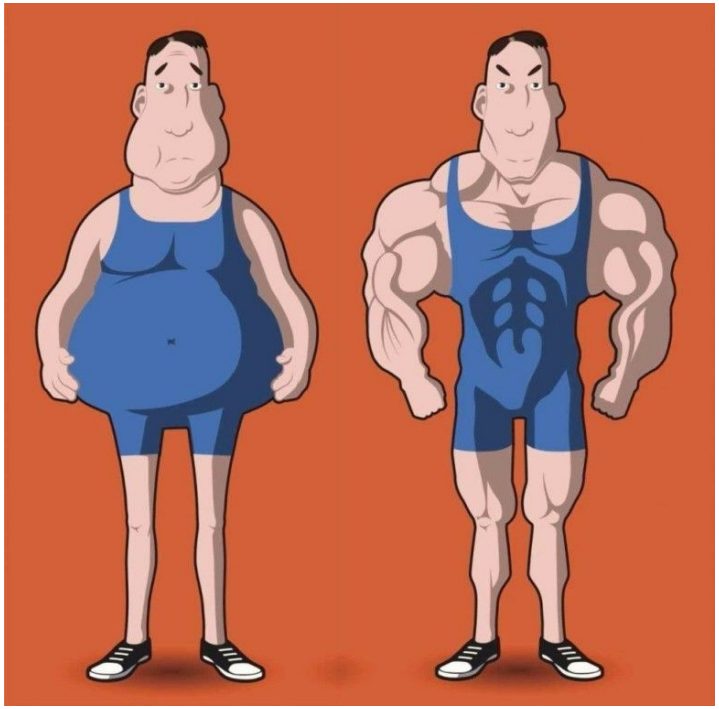
когда каждому возможному значению случайной величины *X* соответствует одно возможное значение случайной величины Y.

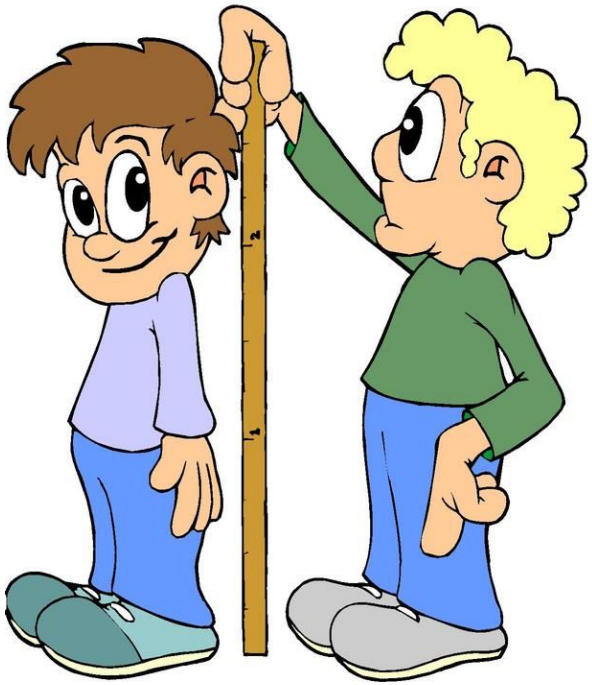
Корреляционная зависимость – это статистическая зависимость, проявляющаяся в том, что при изменении одной из величин изменяется среднее значение другой:

$$\overline{y} = f(x)$$

Например, рост и масса.

При одном и том же росте масса различных индивидуумов может быть различна, но между средними значениями этих показателей имеется определенная зависимость.





Зависимость между случайными величинами X и Y в теории вероятностей и математической статистике описывается, в первую очередь, такими характеристиками, как корреляционный момент K_{xy} и коэффициента корреляции r_{xy}.

Статистическую взаимосвязь составляющих системы случайных величин характеризует корреляционный момент (момент связи).

$$K_{XY} = \operatorname{cov}(X, Y) = M[(X - m_X)] \cdot M[(Y - m_Y)]$$

Для компактной записи результаты расчётов представляют в виде корреляционной матрицы:

$$(K_{xy})=\begin{pmatrix} D_x & K_{xy} \\ K_{yx} & D_y \end{pmatrix}$$

где Dx и Dy – дисперсии случайных величин X и Y.

Для изучения корреляционной связи, данные о статистической зависимости удобно задавать в виде корреляционной таблицы:

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Для наглядности полученного материала каждую пару можно представить в виде точки на координатной плоскости.

По оси абсцисс откладывают значения одного вариационного ряда x_i, а по оси ординат другого y_i.

ПОЛЕ КОРРЕЛЯЦИИ

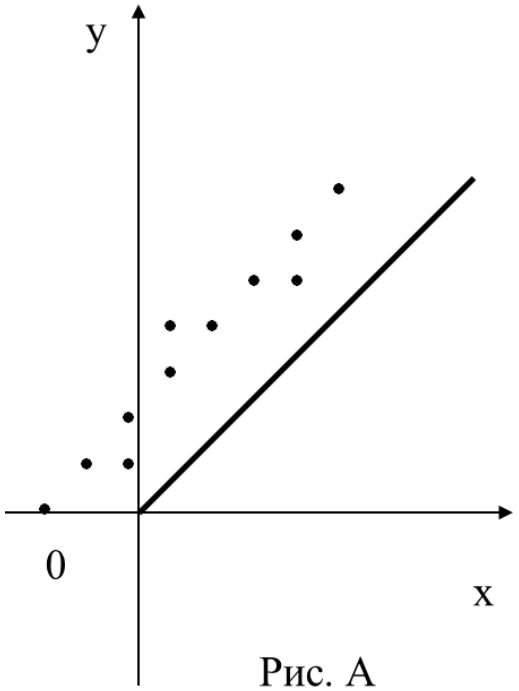


Рис. А

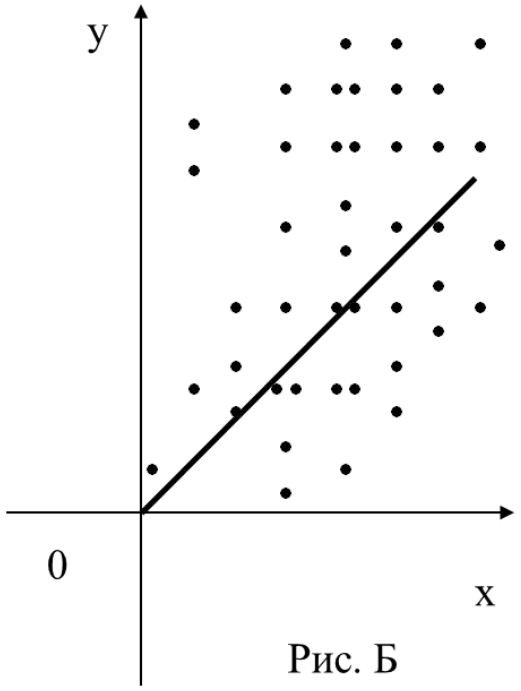


Рис. Б

$$(K_{xy}) = \begin{pmatrix} D_x & K_{xy} \\ K_{yx} & D_y \end{pmatrix}$$

где σ_x *и* σ_y *– средние квадратические отклонения случайных величин X и Y*

$$\sigma_x = \sqrt{D_x} \qquad \sigma_y = \sqrt{D_y}$$

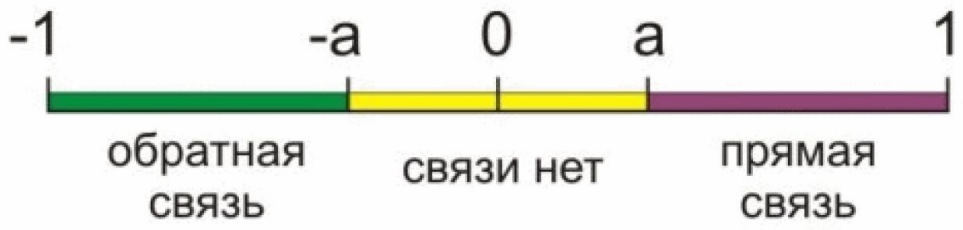
Выборочный коэффициент линейной корреляции **r** характеризует тесноту линейной связи между количественными признаками в выборке:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$(K_{xy}) = \begin{pmatrix} D_x & K_{xy} \\ K_{yx} & D_y \end{pmatrix}$$

$$(K_{xy}) = \begin{pmatrix} D_x & K_{xy} \\ K_{yx} & D_y \end{pmatrix}$$

Интерпретация коэффициента корреляции



Свойства коэффициента корреляции:

1. Коэффициент корреляции принимает значения на отрезке [-1;1].

В зависимости от того, насколько модуль **r** приближается к 1, различают связи:

- **r < 0,3** – слабая связь;
- **r = 0,3-0,5** – умеренная связь;
- **r = 0,5-0,7** – значительная;
- **r = 0,7-0,8** – достаточно тесная;
- **r = 0,8 – 0,9** – тесная (сильная);
- **r > 0,9** – очень тесная.

2. Если случайные величины между собой связаны линейно, то $|r_{XY}|=1$

4. Если случайные величины независимые, то $r_{XY}=0$

5. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится

ПРИМЕР 1

Имеются данные о результате экспериментальных замеров прочности шва для ниток различной линейной плотности

Xi	Yi
23	14,8
38	27,5
43	28,4
48	30,4

где Xi – линейная плотность нитей

Yi - прочность шва

$$(K_{xy}) = \begin{pmatrix} D_x & K_{xy} \\ K_{yx} & D_y \end{pmatrix}$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

i	Xi	Yi	Xi - Xcp	(Xi - Xcp)^2	Yi - Ycp	(Yi - Ycp)^2	(Xi - Xcp)(Yi - Ycp)
1	23	14,8	-15	225	-10,475	109,725625	157,125
2	38	27,5	0	0	2,225	4,950625	0
3	43	28,4	5	25	3,125	9,765625	15,625
4	48	30,4	10	100	5,125	26,265625	51,25
Сумма	152	101,1	0	350	0	150,7075	224

X ср	38
Y ср	25,275

Итак, получаем

r	0,975319

ПРИМЕР 2

Имеются данные о рейтинге авиакомпании по 5 бальной шкале (Xi) и оценке ее безопасности по 10 бальной шкале (Yi)

Xi	Yi
1	3
2	5
3	7
4	8

Заполним таблицу по формулам

$$\bar{x} = \frac{\sum_{t=1}^{n} x_i}{n} \; ; \quad \bar{y} = \frac{\sum_{t=1}^{n} y_i}{n} \quad \text{и} \quad r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

i	Xi	Yi	Xi - Xcp	(Xi - Xcp)^2	Yi - Ycp	(Yi - Ycp)^2	(Xi - Xcp)(Yi - Ycp)
1	1	3					
2	2	5					
3	3	7					
4	4	8					
Сумма							

X cp	
Y cp	