



Математика, часть 2

Корреляционный анализ данных

Лекция № 9



Система двух и более случайных величин

Данные, содержащими две или n случайных величин называются

- двумерными
- n - мерными

(X, Y) - Обозначение двумерной случайной величины

X, Y – составляющие или компоненты случайной величины (X, Y) .

Примеры ДСВ:

- геологическая проба руды, содержащая золото – X и серебро – Y
- результаты геохимического опробования образцов железорудного месторождения, где кроме содержания железа определяется содержание марганца, титана, висмута, ванадия и других компонентов.

Необходимо различать

- дискретные случайные величины
- непрерывные случайные величины



Изображение системы из двух случайных величин

Графически систему из двух случайных величин X_1 и Y_1 можно представить случайной точкой на плоскости XOY с координатами X_1 и Y_1 .

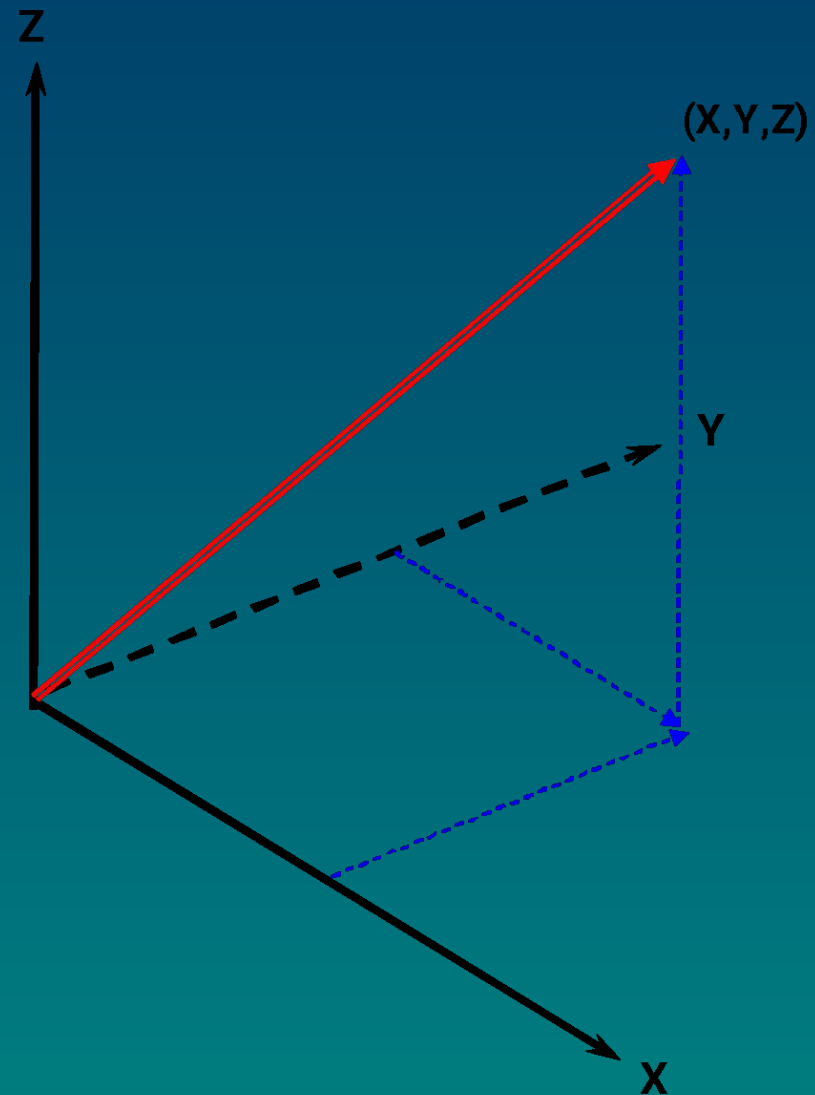




Изображение системы из трёх случайных величин

Трёхмерная случайная величина изображается

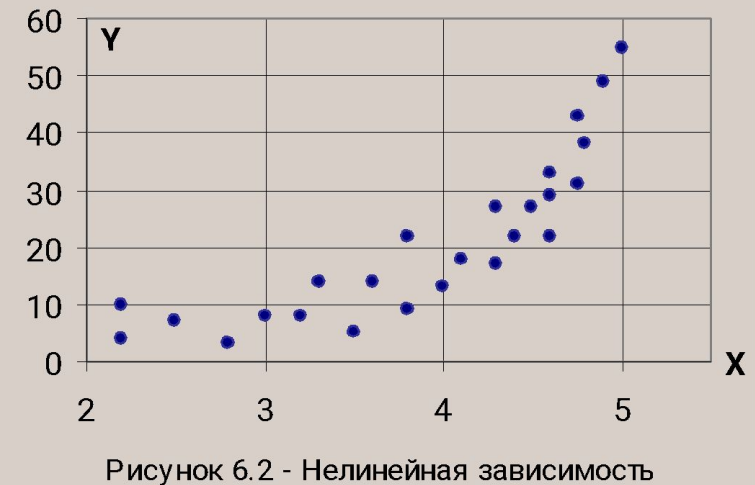
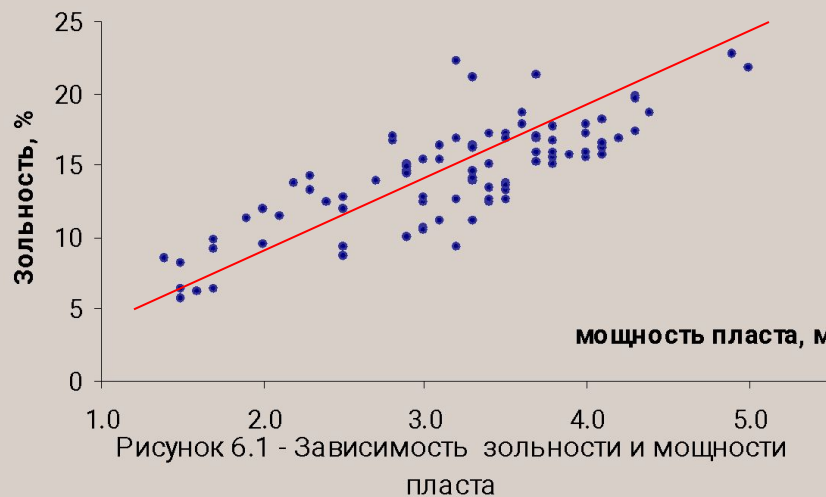
1. точкой в трёхмерном пространстве с координатами X, Y, Z
2. вектором.





2 задачи корреляционного анализа

1. **определение формы** корреляционной зависимости, иначе говоря, необходимо установить какой вид имеет функция регрессии (линейную или нелинейную)
2. **оценка силы (тесноты)** корреляционной связи с помощью коэффициента корреляции.





Закон распределения дискретной двумерной СВ

1. Законом распределения дискретной двумерной случайной величины называют перечень возможных значений этой величины (т. е. пар чисел (x_i, y_j) и их вероятностей $p(x_i, y_j)$ ($i=1, 2, \dots, n; j=1, 2, \dots, m$). Обычно закон распределения задают в виде таблицы с двойным входом

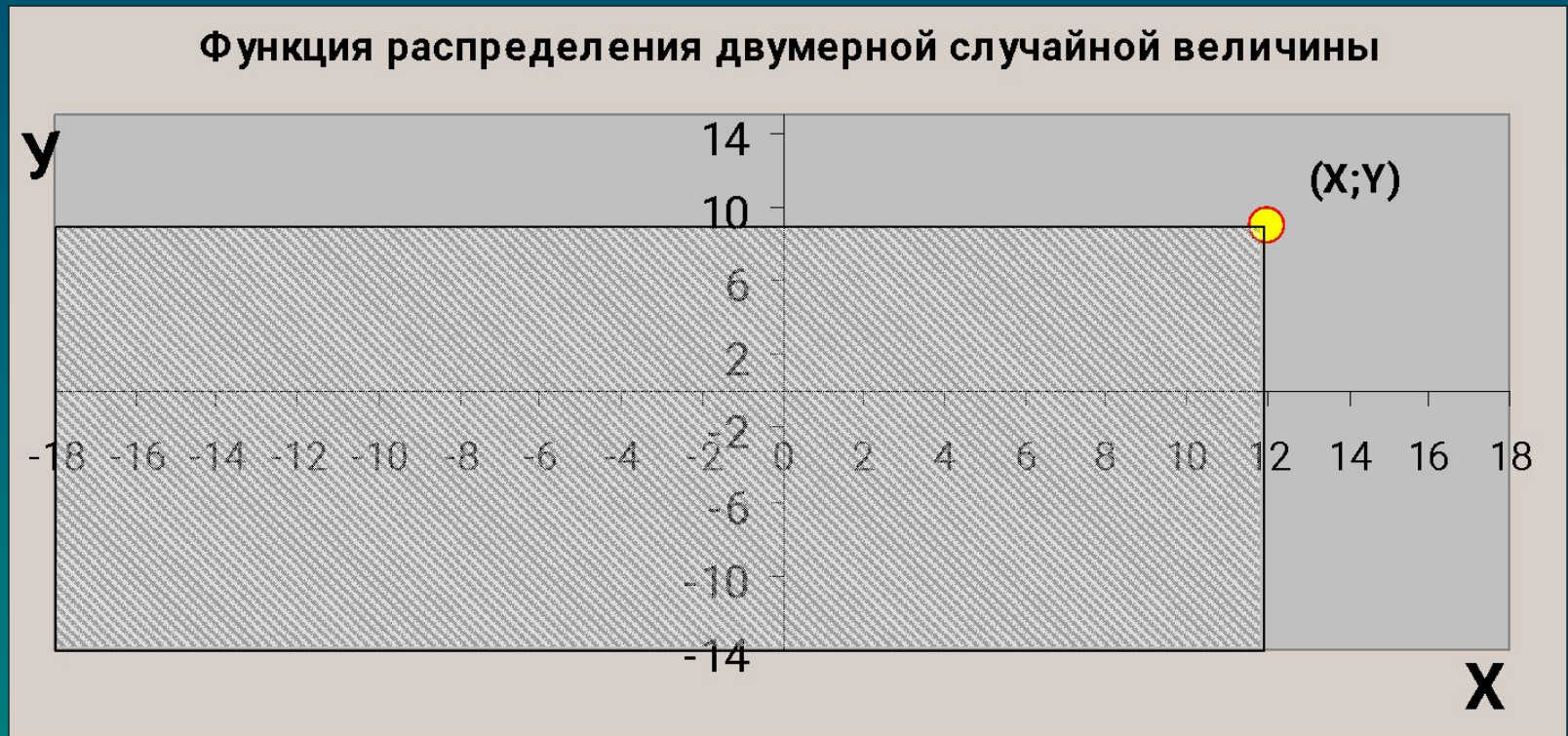
$Y \backslash X$	x_1	x_2	\dots	x_l	\dots	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	\dots	$p(x_l, y_1)$	\dots	$p(x_n, y_1)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
y_j	$p(x_1, y_j)$	$p(x_2, y_j)$	\dots	$p(x_l, y_j)$	\dots	$p(x_n, y_j)$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	\dots	$p(x_l, y_m)$	\dots	$p(x_n, y_m)$



Интегральная функция распределения дискретной двумерной СВ

1. Интегральной функцией распределения двумерной случайной величины (X, Y) называют функцию $F(x, y)$, определяющую для каждой пары чисел x, y вероятность того, что X примет значение, меньшее x и при этом Y примет значение, меньшее y :

2.
$$F(x, y) = P(X < x, Y < y),$$



Свойства интегральной функции распределения двумерной СВ

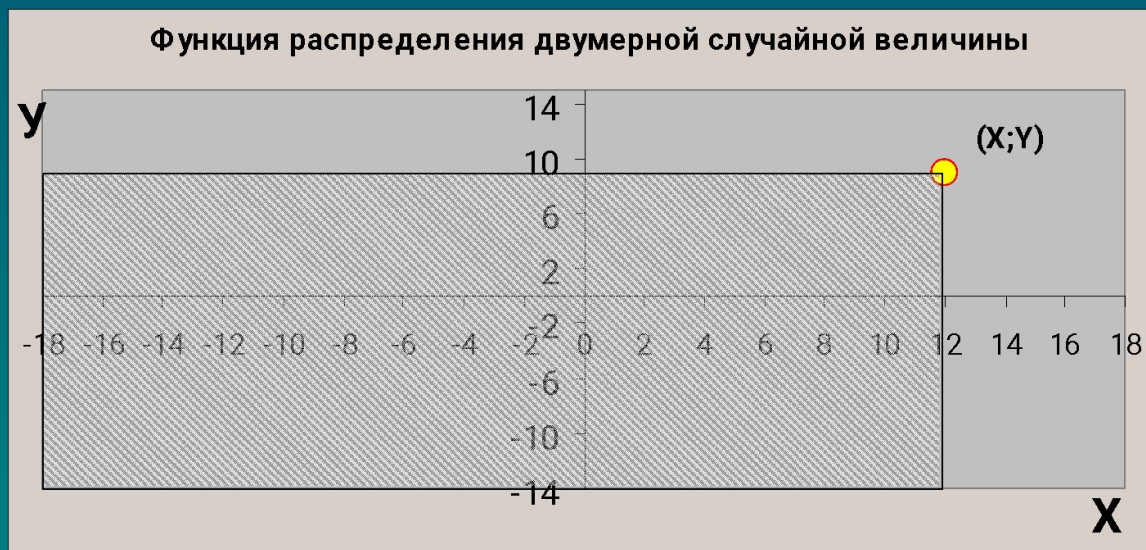
1. **Свойство 1.** Значения интегральной функции удовлетворяют двойному неравенству

$$0 \leq F(x, y) \leq 1$$

2. **Свойство 2.** $F(x, y)$ есть неубывающая функция по каждому аргументу

$$F(x_2, y) \geq F(x_1, y), \text{ если } x_2 > x_1;$$

$$F(x, y_2) \geq F(x, y_1), \text{ если } y_2 > y_1.$$





Свойства интегральной функции распределения двумерной СВ

3. Свойство 3. Имеют место предельные соотношения:

- 1) $F(-\infty, y) = 0,$
- 2) $F(x, -\infty) = 0,$
- 3) $F(-\infty, -\infty) = 0,$
- 4) $F(\infty, \infty) = 1.$



4. Свойство 4.

а) При $y = \square$ интегральная функция системы становится интегральной функцией составляющей X :

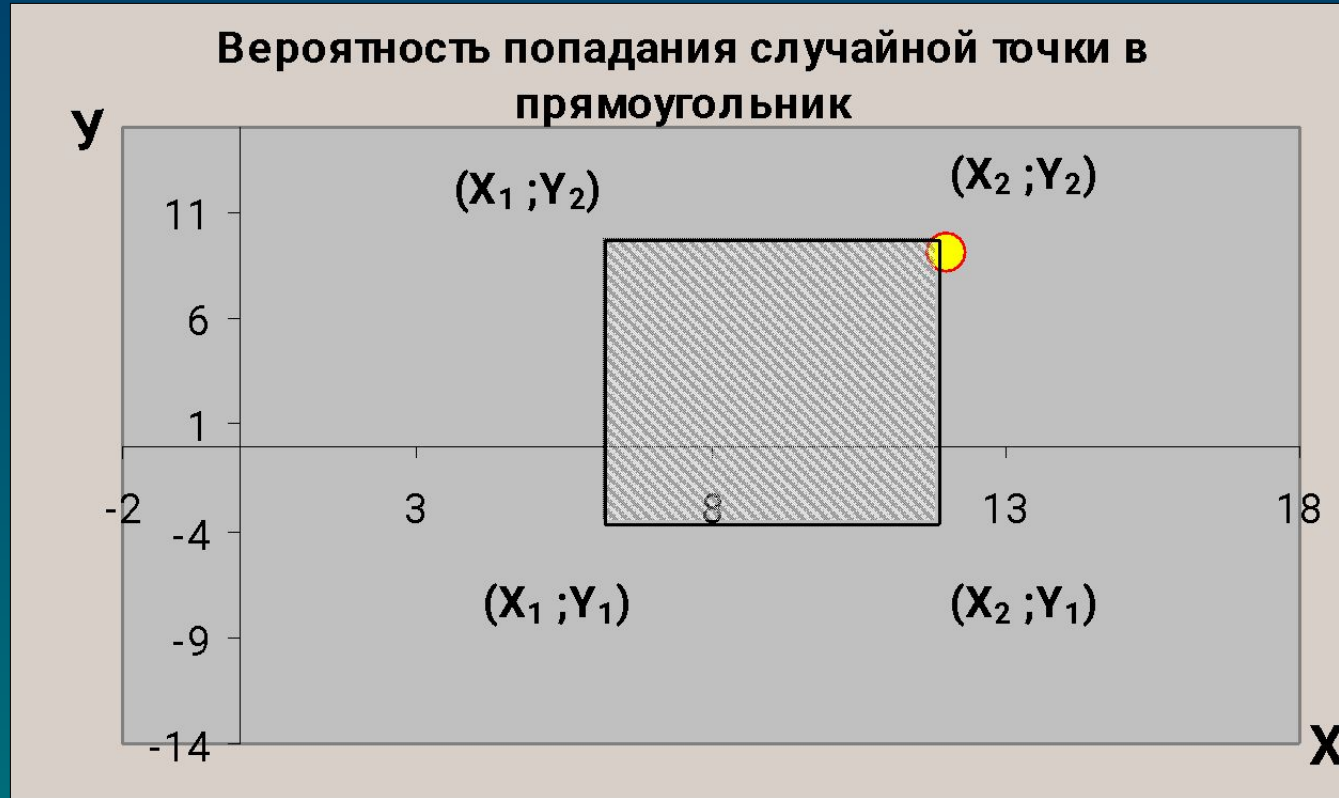
$$F(x, \square) = F_1(x).$$

б) При $x = \square$ интегральная функция системы становится интегральной функцией составляющей Y :

$$F(\infty, y) = F_2(y).$$



Свойства интегральной функции распределения двумерной СВ



$$P(x_1 < X < x_2, y_1 < Y < y_2) = \\ = [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)]$$



Двумерная плотность вероятности

Дифференциальной функцией распределения $f(x, y)$ двумерной непрерывной случайной величины (X, Y) называют вторую смешанную частную производную от интегральной функции:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x * \partial y}$$

Геометрически эту функцию можно истолковать как поверхность, которую называют *поверхностью распределения*.

Зная дифференциальную функцию $f(x, y)$, можно найти интегральную функцию $F(x, y)$ по формуле

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(x, y) dx dy,$$



Свойства дифференциальной функции распределения

1. **Свойство 1.** Дифференциальная функция неотрицательна:

$$f(x, y) \geq 0.$$

2. **Свойство 2.** Двойной несобственный интеграл с бесконечными пределами от дифференциальной функции равен единице:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$$

Отыскание дифференциальных функций

$$f_1(x) = \frac{dF_1(x)}{dx}$$

$$\frac{dF_1}{dx} = \int_{-\infty}^{\infty} f(x, y) dy,$$

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy$$

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

$$F_1(x) = F(x, \infty)$$

$$F_1(x) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(x, y) dx dy.$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx$$



Зависимые и независимые СВ

две случайные величины называются независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина.

Теорема. Для того чтобы случайные величины X и Y были независимыми, необходимо и достаточно, чтобы интегральная функция системы (X, Y) была равна произведению интегральных функций составляющих:

$$F(x, y) = F_1(x) \cdot F_2(y).$$

Следствие. Для того чтобы непрерывные случайные величины X и Y были независимыми, необходимо и достаточно, чтобы дифференциальная функция системы (X, Y) была равна произведению дифференциальных функций составляющих:

$$f(x, y) = f_1(x) \cdot f_2(y).$$

Корреляционный момент

Корреляционным моментом μ_{xy} случайных величин X и Y называют математическое ожидание произведения отклонений этих величин:

$$\mu_{xy} = M[(X - M(X))(Y - M(Y))].$$

Для вычисления корреляционного момента дискретных величин пользуются формулой

$$\mu_{xy} = \sum_{i=1}^n \sum_{j=1}^m (x_i - M(X))(y_j - M(Y)) p(x_i, y_j),$$

а для непрерывных

$$\mu_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - M(X))(y - M(Y)) f(x, y) dx dy.$$

Теорема. Корреляционный момент двух независимых случайных величин X и Y равен нулю.

$$\begin{aligned} \mu_{xy} &= M[(X - M(X)) \cdot (Y - M(Y))] = \\ &= M[X - M(X)] \cdot M[Y - M(Y)] = 0. \end{aligned}$$



Коэффициент корреляции. Зависимые СВ

Две случайные величины могут быть связаны между собой функциональной зависимостью, статистической либо независимыми между собой.

Строгая функциональная зависимость случайных величин X и Y встречается крайне редко и может быть записана, например, в виде:

$$Y = aX + b \quad (6.1)$$

Эта запись означает, что каждому значению X соответствует только одно значение Y . Функциональную зависимость можно ещё назвать теоретической.

Как правило, случайные величины подвержены влиянию не одного фактора, а целого набора случайных факторов. В результате одному значению X соответствует не одно, а множество значений Y . Это множество называется законом распределения случайной величины Y . В таком случае говорят о *статистической (корреляционной) зависимости*, при которой изменение случайной величины X вызывает изменение некоторых параметров распределения другой - Y , в частности, среднего значения \bar{Y} .



Пример корреляционной зависимости

Пример случайной величины Y , которая не связана с величиной X функционально, а связана корреляционно.

Одинаковые по форме образцы руды с одинаковым содержанием полезного компонента X характеризуются разной плотностью. Это объясняется тем, что на плотность образца породы, кроме удельного веса полезного компонента, влияют такие *случайные* факторы, как разная пористость, разное количество включений других минералов и др.

Таким образом, конкретная плотность Y руды не связана функционально с конкретным содержанием полезных компонентов. Однако практика показывает, что на полиметаллических месторождениях (свинцово-цинковых) в среднем на плотность руды Y влияет суммарное содержание свинца и цинка в пробе X , т.е. плотность Y связана с X корреляционной зависимостью.



Условные средние

Условным средним \bar{y}_x называют среднее арифметическое значений Y , соответствующих значению случайной величины $X = x$.

Допустим, что в процессе эксперимента мы отобрали 3 образца с одинаковым значением содержания и определили плотность каждого из них. Так для содержания $x_1 = 35\%$ получены три значения плотности $y_1 = 3.30 \text{ т/м}^3$, $y_2 = 3.26 \text{ т/м}^3$, $y_3 = 3.28 \text{ т/м}^3$. Среднее арифметическое значение плотности определяется по формуле 3.2 и будет равно:

$$\bar{y}_{35} = \frac{3.30 + 3.26 + 3.28}{3} = 3.28.$$

Число \bar{y}_{35} называется *условным* средним, чёрточка над буквой служит обозначением среднего арифметического, цифра 35 означает, что рассматриваются только те значения плотности Y , которые соответствуют содержанию полезных компонентов в образце породы $x_1 = 35\%$.



Корреляционная зависимость

Если каждому значению x соответствует одно значение условной средней \bar{y}_x , то считается, что случайная величина Y зависит от X корреляционно.

Корреляционной зависимостью Y от X называют функциональную зависимость условной средней \bar{y}_x от x :

$$\bar{y}_x = f(x), \quad (5.1)$$

а уравнение (5.1) уравнением регрессии Y по X . Функция $f(x)$ называется регрессией Y по X , а её график - линией регрессии Y по X .

Поскольку случайные величины X, Y связаны корреляционно, то по аналогии можно определить корреляционную зависимость X по Y . Все вышеприведённые определения будут справедливы для данного случая, если в определениях поменять местами случайные величины X, Y .



Числовые характеристики СВ

Начальным моментом порядка k, s называется математическое ожидание произведения соответствующих степеней случайных величин:

$$m_{k,s} = M[X^k Y^s] \quad (6.13)$$

Для дискретных случайных величин выражение 6.13 имеет вид:

$$m_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{i,j} \quad (6.14)$$

где: $p_{i,j} = P[(X = x_i)(Y = y_j)]$ - вероятности.

Для непрерывных случайных величин начальный момент порядка k, s запишется:

$$m_{k,s} = \int_{-\infty}^{\infty} x^k y^s f(x, y) dx dy \quad (6.15)$$

где: $f(x, y)$ - плотность распределения системы двух случайных величин.



Числовые характеристики СВ

Центральным моментом порядка k, s называется математическое ожидание соответствующих степеней центрированных случайных величин X, Y :

$$\mu_{k,s} = M[X^k Y^s] = M\left[\left(X - m_{1,0}\right)\left(Y - m_{0,1}\right)\right]^2, \quad (6.16)$$

где: $m_{1,0}, m_{0,1}$ - математические ожидания случайных величин X, Y .

(второй смешанный центральный момент).

$$\mu_{11} = K_{xy} = m_{11} - m_{10}m_{01}, \quad (6.17)$$

и далее – коэффициент корреляции:

$$r_{xy} = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \frac{m_{11} - m_{10}m_{01}}{\sigma_x \sigma_y}, \quad (6.18)$$

где: μ_{20} - дисперсия случайной величины X ;

μ_{02} - дисперсия случайной величины Y ;

σ_x, σ_y - стандарты СВ X, Y ; $\sigma_x = \sqrt{\mu_{20}}$, $\sigma_y = \sqrt{\mu_{02}}$.



Числовые характеристики СВ

Дисперсия для случайных величин X, Y вычисляется по следующим формулам:

$$\mu_{20} = m_{20} - m_{10}^2 \quad (6.19)$$

$$\mu_{02} = m_{02} - m_{01}^2 \quad (6.20)$$

Сила связи между X, Y оценивается при помощи выборочного коэффициента корреляции r_{xy} :

$$r_{xy} = \frac{\mu_{11}}{\sigma_x \sigma_y}, \quad (6.2)$$

где: μ_{11} - второй смешанный центральный (корреляционный) момент;

σ_x, σ_y - стандарты (СКО) по x и y соответственно



Числовые характеристики СВ

Корреляционный момент K_{xy} , или его ещё называют *моментом связи*, вычисляется по одной из двух равноценных формул:

$$K_{xy} = \mu_{11} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N} \quad (6.3)$$

$$K_{xy} = \mu_{11} = \frac{\sum_{i=1}^N x_i y_i}{N} - \bar{x} * \bar{y} \quad (6.4)$$

Выборочное уравнение регрессии (Y по X) имеет вид:

$$\bar{y}_x - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad (6.5)$$



Уравнение регрессии

Обычно в таком виде уравнение 6.5 не применяется, поэтому его приводят к линейному виду 6.1. Угловым коэффициентом a и свободным членом b можно определить из выражений:

$$a = r_{xy} \frac{\sigma_y}{\sigma_x} \quad (6.6)$$

$$b = \bar{y} - r_{xy} \frac{\sigma_y}{\sigma_x} \bar{x} \quad (6.7)$$

Окончательно уравнение регрессии (Y по X): $\bar{y}_x = ax + b$. Выборочное уравнение регрессии (X по Y) представлено формулой 6.8, но его, как и выражение 6.5, тоже приводят к линейному виду $\bar{x}_y = a'y + b'$.

$$\bar{x}_y - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \quad (6.8)$$



Погрешность коэффициента корреляции

Погрешность определения коэффициента корреляции σ_r зависит от объёма выборки из генеральной совокупности. С увеличением объёма выборки погрешность будет уменьшаться. Эта зависимость выражается следующей формулой:

$$\sigma_r = \frac{1 - r_{xy}^2}{\sqrt{N}} \quad (6.9)$$

Надёжность коэффициента корреляции оценивается при помощи следующего отношения:

$$\varphi = \frac{|r_{xy}|}{\sigma_r} \quad (6.10)$$

Если $\varphi \geq 3$ то, согласно теореме Ляпунова с вероятностью $P=0,997$ можно утверждать, что связь между изучаемыми случайными величинами надёжная и наоборот.



Числовые характеристики СВ

Чтобы оценить точность уравнения регрессии определяют среднее квадратичное отклонение предсказанных значений \bar{y}_x от исходных y_i . Это можно выполнить по одной из двух формул:

$$\sigma_{y/x} = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y}_{x_i})^2}{N}}, \quad (6.11)$$

$$\sigma_{y/x} = \sigma_y \sqrt{1 - r_{xy}^2}. \quad (6.12)$$



Корреляционный анализ при большом числе данных

$$\Delta x = \frac{x_{\max} - x_{\min}}{1 + 3.2 * \lg N}$$

$$\Delta y = \frac{y_{\max} - y_{\min}}{1 + 3.2 * \lg N}$$

1. Вычисляют классовый интервал по формуле Стержеса
2. Строят корреляционную решётку
3. Заносят все пары исходных данных в корреляционную решётку
4. Находят суммы, суммы произведений
5. Вычисляют средние и стандарты для X и Y, корреляционный момент и коэффициент корреляции
6. Рассчитывают уравнения регрессии, СГО
7. Выполняют оценку точности найденного уравнения регрессии



Корреляционная решётка

Таблица 6.3 Корреляционная решётка

		Условное значение класса α_x											
		-3	-2	-1	0	1	2	3	4				
α_y	Классы	28,2 – 33,8	33,8 – 39,4	39,4 – 45,0	45,0 – 50,6	50,6 – 56,2	56,2 – 61,8	61,8 – 67,4	67,4 – 73,0	n_y	$\sum n_y$	$\sum n_y^2$	$\sum n_{xy}$
-3	3,12 – 3,32	14	5							19	-57	171	156
-2	3,32 – 3,52	126	30							22	-44	88	72
-1	3,52 – 3,72	18	36	18	0					7	-7	7	-1
0	3,72 – 3,92				0	6	1			12	0	0	0
1	3,92 – 4,12				2	10				8	8	8	11
2	4,12 – 4,32				5	5	3			14	28	56	68
3	4,32 – 4,52				2	1	7	5	1	15	45	135	144
4	4,52 – 4,72						1	10	4	3	12	48	44
	n_x						12	32	2	100	-15	513	494
	$n_x \alpha_x$	17	14	9	9	17	11	16	7	27			
	$n_x \alpha_x^2$	-51	-28	-9	0	17	22	48	28				
	$\sum n_{xy} \alpha_x \alpha_y$	153	56	9	0	17	44	144	112	535			
		144	66	18	0	6	40	132	88	494			

Расчёт начальных моментов

Средние значения для случайных величин X и Y определяются из выражений:

$$m_{10} = x_0 + m'_{10} * \Delta x \quad (6.21)$$

$$m_{01} = y_0 + m'_{01} * \Delta y \quad (6.22)$$

Первые **условные начальные моменты** вычисляются по следующим формулам:

$$m'_{10} = \frac{\sum n_x \alpha_x}{N}, \quad m'_{01} = \frac{\sum n_y \alpha_y}{N} \quad (6.23)$$

где: N – количество пар значений X, Y .

Чтобы получить $\sum n_x \alpha_x$, необходимо умножить число попавших значений n_x в класс на его условное значение α_x и просуммировать.



Расчёт вторых условных начальных моментов

Вторые условные начальные моменты в соответствии с их определением найдутся из следующих выражений:

$$m'_{20} = \frac{\sum n_x \alpha_x^2}{N}, \quad m'_{02} = \frac{\sum n_y \alpha_y^2}{N} \quad (6.24)$$

Дисперсию и стандарты (СКО) удобнее вычислять через вторые условные центральные моменты:

$$D_x = \mu_{20} = \left(m'_{20} - m'_{10}{}^2 \right) * \Delta x^2, \quad D_y = \mu_{02} = \left(m'_{02} - m'_{01}{}^2 \right) * \Delta y^2 \quad (6.25)$$

$$\sigma_x = \sqrt{D_x}, \quad \sigma_y = \sqrt{D_y} \quad (6.26)$$



Расчёт вторых условных начальных моментов

Коэффициент корреляции для принятого способа обработки вычисляют через условные начальные моменты и условные стандарты σ'_x, σ'_y по формуле:

$$r_{xy} = \frac{m'_{11} - m'_{10}m'_{01}}{\sigma'_x \sigma'_y} \quad (6.27)$$

где: m'_{11} - второй смешанный условный начальный момент:

$$m'_{11} = \frac{\sum n_{xy} \alpha_x \alpha_y}{N} \quad (6.28)$$



Пример

Вычислим коэффициент корреляции по данным табл. 6.3.

$$r_{xy} = \frac{\frac{494}{100} - \frac{27}{100} * \frac{-15}{100}}{2.30 * 2.26} = 0.959$$

Погрешность и значимость коэффициента корреляции определим по формулам 6.9, 6.10:

$$\sigma_r = (1 - 0.959^2) / \sqrt{100} = 0.008 \quad \varphi = |0.959| / 0.008 = 120$$

На основании этих расчётов делаем следующие выводы:

Зависимость между содержанием железа и плотностью руды существенная, так как $r_{xy} > 0.5$;

Коэффициент корреляции вычислен надёжно, поскольку $\varphi > 3$;



Пример

Можно составить уравнение регрессии $\bar{y}_x = f(x)$.

$$\bar{y}_x - 3.79 = 0.959 \frac{0.45}{12.86} (x - 49.37)$$

После упрощений и преобразований будем иметь:

$$\bar{y}_x = 0.033707 * x + 2.12602 \quad (6.29)$$

Уравнение регрессии X по Y получим из выражения 6.8:

$$\bar{x}_y = 27.30377 * y - 54.1153 \quad (6.30)$$