Система двух случайных величин

Двумерная случайная величина

До сих пор мы рассматривали дискретные случайные величины, которые называют одномерными: их возможные значения определялись одним числом. Кроме одномерных величин рассматривают также величины, возможные значения которых определяются несколькими числами. Двумерную случайную величину обозначают через (X, Y); каждая из величин X и Y называется компонентой (составляющей). Обе величины X и Y, рассматриваемые одновременно, образуют систему двух случайных величин. Например, при штамповке стальных пластинок их длина и ширина представляют собой двумерную случайную величину.

Определение 1. Законом распределения двумерной случайной величины (X, Y) называют множество возможных пар чисел (x_i, y_j) и их вероятностей $p(x_i, y_j)$. Двумерную случайную величину можно трактовать как случайную точку $A(X, Y)$ на координатной плоскости.

Закон распределения двумерной случайной величины обычно задается в виде таблицы, в строках которой указаны возможные значения x_i случайной величины X, а в столбцах — возможные значения y_j случайной величины Y, на пересечениях строк и столбцов указаны соответствующие вероятности p_{ij}. Пусть случайная величина X может принимать n значений, а случайная величина Y - m значений. Тогда закон распределения двумерной случайной величины (X, Y) имеет вид

Y/X	x_1	x_2	$\dots$	x_i	$\dots$	x_n
y_1	p_{11}	p_{12}	$\dots$	p_{1i}	$\dots$	p_{1n}
y_2	p_{21}	p_{22}	$\dots$	p_{2i}	$\dots$	p_{2n}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_j	p_{j1}	p_{j2}	$\dots$	p_{ji}	$\dots$	p_{jn}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_m	p_{m1}	p_{m2}	$\dots$	p_{mi}	$\dots$	p_{mn}

(18.21)

Из этой таблицы можно найти законы распределения каждой из случайных компонент. Например, вероятность того, что случайная величина X примет значение x_k, равна, согласно теореме сложения вероятностей независимых событий,

$$P(x_k) = p_{1k} + p_{2k} + \ldots + p_{mk}, \quad k = 1, 2, \ldots, n. \qquad (18.22)$$

Иными словами, для нахождения вероятности $P(x_k)$ нужно просуммировать все m вероятностей по k-му столбцу таблицы (18.21). Аналогично получается вероятность того, что случайная величина Y примет возможное значение y_r: $P(y_r)$ получается суммированием всех n вероятностей r-й строки таблицы

(18.21) (r = 1, 2, ... , m). Отсюда следует, что сумма всех вероятностей в законе распределения (18.21) равна единице:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1. \qquad (18.23)$$

Корреляционный момент

Определение 2. *Корреляционным моментом* случайных величин X и Y (или *ковариацией)* называется математическое ожидание произведений их отклонений:

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

Корреляционный момент служит для описания связи между случайными величинами X и Y. Из свойств математического ожидания легко убедиться в том, что μ_{xy} можно записать в следующем виде:

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

Для непосредственного вычисления корреляционного момента (ковариации) используется формула (см. распределение (18.21))

$$\mu_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{ij} - M(X)M(Y). \qquad (18.24)$$

ТЕОРЕМА 3. *Корреляционный момент двух независимых случайных величин X и Y равен нулю.*

Если корреляционный момент μ_{xy} не равен нулю, то, стало быть, величины X и Y являются зависимыми.

Коэффициент корреляции

Из определения корреляционного момента следует, что его размерность равна произведению размерностей величин X и Y; например, если X и Y измерены в сантиметрах, то μ_{xy} имеет размерность см^2.

Это обстоятельство затрудняет сравнение корреляционных моментов различных систем случайных величин. Для устранения этого недостатка вводят безразмерную числовую характеристику — *коэффициент корреляции,* величина которого не зависит от выбора системы измерения случайных величин.

Определение 3. Коэффициентом корреляции случайных величин X и Y называется отношение их корреляционного момента к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \mu_{xy}/(\sigma_x\sigma_y) = \mu_{xy}/\sqrt{D(X)D(Y)}. \qquad (18.25)$$

Из определения и свойств математического ожидания и дисперсии следует важный вывод, что абсолютная величина коэффициента корреляции не превосходит единицы:

$$|r_{xy}| \leqslant 1 \text{ или } -1 \leqslant r_{xy} \leqslant 1. \qquad (18.26)$$

Определение 4. Две случайные величины X и Y называются *коррелированными,* если их корреляционный момент (коэффициент корреляции) отличен от нуля; если же их корреляционный момент равен нулю, то X и Y называются *некоррелированными.*

Таким образом, две коррелированные случайные величины (т.е. при $r_{xy} \neq 0$) являются также и зависимыми. Обратное утверждение неверно, т.е. две зависимые величины могут быть как коррелированными, так и некоррелированными.

Линейная регрессия

Пусть (X, Y) — двумерная случайная величина, где X и Y — зависимые случайные величины. Оказывается возможным приближенное представление величины Y в виде линейной функции величины X:

$$Y \approx g(x) = aX + b, \qquad (18.27)$$

где a и b — параметры, подлежащие определению. Обычно эти величины определяются с помощью метода наименьших квадратов (см. п. 8.5).

Определение 5. Функция (18.27) называется *наилучшим приближением* в смысле метода наименьших квадратов, если математическое ожидание $M[Y — g(X)]^2$ принимает наименьшее возможное значение. Функцию $g(x)$ называют *среднеквадратической регрессией* Y на X.

ТЕОРЕМА 4. *Линейная средняя квадратическая регрессия Y на X имеет вид*

$$g(X) = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(X - m_x), \qquad (18.28)$$

где r_{xy} *определяется формулой* (18.25), $m_y = M(Y)$ *и* $m_x = M(X)$ — *математические ожидания соответственно случайных величин Y и X.*

Коэффициент $b = r_{xy}\sigma_y / \sigma_x$ называют *коэффициентом регрессии* Y на X, а прямую

$$y - m_y = r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x),$$

реализующую линейную зависимость (18.28) случайной величины Y от случайной величины X, называют *прямой среднеквадратической регрессии* X на Y.

Аналогичную форму записи имеет прямая среднеквадратическая регрессия X на Y:

$$x - m_x = r_{xy}\frac{\sigma_x}{\sigma_y}(y - m_y).$$

Рассмотрим пример:

Дано распределение вероятностей двумерной случайной величины

y/x	0,1	0,2	0,3

21	0,01	0,02	0,3
39	0,03	0,1	0,1
82	0,2	0,2	0,04

Найти одномерные законы распределения составляющих *X*, *Y*, вычислить их числовые характеристики, коэффициенты корреляции. Записать уравнения линейной регрессии. Найти условные законы распределения, условные математические ожидания и построить линии регрессии.

Решение

1. Составим одномерные законы распределения.

В нашем случае возможные значения случайной величины *X:* $x_1 = 0{,}1$, $x_2 = 0{,}2$, $x_3 = 0{,}3$. Тогда, согласно формуле (18.22), имеем $P(x_1) = 0{,}01 + 0{,}03 + 0{,}2 = 0{,}24$, $P(x_2) = 0{,}02 + 0{,}1 + 0{,}2 = 0{,}32$, $P(x_3) = 0{,}3 + 0{,}1 + 0{,}04 = 0{,}44$. Отсюда получаем закон распределения *X:*

X	0,1	0,2	0,3
P	0,24	0,32	0,44

Контроль 0,24 + 0,32 + 0,44 = 1.

Аналогично получаем и для распределения *Y*: $y_1 = 21$, $y_2 = 39$, $y_3 = 82$; $P(y_1) = 0{,}01 + 0{,}02 + 0{,}3 = 0{,}33$, $P(y_2) = 0{,}03 + 0{,}10 + 0{,}10 = 0{,}23$, $P(y_3) = 0{,}2 + 0{,}2 + 0{,}04 = 0{,}44$.

Y	21	39	82
P	0,33	0,23	0,44

Контроль 0,33 + 0,23 + 0,44 = 1.

2. Найдем числовые характеристики.

Числовые характеристики случайных величин *X*, *Y* находятся по стандартным формулам:

математическое ожидание	$M(\mathrm{X}) = m_x = \sum_{i=1}^{n} x_i p_i$	$M(\mathrm{Y}) = m_y = \sum_{j=1}^{n} y_j p_j$
дисперсия	$D(X) = M(X^2) - m_x^2$	$D(Y) = M(Y^2) - m_y^2$
среднее квадратическое отклонение	$\sigma(X) = \sigma_x = \sqrt{D(\mathrm{X})}$	$\sigma(Y) = \sigma_y = \sqrt{D(Y)}$

M(*X*) = 0,1*0,24 + 0,2*0,32 + 0,3*0,44 = 0,22

$$M(X^2) = 0{,}1^2 * 0{,}24 + 0{,}2^2 * 0{,}32 + 0{,}3^2 * 0{,}44 = 0{,}0548$$

$$D(X) = 0{,}0548 - 0{,}22^2 = 0{,}0064$$

$$\sigma(X) = \sqrt{0{,}0064} = 0{,}08$$

M(Y)=21*0,33+39*0,23+82*0,44=51,98

$$M(Y^2) = 21^2 * 0{,}33 + 39^2 * 0{,}23 + 82^2 * 0{,}44 = 3453{,}92$$

$$D(Y) = 3453{,}92 - 51{,}98^2 = 752{,}0$$

$$\sigma(Y) = \sqrt{752{,}0} = 27{,}42$$

3. Вычислим корреляционный момент.

Составим ряд распределения произведения *XY*

XY	2,1	3,9	8,2	4,2	7,8	16,4	6,3	11,7	24,6
P	0,01	0,03	0,2	0,02	0,1	0,2	0,3	0,1	0,04

M(*XY*)=2,1*0,01 + 3,9*0,03 + 8,2*0,2 + 4,2*0,02 + 7,8*0,1 + 16,4*0,2 + 6,3*0,3 + 11,7*0,1 + 24,6*0,04 = 9,966

$\mu_{xy} = 9{,}966 - 0{,}22 * 51{,}98 = -1{,}47$ корреляционный момент

$$r_{xy} = \frac{M(XY) - m_x \cdot m_y}{\sigma_x \cdot \sigma_y} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y} = \frac{-1{,}47}{0{,}08 \cdot 27{,}42} = -0{,}67$$

Уравнение линейной регрессии *Y* на *X*

$$y - 51{,}98 = -0{,}67\frac{27{,}42}{0{,}08}(x - 0{,}22)$$

После раскрытия скобок и приведения подобных слагаемых получим следующее уравнение:

$$y = 102{,}5 - 229{,}5x$$

Уравнение линейной регрессии X на Y

$$x - 0{,}22 = -0{,}67\frac{0{,}08}{27{,}42}(y - 51{,}98)$$

После раскрытия скобок и приведения подобных слагаемых получим следующее уравнение:

$$x = 0{,}322 - 0{,}002y$$

4. Найдем условные законы распределения величины *Y*:

Y	21	39	82
P(*Y*/*X*=0,1)	1/24	3/24	20/24
P(*Y*/*X*=0,2)	2/32	10/32	20/32
P(*Y*/*X*=0,3)	30/44	10/44	4/44

$$P(Y = 21/\, X = 0{,}1) = \frac{P(X = 0{,}1; Y = 21)}{P(x = 0{,}1)} = \frac{0{,}01}{0{,}24} = \frac{1}{24}$$

$$P(Y = 39/\, X = 0{,}1) = \frac{P(X = 0{,}1; Y = 39)}{P(x = 0{,}1)} = \frac{0{,}03}{0{,}24} = \frac{3}{24}$$

$$P(Y=82/\,X=0{,}1)=\frac{P(X=0{,}1;Y=82)}{P(x=0{,}1)}=\frac{0{,}2}{0{,}24}=\frac{20}{24}$$

$$P(Y=21/\,X=0{,}2)=\frac{P(X=0{,}2;Y=21)}{P(x=0{,}2)}=\frac{0{,}02}{0{,}32}=\frac{2}{32}$$

$$P(Y=39/\,X=0{,}2)=\frac{P(X=0{,}2;Y=39)}{P(x=0{,}2)}=\frac{0{,}1}{0{,}32}=\frac{10}{32}$$

$$P(Y=82/\,X=0{,}2)=\frac{P(X=0{,}2;Y=82)}{P(x=0{,}2)}=\frac{0{,}2}{0{,}32}=\frac{20}{32}$$

$$P(Y=21/\,X=0{,}3)=\frac{P(X=0{,}3;Y=21)}{P(x=0{,}3)}=\frac{0{,}3}{0{,}44}=\frac{30}{44}$$

$$P(Y=39/\,X=0{,}3)=\frac{P(X=0{,}3;Y=39)}{P(x=0{,}3)}=\frac{0{,}1}{0{,}44}=\frac{10}{44}$$

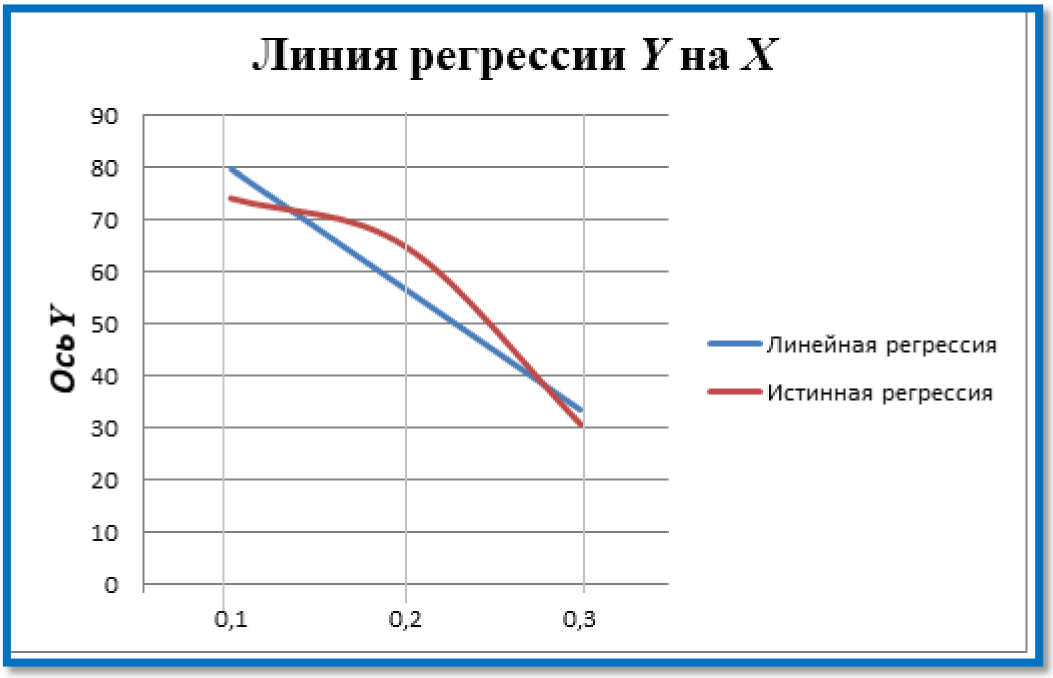
Вычислим условные математические ожидания

$$M(Y/X=0{,}1)=21*\frac{1}{24}+39*\frac{3}{24}+82*\frac{20}{24}=\frac{1778}{24}=74{,}1$$

$$M(Y/X=0{,}2)=21*\frac{2}{32}+39*\frac{10}{32}+82*\frac{20}{32}=\frac{2072}{32}=64{,}75$$

$$M(Y/X=0{,}3)=21*\frac{30}{44}+39*\frac{10}{44}+82*\frac{4}{44}=\frac{1348}{44}=30{,}64$$

Итак, мы нашли зависимость условного математического ожидания от значений случайной величины *X*, т.е. регрессию *Y* на *X*. Изобразим на координатной плоскости линии истинной и линейной регрессии.



5. Аналогично составим условные законы распределения X на Y

X	0,1	0,2	0,3
P(X/Y=21)	1/33	2/33	30/33
P(X/Y=39)	3/23	10/23	10/23
P(X/Y=82)	10/22	10/22	2/22

$$P(X=0{,}1/Y=21)=\frac{P(X=0{,}1;Y=21)}{P(Y=21)}=\frac{0{,}01}{0{,}33}=\frac{1}{33}$$

$$P(X=0{,}2/Y=21)=\frac{P(X=0{,}2;Y=21)}{P(Y=21)}=\frac{0{,}02}{0{,}33}=\frac{2}{33}$$

$$P(X=0{,}3/Y=21)=\frac{P(X=0{,}3;Y=21)}{P(Y=21)}=\frac{0{,}3}{0{,}33}=\frac{30}{33}$$

$$P(X=0{,}1/Y=39)=\frac{P(X=0{,}1;Y=39)}{P(Y=39)}=\frac{0{,}03}{0{,}23}=\frac{3}{23}$$

$$P(X=0{,}2/Y=39)=\frac{P(X=0{,}2;Y=39)}{P(Y=39)}=\frac{0{,}1}{0{,}23}=\frac{10}{23}$$

$$P(X=0{,}3/Y=39)=\frac{P(X=0{,}3;Y=39)}{P(Y=39)}=\frac{0{,}1}{0{,}23}=\frac{10}{23}$$

$$P(X=0{,}1/Y=82)=\frac{P(X=0{,}1;Y=82)}{P(Y=82)}=\frac{0{,}20}{0{,}44}=\frac{10}{22}$$

$$P(X=0{,}2/Y=82)=\frac{P(X=0{,}2;Y=82)}{P(Y=82)}=\frac{0{,}20}{0{,}44}=\frac{10}{22}$$

$$P(X=0{,}3/Y=82)=\frac{P(X=0{,}3;Y=82)}{P(Y=82)}=\frac{0{,}04}{0{,}44}=\frac{2}{22}$$

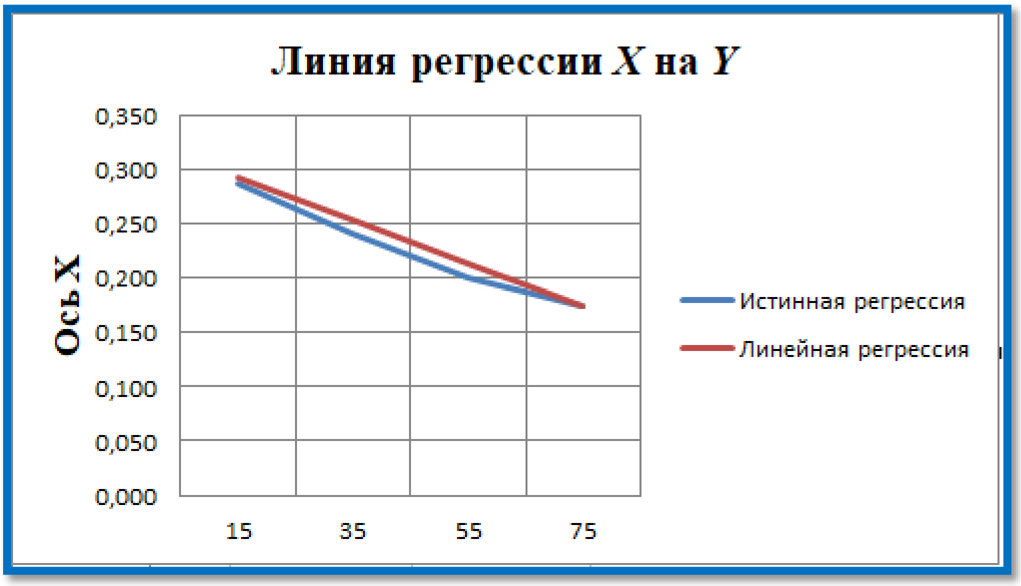
Вычислим условные математические ожидания

$$M(X/Y=21)=0{,}1*\frac{1}{33}+0{,}2*\frac{2}{33}+0{,}3*\frac{30}{33}=\frac{95}{330}=0{,}29$$

$$M(X/Y=39)=0{,}1*\frac{3}{23}+0{,}2*\frac{10}{23}+0{,}3*\frac{10}{23}=\frac{53}{230}=0{,}23$$

$$M(X/Y=82)=0{,}1*\frac{10}{22}+0{,}2*\frac{10}{22}+0{,}3*\frac{2}{22}=\frac{36}{220}=0{,}16$$

Изобразим на координатной плоскости линии истинной и линейной регрессии.



Линия регрессии X на Y
Ось X
0,350
0,300
0,250
0,200
0,150
0,100
0,050
0,000
15
35
55
75
Истинная регрессия
Линейная регрессия