1 Случайный вектор

1.1 Совместное распределение координат случайного вектора

Случайный n***-мерный вектор*** (или многомерная случайная величина) определяется упорядоченным набором n случайных величин:

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_n).$$

В частности, двумерный случайный вектор определяется парой случайных величин: (ξ, η).

Выделим раздельно два вида случайного вектора в зависимости от множества значений его координат: 1) непрерывное множество значений каждой координаты случайного вектора и 2) дискретное (конечное или бесконечной) множество значений каждой координаты. Ради краткости в первом случае такой вектор назовем ***непрерывным случайным вектором,*** во втором **–** ***дискретным случайным вектором.***

Определение. ***Функцией распределения*** n-мерного случайного вектора называется функция n аргументов, которая определяется как вероятность пересечения событий $\xi_k < x_k$, $k = 1, 2, \dots, n$:

$$F(x_1, x_2, \dots x_n) = P\left(\bigcap_{k=1}^{k=n} (\xi_k < x_k)\right). \qquad (1)$$

Распределение вероятностей, определяемое такой функцией (1), называется ***совместным распределением координат*** n***-мерного случайного вектора.***

В частности, функция совместного распределения двумерного случайного вектора (ξ, η):

$$F(x, y) = P((\xi < x) \cap (\eta < y)). \qquad (1.2)$$

Из этого определения непосредственно следуют свойства функции распределения.

Свойство 1. Функция распределения монотонно не убывает по каждому аргументу.

Свойство 2. Справедливы предельные равенства:

$$\lim_{\substack{x \to -\infty \\ y \to -\infty}} F(x, y) = 0, \quad \lim_{\substack{x \to +\infty \\ y \to +\infty}} F(x, y) = 1. \qquad (1.3)$$

Свойство 3.

$$\lim_{y \to +\infty} F(x, y) = F_1(x) = P(\xi < x), \; \lim_{x \to +\infty} F(x, y) = F_2(y) = P(\eta < y).$$

Определение. ***Плотностью распределения*** двумерного случайного вектора называется вторая смешанная производная функции распределения:

$$f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y}, \tag{1.4}$$

если она существует. Ясно, что такая производная не всегда существует. В частности, для дискретного случайного вектора плотность распределения не существует.

Приведем выражение функции распределения через плотность распределения:

$$F(x,y)=\iint\limits_D f(x,y)dxdy=\int\limits_{-\infty}^{x}dx\int\limits_{-\infty}^{y}f(x,y)dy. \tag{1.5}$$

Свойства плотности распределения следуют непосредственно из ее определения и свойств функции распределения.

Свойство 1. В каждой точке (x,y) справедливо $f(x,y)\geq 0$.

Свойство 2. $\int\limits_{-\infty}^{+\infty}dx\int\limits_{-\infty}^{+\infty}f(x,y)dy=1.$ (1.6)

Свойство 3. $P((x,y)\in D)=\iint\limits_D f(x,y)dy.$ (1.7)

В случае, когда координаты случайного вектора есть дискретные случайные величины, ***совместное распределение двумерного случайного дискретного вектора*** удобно представлять в виде следующей таблицы.

Таблица 1

η \ ξ	x_1	x_2	...	x_{n_1}
y_1	p_{11}	p_{12}	...	p_{1n_1}
y_2	p_{21}	p_{22}	...	p_{2n_1}
...	...	...	...	...
y_{n_2}	$p_{n_2 1}$	$p_{n_2 2}$	...	$p_{n_2 n_1}$

Обозначения: $\{x_i\}$, $\{y_j\}$ – множества значений случайных величин ξ и η соответственно; $p_{i\,j}=P(\xi=x_i,\eta=y_j)$ – вероятность того, что случайный вектор (ξ,η) принимает значение (x_i, y_j).

Приведем свойства распределения дискретного случайного вектора, аналогичные приведенным выше свойствам 1 – 3 случайного вектора с координатами непрерывного типа. Распределение задано табл. 1.

Свойство 1. Для всех $i, j,\ 1 \le i \le n_1, 1 \le j \le n_2$ справедливо $0 \le p_{ij} \le 1$.

Свойство 2. $\sum_{i=1}^{i=n_1} \sum_{j=1}^{j=n_2} p_{ij} = 1$.

Свойство 3. $P(m_1 \le x_i \le m_2, k_1 \le y_j \le k_2) = \sum_{i=m_1}^{i=m_2} \sum_{j=k_1}^{j=k_2} p_{ij}$

Геометрический смысл двумерного случайного вектора состоит в том, что множество значений такого вектора определяет множество точек, координаты которых есть случайные величины (ξ, η).

Примером случайного вектора может служить пара (A, φ), где A – амплитуда, φ – фаза простого гармонического колебания с определенной частотой ω: $A\sin(\omega t + \varphi)$. Амплитуда A распределена по нормальному закону $N(a, \sigma)$, фаза φ распределена равномерно на отрезке $[0, \pi]$.

1.2 Частные распределения координат случайного вектора

Для изучения каждой координаты случайного вектора вводятся частные распределения его координат.

Определение. ***Частное распределение координаты*** ξ случайного вектора (ξ, η) определяется функцией: $F(x) = \lim_{y \to +\infty} F(x, y) = P(\xi < x)$.

Аналогично определяется частное распределение второй координаты η:

$F_2(y) = \lim_{x \to +\infty} F(x, y) = P(\eta < y)$.

Это определение верно и для непрерывного, и для дискретного случайного вектора. Однако в дискретном случае более удобно для вычислений пользоваться другой формулировкой, основанной на задании случайного вектора с помощью таблицы распределения, приведенной в п. 1.1 (см. табл.1). Частные распределения координат также получаем в виде таблиц:

Таблица 2

ξ	x_1	x_2	$\cdots$	x_{n_1}
P_i'	P_1'	P_2'	$\cdots$	P_{n_2}'

Таблица 3

η	y_1	y_2	$\cdots$	y_{n_2}
P_j''	P_1''	P_2''	$\cdots$	P_{n_2}'

Здесь элементы второй строки таблицы определяются формулами:

$$P_i' = \sum_{j=1}^{j=n_2} p_{ij}\ ,\ i = 1,2,\ldots,n_1; \qquad P_j'' = \sum_{i=1}^{i=n_1} p_{ij} \quad j = 1,2,\ldots,n_2. \quad (1.8)$$

Пример. Распределение случайного вектора (ξ,η) задано таблицей:

Найти частные распределения координат.

1.3 Условные распределения координат случайного вектора

Рассмотрим теперь распределение условных вероятностей каждой из координат при гипотезе, что другая координата приняли некоторое значение.

Определение. ***Условным распределением координаты*** ξ двумерного случайного вектора (ξ,η) при гипотезе $\eta = y$ называется распределение условных вероятностей $P(\xi < x / \eta = y)$. Аналогично для второй координаты η.

Такие распределения координат ***непрерывного*** случайного вектора определяются плотностями

$$f_1(x/\eta = y) = \frac{f(x,y)}{f_2(y)}, \quad f_2(y/\xi = x) = \frac{f(x,y)}{f_1(x)}, \qquad (1.9)$$

где $f(x,y)$ – плотность совместного распределения, $f_1(x)$, $f_2(y)$ – плотности частных распределений координат (ξ,η)..

Условные распределения координат ***дискретного*** случайного вектора определяются условными вероятностями:

$$P(\xi = x_i / \eta = y_j) = \frac{p_{ij}}{p_j''}, \quad P(\eta = y_j / \xi = x_i) = \frac{p_{ij}}{p_i''}. \qquad (1.10)$$

Вводим более краткие обозначения:

- $P_{i/j}' = \dfrac{p_{ij}}{P_j''}$ – условная вероятность того, что первая координата принимает значение с номером i при гипотезе, что вторая координата приняла значение с номером j;
- $P_{j/i}'' = \dfrac{p_{ij}}{P_i'}$ – условная вероятность того, что вторая координата принимает значение с номером j при гипотезе, что первая координата приняла значение с номером i

Приведем таблицы условных распределений координат двумерного случайного вектора: птаблица условных распределений первой координатs (табл. 4) и таблица условных распределений второй координаты (табл. 5).

Таблица 4

η \ ξ	x_1	x_2	$\cdots$	x_{n_1}
y_1	$P'_{1/1}$	$P'_{2/1}$	$\cdots$	$P'_{n_1/1}$
y_2	$P'_{1/2}$	$P'_{2/2}$	$\cdots$	$P'_{n_1/2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_{n_2}	P'_{1/n_2}	P'_{2/n_2}	$\cdots$	p_{n_1/n_2}

Таблица 5

η \ ξ	x_1	x_2	$\cdots$	x_{n_1}
y_1	$P''_{1/1}$	$P'_{2/1}$	$\cdots$	$P'_{n_1/1}$
y_2	$P''_{2/1}$	$P'_{2/2}$	$\cdots$	$P'_{n_1/2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_{n_2}	P'_{1/n_2}	P'_{2/n_2}	$\cdots$	p_{n_1/n_2}

Таким образом, со случайным двумерным вектором связаны три распределения:

1) Наиболее полной его характеристикой является функция ***совместного распределения*** координат случайного вектора.

2) Через совместное распределение вводятся ***частные распределения*** его координат.

3) ***Условные распределения*** координат вводятся с помощью совместного распределения и частных распределений.

1.4 Случайный вектор с независимыми координатами

Определение. Случайные величины ξ, η называются ***независимыми***, если функция совместного распределения случайного вектора (ξ, η) равна произведению функций частных распределений:

$$F(x, y) = F_1(x) F_2(y). \tag{4.1}$$

Очевидные свойства случайного вектора с независимыми координатами следуют из определения независимых случайных величин.

Свойство 1. Плотность совместного распределения координат непрерывного случайного вектора с независимыми координатами равна произведению плотностей частный распределений его координат:

$$f(x,y) = f_1(x)\,f_2(y). \tag{4.2}$$

Для дискретного вектора:

$$p_{ij} = p_i' p_j''. \tag{4.3}$$

Свойство 2. Условные плотности координат случайного вектора равны плотностям частных распределений его координат:

$$f_1(x/\eta = y) = f_1(x), \quad f_2(y/\xi = x) = f_2(y). \tag{4.4}$$

Для дискретного вектора это свойство можно выразить так:

$$P(\xi = x_i / \eta = y_j) = p_{ij}, \quad P(\eta = y_j / \xi = x_i) = p_{ij} \tag{4.5}$$

Иначе говоря, для вектора с независимыми координатами условные распределения совпадают с частными.

1.5. Моменты распределений

Для введенных выше распределений вводятся соответствующие им моменты. Сохраняя прежние [2] обозначения, приведем формулы, определяющие начальные и центральные моменты до второго порядка включительно:

- m_{10}, m_{01} – **первый *начальный момент*** (математическое ожидание) соответственно первой и второй координаты;
- m_{11} – ***второй смешанный начальный момент***;
- μ_{20}, μ_{02} – ***второй центральный момент*** (дисперсия) соответственно первой и второй координаты;
- μ_{11} – ***второй смешанный центральный момент***.

Ниже приводятся определения этих моментов для совместных распределений непрерывного и дискретного типов.

Непрерывный вектор:

$$m_{10} = \iint\limits_{\substack{-\infty<x<+\infty \\ -\infty<y<+\infty}} x\, f(x,y)dxdy, \quad m_{01} = \iint\limits_{\substack{-\infty<x<+\infty \\ -\infty<y<+\infty}} y\, f(x,y)dxdy.$$

$$\mu_{20} = \iint\limits_{\substack{-\infty<x<+\infty \\ -\infty<y<+\infty}} (x-m_{10})^2\, f(x,y)dxdy,$$

$$\mu_{02} = \iint\limits_{\substack{-\infty<x<+\infty \\ -\infty<y<+\infty}} (y-m_{01})^2\, f(x,y)dxdy.$$

$$\mu_{11} = \iint\limits_{\substack{-\infty<x<+\infty \\ -\infty<y<+\infty}} (x-m_{10})(y-m_{01})\, f(x,y)dxdy.$$

Дискретный вектор:

$$m_{10} = \sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} x_i p_{i\,j}, \quad m_{01} = \sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} y_j p_{i\,j}$$

$$\mu_{20} = \sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} (x_i - m_{10})^2 p_{i\,j}, \quad \mu_{02} = \sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} (y_j - m_{01})^2 p_{i\,j},$$

$$\mu_{11} = \sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} (x_i - m_{01})(y_j - m_{01}) p_{i\,j}.$$

2. Регрессионный анализ

2.1 Функции, линии и поверхности регрессии

Рассмотрим вначале двумерный случайный вектор. Условным распределениям координат такого вектора (ξ, η) соответствуют условные математические ожидания его компонент:

- $\varphi(x) = M(\eta/\xi=x)$ – условное математическое ожидание случайной величины η при условии, что другая случайная величина ξ приняла значение x,

- $\psi(y) = M(\xi/\eta=x)$ – условное математическое ожидание случайной величины ξ при условии, что другая случайная величина η приняла значение y.

Функции $\varphi(x)$ и $\psi(y)$ называются ***функциями регрессии*** случайного вектора (ξ, η), а уравнения

$$y = \varphi(x), \quad x = \psi(y) \tag{2.1}$$

– ***уравнениями регрессии***. Первое из них – уравнение регрессии η на ξ, второе – ξ на η. Геометрическое место соответствующих этим уравнениям точек называются ***линиями регрессии***.

Ясно, что для случайного вектора с независимыми компонентами линиями регрессии будут две взаимно перпендикулярные прямые, параллельные координатным осям и пересекающиеся в центре распределения – в точке с координатами $M\xi$, $M\eta$.

В некоторых случаях функции регрессии оказываются линейными. Тогда говорят, что имеет место ***линейная регрессия***.

Аналогично определяются уравнения регрессии трехмерного случайного вектора (ξ, η, ζ):

$$z = \varphi(x, y), \quad y = \psi(x, z), \quad x = \chi(y, z), \tag{2.2}$$

где $\varphi(x, y)$ – условное математическое ожидание случайной величи-ны ζ при условии что $\xi = x$, $\eta = y$; аналогично $y = \psi(x, z)$, $x = \chi(y, z)$ для η, ζ. Геометрическим образом уравнений (2.2) регрессии трехмерного случайного вектора (ξ, η, ζ) являются некоторые поверхности, которые называются ***поверхностями регрессии***.

Таким же образом вводятся уравнения регрессии случайного вектора $(\xi_1,\xi_2, \ldots , \xi_n)$ произвольного числа измерений.

Имеет место некоторая неоднозначность термина ***уравнение регрессии***.

2.2 Экстремальное свойство функции регрессии

Поставим следующую задачу. Задано совместное распределение двумерного случайного вектора (ξ, η). Требуется найти такие функции

$$y = f(x), \; x = g(y), \tag{2.3}$$

которые наилучшим в смысле среднеквадратического отклонения образом описывают зависимость между компонентами случайного вектора. Иначе говоря, определить функции (3.1) так, чтобы соответствующие моменты $M(\eta - f(\xi))^2$ и $M(\xi - g(y))^2$ достигали своего наименьшего значения. Для этого надо найти минимум функционалов

$$Q_1[f(x)] = M(\eta - f(\xi))^2, \quad Q_2[g(y)] = M(\xi - g(y))^2. \qquad (2.4)$$

Теорема. Минимум функционалам (3.4) доставляют функции регрессии (2.1).

Докажем это, ограничиваясь ради краткости и простоты рассмотрением двумерного дискретного распределения. Примем обозначения:

- $\{x_i\}$, $\{y_j\}$ – множества значений случайных величин ξ и η соответственно,
- $p_{ij} = P(\xi = x_i, \eta = y_j)$ – вероятность того, что случайный вектор (ξ,η) принимает значение (x_i, y_j),
- $P(\xi = x_i / \eta = y_j)$ – условная вероятность того, что $\xi = x_i$ при условии $\eta = y_j$,
- $P(\eta = y_j / \xi = x_i)$ – условная вероятность того, что $\eta = y_j$ при условии $\xi = x_i$.

Первое из минимизируемых выражений принимает вид

$$Q_1[f(x)] = M(\eta - f(\xi))^2 = \sum_{i,j} (y_j - f(x_i))^2 p_{ij} =$$
$$= \sum_{i,j} (y_j - f(x_i))^2 P(\xi = x_i, \eta = y_j). \qquad (2.5)$$

Заменяя вероятность $p_{i\,j}$ соответствующим произведением вероятностей (теорема умножения)

$$p_{ij} = P(\xi = x_i, \eta = y_j) = P(\xi = x_i)\, P(\eta = y_j / \xi = x_i),$$

получим:

$$M(\eta - f(\xi))^2 = \sum_i P(\xi = x_i) \sum_j (y_j - f(x_i))^2\, P(\eta = y_j / \xi = x_i). \quad (3.4)$$

Преобразуем внутреннюю сумму последнего выражения так:

$$\sum_j (y_j - f(x_i))^2\, P(\eta = y_j / \xi = x_i) =$$
$$= \sum_j (y_j - \varphi(x_i))^2\, P(\eta = y_j / \xi = x_i) +$$
$$+ \sum_j (\varphi(x_i) - f(x_i))^2\, P(\eta = y_j / \xi = x_i) +$$

$$+ 2\sum_j ((y_j - \varphi(x_i))(\varphi(x_i) - f(x_i)(\mathrm{P}(\eta = y_j / \xi = x_i).$$

Покажем, что последняя сумма равна нулю:

$$\sum_j (y_j - \varphi(x_i))(\varphi(x_i) - f(x_i)(\mathrm{P}(\eta = y_j / \xi = x_i) =$$

$$=(\varphi(x_i) - f(x_i))(\sum_j y_j \mathrm{P}(\eta = y_j /\xi = x_i) - \varphi(x_i)\sum_j \mathrm{P}(\eta = y_j /\xi = x_i) =$$

$$= (\varphi(x_i) - f(x_i))(\varphi(x_i) - \varphi(x_i)) = 0.$$

Здесь мы воспользовались тем, что $\varphi(x)$ – функция регрессии и по определению она равна условному математическому ожиданию:

$$\sum_j y_j \mathrm{P}(\eta = y_j / \xi = x_i) = \varphi(x_i),$$

а также очевидным равенством

$$\sum_j \mathrm{P}(\eta = y_j / \xi = x_i) = 1.$$

Теперь минимизируемое выражение (3.3) принимает вид:

$$\mathrm{Q}_1[f(y)] = \mathrm{M}(\eta - f(\xi))^2 =$$

$$= \sum_i \mathrm{P}(\xi = x_i)\sum_j ((y_j - \varphi(x_i))^2 + (\varphi(x_i) - f(x_i))^2)\mathrm{P}(\eta = y_j /\xi = x_i), \quad (2.6)$$

и становится ясно, что минимальное значение этому выражению среди всех возможных функций $f(x)$ доставляет функция регрессии $\varphi(x)$:

$$f(x) = \varphi(x).$$

Аналогично можно доказать, что минимальное значение функционалу $\mathrm{Q}_2[g(y)]$ доставляет функция регрессии $\psi(y)$:

$$g(y) = \psi(y).$$

Для функций регрессии n-мерного случайного вектора также справедливо экстремальное свойство, доказанное выше для случая двух измерений. Это свойство можно выразить с помощью следующего равенства:

$$\min_f \mathrm{Q}[f(\xi_1, \xi_2, ..., \xi_{n-1})] = \min_f \mathrm{M}(\xi_n - f(\xi_1, \xi_2, ... ,\xi_{n-1}))^2 =$$

$$= \mathrm{M}(\xi_n - \varphi(\xi_1, \xi_2, ... ,\xi_{n-1}))^2, \quad (2.7)$$

где $\varphi(x_1, x_2, \ldots, x_{n-1})$ – функция регрессии ξ_n на $\xi_1, \xi_2, \ldots, \xi_n$.

Экстремальное свойство функций регрессии может служить обоснованием прогноза значений случайной величины на основании известных значений других координат случайного вектора.

2.3. Парная линейная регрессия

Мы убедились, что наилучшим в смысле среднеквадратичного приближения одной случайной величины с помощью функции остальных координат случайного вектора является функция регрессии. Для двумерного случайного вектора наименьшее среди всех возможных функций $y = f(x)$, $x = g(y)$ значение критериям $Q_1[f(x)] = \mathrm{M}(\eta - f(\xi))^2$, $Q_2[g(y)] = \mathrm{M}(\xi - g(y))^2$ доставляют функции регрессии.

Естественно поставить задачу выбора оптимальных по критериям $Q_1[f(x)]$ и $Q_2[g(y)]$ соответственно фнкций $y = f(x)$, $x = g(y)$, ограничиваясь некоторым определенным заранее классом функций. Решение задачи в такой постановке известно под названием ***метода наименьших квадратов***. В этом случае функции $y = f(x)$ и $x = g(y)$, вообще говоря, не являются функциями регрессии в смысле данного выше определения (п. 2.1). Однако в приложениях для них сохраняют термин ***функции регрессии*** и ***линии регрессии*** с указанием соответствующего класса функций: линейная регрессии, параболическая регрессия и т.д. Разъяснение неоднозначности этих терминов можно найти на примере, который приводится п. 2.3. Особый интерес представляет собой среднеквадратическая регрессия в классе тригонометрических многочленов. Этому вопросу посвящается отдельный раздел в курсе математического анализа.

Мы ограничимся рассмотрением простейшего случая минимизации среднего квадрата в классе линейных функций. В этой ситуации говорят о ***линейной регрессии***. Заметим, что к линейной регрессии сводятся также случаи многих других классов функций.

Линейная среднеквадратичная регрессия в двумерном случае, в приложениях называемая парной линейной регрессией, приводит к следующей задаче.

Предположим, что для двумерного случайного вектора (ξ, η) известны все моменты до второго порядка включительно и дисперсии его компонент не равны нулю: $\mathrm{D}\xi \neq 0$, $\mathrm{D}\eta \neq 0$. Требуется найти такую линейную функцию

$$f(x) = \alpha x + \beta, \tag{2.8}$$

которая доставляет минимальное значение среднему квадрату

$$Q_1[(\alpha, \beta)] = M(\eta - (\alpha\xi + \beta))^2. \qquad (2.9)$$

Иначе говоря, требуется найти параметры α_0, β_0 так, чтобы выполнялось условие:

$$\min_{\alpha,\, \beta} M(\eta - (\alpha\xi + \beta))^2 = M(\eta - (\alpha_0\xi + \beta_0))^2. \qquad (2.10)$$

Переходим к принятым обозначениям для моментов двумерного случайного вектора:

- $m_{kl} = M(\xi^k \eta^l)$ – смешанный начальный момент порядка k, l;
- $\mu_{kl} = M((\xi - M\xi)^k (\eta - M\eta)^l)$ – смешанный центральный момент порядка k, l.

Преобразуем средний квадрат отклонения (2.9), используя введенные обозначения моментов:

$$M(\eta - (\alpha\xi + \beta))^2 =$$
$$= M((\eta - m_{01}) - \alpha(\xi - m_{10}) + m_{01} - (\alpha\, m_{10} + \beta))^2,$$

что приводит к выражению:

$$M(\eta - (\alpha\xi + \beta))^2 =$$
$$= \mu_{02} - \frac{\mu_{11}^2}{\mu_{20}} + \left(\sqrt{\mu_{20}}\beta - \frac{\mu_{11}}{\sqrt{\mu_{20}}}\right)^2 + (m_{01} - (\alpha\, m_{10} + \beta))^2. \qquad (2.11)$$

В соответствии с поставленной задачей минимизации выражения (2.9) за счет выбора α и β приходим к условиям:

$$\begin{cases} m_{10}\alpha + \beta = m_{01}, \\ \sqrt{\mu_{20}}\,\alpha = \dfrac{\mu_{11}}{\sqrt{\mu_{20}}}. \end{cases} \qquad (2.12)$$

Решением этой системы и будут требуемые значения α и β:

$$\alpha = \frac{\mu_{11}}{\mu_{20}}, \quad \beta = m_{01} - \frac{\mu_{11}}{\mu_{20}} m_{10}. \qquad (2.13)$$

Таким образом, доказана следующая теорема.

Теорема. Если случайный вектор (ξ, η) имеет все моменты до второго порядка включительно и при этом μ_{20} и μ_{02} не равны нулю, то среди всех линейных функций $y = \alpha x + \beta$ существует единственная линейная функция, при которой критерий

$$Q_1[(\alpha, \beta)] = \mathrm{M}(\eta - (\alpha\xi + \beta))^2$$

принимает наименьшее значение, которое достигается при следующих значениях α и β:

$$\alpha_0 = \frac{\mu_{11}}{\mu_{20}}, \quad \beta_0 = m_{01} - \frac{\mu_{11}}{\mu_{20}} m_{10}. \tag{2.14}$$

Иначе говоря, прямая линия $y = \alpha_0 x + \beta_0$ является прямой наилучшего среди всех прямых среднеквадратичного приближения зависимости координаты η от координаты ξ по критерию $Q_1[(\alpha, \beta)]$.

Абсолютно симметричный результат получим, выбирая линейную функцию $x = \gamma y + \delta$ наилучшего среднеквадратичного приближения зависимости координаты ξ от координаты η, имея в виду критерий

$$Q_2[(\gamma, \delta)] = \mathrm{M}(\xi - (\gamma\eta + \delta))^2.$$

Наименьшему значению критерия $Q_2[(\gamma, \delta)]$ соответствуют значения

$$\gamma_0 = \frac{\mu_{11}}{\mu_{02}}, \qquad \delta_0 = m_{10} - \frac{\mu_{11}}{\mu_{02}} m_{01}. \tag{2.15}$$

Для более удобного выражения α_0, β_0 преобразуем (1.15) и воспользуемся формулой

$$r(\xi, \eta) = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}}, \tag{2.16}$$

котрая определяет характеристику $r = r(\xi, \eta)$, называемую ***коэффициентом корреляции*** случайных величин ξ и η. В теории вероятностей мы уже познакомились с этой характеристикой. Более подробно вероятностный смысл и свойства этой характеристики мы рассмотрим в п. 2.2.1. Формулы (2.15) с учетом выражения (2.16) принимают вид:

$$\alpha_0 = \sqrt{\frac{\mu_{02}}{\mu_{20}}}\, r, \quad \beta_0 = m_{01} - m_{10}\alpha_0. \tag{2.17}$$

Уравнение прямой линии среднеквадратичной регрессии η на ξ

$$y = \alpha_0 x + \beta_0,$$

учитывая полученные значения α_0, β_0 (1.18), можем теперь переписать в виде:

$$\frac{y - m_{01}}{\sqrt{\mu_{02}}} = r\frac{x - m_{10}}{\sqrt{\mu_{20}}} . \tag{2.18}$$

Точно таким же образом получим уравнение прямой линии средней квадратичной регрессии ξ на η:

$$r\frac{y - m_{01}}{\sqrt{\mu_{02}}} = \frac{x - m_{10}}{\sqrt{\mu_{20}}} . \tag{2.19}$$

Теперь можно определить соответствующие минимальные значения критериев $Q_1[(\alpha, \beta)]$ и $Q_2[(\alpha, \beta)]$:

$$Q_1[(\alpha, \beta)] = \mathrm{M}(\eta - (\alpha\xi + \beta))^2, \quad Q_2[(\gamma, \delta)] = \mathrm{M}(\xi - (\gamma\eta + \delta))^2.$$

Из (2.11), (2.12) и (2.16) следует:

$$\min_{\alpha,\beta} \mathrm{M}(\eta - (\alpha\xi + \beta))^2 = \mu_{02}(1 - r^2), \tag{2.20}$$

$$\min_{\gamma,\delta} \mathrm{M}(\xi - (\gamma\eta + \delta))^2 = \mu_{20}(1 - r^2). \tag{2.21}$$

Эти минимальные значения среднеквадратичных отклонений называются ***остаточными дисперсиями.*** Более подробно об этих характеристиках см. п. 3.2. Приведем предельные значения остаточной дисперсии в зависимости от коэффициента корреляции r. Наибольшие значения:

$$\min_{\alpha,\beta} \mathrm{M}(\eta - (\alpha\xi + \beta))^2 = \mu_{02}, \tag{2.22}$$

$$\min_{\gamma,\delta} \mathrm{M}(\xi - (\gamma\eta + \delta))^2 = \mu_{20}, \tag{2.23}$$

если случайные величины ξ и η не коррелированы, т. е. $r = 0$.

Наименьшие значения:

$$\min_{\alpha,\beta} \mathrm{M}(\eta - (\alpha\xi + \beta))^2 = 0; \tag{2.24}$$

$$\min_{\gamma,\delta} \mathrm{M}(\xi - (\gamma\eta + \delta))^2 = 0. \tag{2.25}$$

Эти значения достигаются, если между случайными величинами имеет место линейная зависимость, т. е. $r^2 = 1$.

Геометрический смысл прямых линий среднеквадратической регрессии состоит в том, что эти прямые доставляют наименьшее среди всех прямых значение среднему квадрату расстояния некоторой системы материальных точек, если это расстояние измерять вдоль координатных осей.

По уравнениям этих прямых (2.18) и (2.19) можно определить угол φ между ними. Известными из аналитической геометрии средствами находим:

$$\operatorname{tg}\varphi = \frac{\sqrt{\mu_{20}\mu_{02}}}{\mu_{20}+\mu_{02}} \cdot \frac{1-r^2}{r}. \qquad (2.26)$$

Если $r^2 = 1$, то обе прямые среднеквадратичной регрессии совпадают. В этом случае, как мы видели из (2.24) и (2.25), остаточные дисперсии равны нулю. Это имеет место тогда (и только тогда), когда случайные величины ξ и η с вероятностью 1 связаны линейной зависимостью, т. е.

$$P(\eta = \alpha\xi + \beta) = 1. \qquad (2.27)$$

Перечислим теперь основные свойства прямых среднеквадратичной регрессии, которые следуют непосредственно из формул (2.19) – (2.22) и (2.26).

<u>Свойство 1</u>. Прямые регрессии проходят через точку (m_{10}, m_{01}).

<u>Свойство 2</u>. Если между случайными величинами имеет место собственная линейная зависимость, то линии регрессии совпадают.

<u>Свойство 3</u>. Если случайные величины не коррелированы, то прямые регрессии ортогональны.

<u>Свойство 4</u>. Угол φ между прямыми регрессии определяется формулой (2.27).

Рассмотрим пример вычисления моментов и уравнений регрессии двумерного случайного вектора.

П р и м е р. Распределение дискретного двумерного случайного вектора (ξ,η) задано таблицей. Здесь (x_i, y_j) – возможные значения координат случайного вектора (ξ,η), $p = P(\xi = 5, \eta = 1$ – неизвестная вероятность.

y_j ⁄ x_i	2	3	4	5
1	–	0,03	0,09	p
2	0,02	0,06	0,13	0,1
3	0,07	0,11	0,08	0,04
4	0,14	0,05	–	–

Вычисления приводятся с точностью до 10^{-4}.

1. Имея в виду свойство 2 дискретного совместного распределения $\sum_{i=1}^{i=n_1}\sum_{j=1}^{j=n_2} p_{ij} = 1$, получаем 0,92+ $p = 1$, поэтому $p = 0.08$, и имеем теперь полную таблицу совместного распределения двумерного случайного вектора.

	2	3	4	5
1	0	0,03	0,09	0,08
2	0,02	0,06	0,13	0,1
3	0,07	0,11	0,08	0,04
4	0,14	0,05	0	0

2. Частные распределения координат находим, пользуясь таблицами 2 и 3 и формулами 2.1. Таблицы частных распределений:

j	1	2	3	4
P_j''	0.2	0.31	0.3	0.19

i	2	3	4	5
P_i'	0.23	0.25	0.3	0.22

3. Находим моменты частных распределений координат.

$$M\eta = 1*0.2 + 2*0.31 + 3*0.3 + 4*0.19 = 2.48; \qquad M\xi = 3{,}51;$$

$$D\xi = M\xi^2 - (M\xi)^2 = 13{,}47 - 3{,}51^2 = 1{,}1499; \quad \sigma_\xi = 1{,}0723.$$

$$D\eta = 7.18 - (2.48)^2 = 7{,}18 - 6{,}1504 = 1{,}0296; \ \sigma_\eta = 1{,}0147.$$

4. Находим ковариацию и коэффициент корреляции.

$$\operatorname{cov}(\xi,\eta) = M(\xi\eta) - (M\xi)(M\eta); \qquad M(\xi\eta) = 8{,}02; \quad \operatorname{cov}(\xi,\eta) = -0{,}68;$$

$$r(\xi,\eta) = -\frac{0.68}{1.0723*1{,}0147} = -0{,}6250$$

Имеет место отрицательная корреляция.

5. Строим условные распределения координат случайного вектора. По формулам (3.2) находим условные вероятности, и результаты оформляем в виде таблиц условных распределений координат.

$$P(\eta = y_j / \xi = x_i) = \frac{p_{i\,j}}{P_i'}$$

$x\backslash y$	2	3	4	5
1	0	0,12	0,3	0,3636
2	0,0870	0,24	0,4333	0,4545
3	0,3043	0,44	0,2666	0,1818
4	0,6087	0,20	0	0

$$P(\eta = y_j / \xi = x_i) = \frac{p_{i\,j}}{P_i'}$$

$x\backslash y$	2	3	4	5
1	0	0,15	0,45	0,40
2	0,0645	0,1935	0,4194	0,3226
3	0,2333	0,3667	0,2667	0,1333
4	0,7368	0,2632	0	0

6. Находим условные математические ожидания и дисперсии координат.

Условные математические ожидания $M(\xi / \eta = y_j)$, j=1, 2, 3, 4, и $M(\eta / \xi = x_i)$, $i = 1, 2, 3, 4$, приводим в следующих таблицах.

$x\backslash y$	2	3	4	5	$M(\xi/\eta{=}y_j)$
1	0	0,15	0,45	0,40	$M(\xi/\eta{=}1)$=4,2500
2	0,0645	0,1935	0,4194	0,3226	$M(\xi/\eta{=}2)$=4,0001
3	0,2333	0,3667	0,2667	0,1333	$M(\xi/\eta{=}3)$=3,3000
4	0,7368	0,2632	0	0	$M(\xi/\eta{=}4)$=2,2632

$x\backslash y$	2	3	4	5
1	0	0,15	0,45	0,40
2	0,0645	0,1935	0,4194	0,3226
3	0,2333	0,3667	0,2667	0,1333
4	0,7368	0,2632	0	0
$M(\eta / \xi = x_i)$	$M(\eta / \xi = 2)$, =3,7761	$M(\eta / \xi = 3)$, =2,6899	$M(\eta / \xi = 4)$, =2,0889	$M(\eta / \xi = 5)$, =1,4451

7. Составляем линейные уравнения наилучшего квадратического приближения (по методу наилучших квадратов) к линиям регрессии.

В соответствии с формулами (1.19) и (1.20) п. 2.3 имеем:

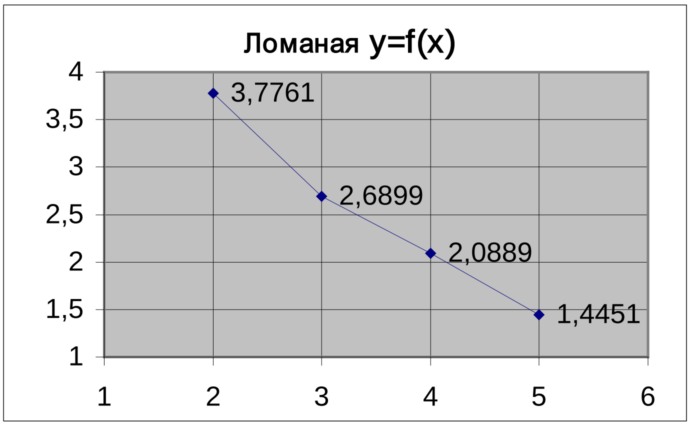
$$\frac{y-2{,}48}{1{,}0147}=-0.625\frac{x-3{,}51}{1{,}0723}; \qquad -0.625\frac{y-2{,}48}{1{,}0147}=\frac{x-3{,}51}{1{,}0723}.$$

После преобразований получаем линейные уравнения регрессии:

$$y=-0{,}615x+5{,}6686; \qquad x=-0.625y+5{,}0823.$$

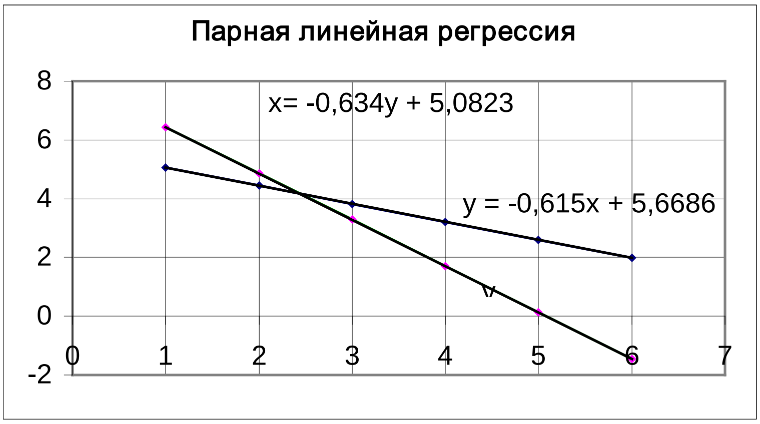
8. Ломаные линии регрессии строим по точкам на основании таблиц условных математических ожиданий.

x	2	3	4	5
y	3,7761	2,6899	2,0889	1,4451



x	4,2500	4,0001	3,3000	1,4451
y	1	2	3	4

1. Уравнения и прямые регрессии (линии наилучшего среднеквадратичного приближения) определяются уравнениями линейной регрессии



3. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

3.1. Линейная зависимость компонент случайного вектора

Рассмотрим теперь n-мерный случайный вектор

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$$

с произвольным распределением, имеющим все моменты до второго порядка включительно. Если $n = 2$, то как в п. 2.3 и 2.4, исследование проводится на ***парную регрессию***. В случае $n > 2$ имеет место ***множественная регрессия.***

Примем обозначения:

- $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)$ – математическое ожидание случайного вектора $\boldsymbol{\xi}$;
- $\boldsymbol{c} = (c_{ij})$ – матрица $n \times n$ вторых моментов случайного вектора $\boldsymbol{\xi}$ (матрица ковариаций).

Ради простоты будем полагать, что центр распределения в начале координат, т. е. $m_k = 0,\ k = 1, 2, \ldots, n.$ Выполнения этого условия легко добиться, переходя к новым координатам и полагая $\dot{\xi}_k = \xi_k - m_k$. Предположим также, что дисперсии всех координат не равны нулю: $\sigma_i^2 = c_{ii} \neq 0$.

Рассмотрим вначале два частных случая.

1. Существует $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ такое, что

$$\mathrm{P}(\boldsymbol{\alpha\xi}^{\mathbf{T}} = \mathbf{0}) = 1. \tag{3.1}$$

Иначе говоря, с вероятностью 1 справедливо равенство

$$\alpha_1\xi_1 + \alpha_2\xi_2 + \ldots + \alpha_n\xi_n = 0. \tag{3.2}$$

В этом случае будем говорить, что между координатами случайного вектора ξ имеет место собственная ***линейная зависимость***.

2. Рассмотрим теперь случай, когда координаты случайного вектора ***некоррелированные***, т. е. для всех i, j $c_{ij} = 0$, если $i \neq j$. Тогда матрица С – диагональная, $\sigma_i^2 = c_{ii} \neq 0$. Поэтому ее определитель не равен нулю. Очевидно, имеет место некоррелированность каждой компоненты случайного вектора с линейной комбинацией остальных.

Справедливы следующие теоремы.

Теорема 1. Чтобы между компонентами случайного вектора имела место собственная линейная зависимость, необходимо и достаточно, чтобы определитель матрицы его ковариаций был равен нулю.

Теорема 2. Чтобы существовала единственная функция линейной среднеквадратичной регрессии случайной величины ξ_n на $\xi_1, \xi_2, \ldots, \xi_{n-1}$, необходимо и достаточно, чтобы определитель матрицы ковариаций случайного вектора $(\xi_1, \xi_2, \ldots, \xi_{n-1})$ был не равен нулю.

Иначе говоря, для существования единственной линейной функции $\alpha_1\xi_1 + \alpha_2\xi_2 + \ldots + \alpha_{n-1}\xi_{n-1}$ наилучшего в смысле среднеквадратичного приближения случайной величины ξ_n, необходимо и достаточно, чтобы компоненты случайного вектора $(\xi_1, \xi_2, \ldots, \xi_{n-1})$ не были связаны линейной зависимостью.

Доказательство. По своему содержанию теоремы аналогичны соответствующим утверждениям линейной алгебры. Как видим, имеет место полная аналогия нашей задачи с задачей представления вектора в линейном векторном пространстве в виде линейной комбинации других векторов. Известно, что наиболее просто коэффициенты линейной комбинации определяются относительно ортогональной системы векторов. В нашем случае такой системой является система случайных некоррелированных величин. Заметим, что второй смешанный центральный момент обладает свойствами, аналогичными свойствам скалярного произведения векторов [11, 13].

Определим коэффициенты α_k^0 наилучшего среднеквадратичного приближения в ситуации, когда компоненты случайного вектора $(\xi_1, \xi_2, \ldots, \xi_{n-1})$ не коррелированы. Этот случай назовем ***ортогональным представ-***

лением случайной величины. Известными из линейной алгебры средствами получим:

$$\alpha_k^{\mathrm{o}} = \frac{c_{kn}}{\sigma_k^2}, \quad k = 1, 2, \ldots, n-1, \tag{2.3}$$

где c_{kn} – ковариация случайных величин ξ_n и ξ_k; σ_k^2 – дисперсия случайной величины ξ_k.

Удобнее записать выражение для коэффициентов (2.3) через коэффициент корреляции r_{nk} случайных величин ξ_n и ξ_k. Ради этого формулу (1.17) запишем в виде:

$$r_{nk} = \frac{c_{nk}}{\sigma_n \sigma_k}. \tag{2.4}$$

Тогда коэффициенты α_k^{o} наилучшего линейного среднеквадра-тичного приближения примут вид:

$$\alpha_k^{\mathrm{o}} = \frac{r_{nk}}{\sigma_k} \sigma_n , \tag{2.5}$$

и получим уравнение линии среднеквадратичной регрессии:

$$\frac{x_n}{\sigma_n} = \sum_{k=1}^{n-1} r_{nk} \frac{x_k}{\sigma_k} . \tag{2.6}$$

В общем случае, если отказаться от условия $M\xi_k = 0$ и положить $M\xi_k = m_k$, $k = 1, 2, \ldots n$, получим уравнение

$$\frac{x_n - m_n}{\sigma_n} = \sum_{k=1}^{n-1} r_{nk} \frac{x_k - m_k}{\sigma_k} . \tag{2.7}$$

Заметим, что уравнения линий регрессии (1.19) и (1.20) суть частный случай этого уравнения, когда $n = 2$.

3.2. Характеристики связи

Рассмотрим некоторые характеристики случайного вектора, особенно важные в задачах регрессионного анализа. Мы назовем их ***характеристиками связи***.

3.2.1. Коэффициент корреляции

Первой и важнейшей характеристикой связи является уже известный нам из теории вероятностей коэффициент корреляции. Здесь мы уже пользовались этой характеристикой в п. 1.4 и 2.1. ***Коэффициентом корреляции*** двух случайных величин ξ и η, имеющих ненулевые дисперсии $D\xi \neq 0$, $D\eta \neq 0$, называется характеристика, определяемая формулой

$$r(\xi,\eta)=\frac{M((\xi-M\xi)(\eta-M\eta))}{\sqrt{M(\xi-M\xi)^2 M(\eta-M\eta)^2}}\ . \tag{2.8}$$

Перечислим основные свойства этой характеристики.

Свойство 1. Коэффициент корреляции любой пары независимых случайных величин ξ, η по абсолютной величине не превосходит единицы: $|r(\xi, \eta)| \leq 1$.

Свойство 2. Коэффициент корреляции независимых случайных величин ξ, η равен нулю: $r(\xi, \eta) = 0$.

Свойство 3. Чтобы случайные величины были связаны собственной линейной зависимостью $P(A\xi + B\eta + C = 0) = 1$, необходимо и достаточно, чтобы их коэффициент корреляции по абсолютной величине равнялся единице: $| r(\xi, \eta) | = 1$.

Свойство 4. Коэффициент корреляции является характеристикой близости стохастической связи случайных величин к собственной линейной зависимости.

Более точно свойство 4 сформулируем ниже, когда будем рассматривать другую характеристику – ***остаточную дисперсию.***

П р и м е р. Даны три независимые случайные величины ξ, η_1, η_2.Найти коэффициент корреляции случайных величин ς и ξ, если $\varsigma = \xi + \eta_1 + \eta_2$ и $\xi \sim N(a_1, \sigma_1)$, $\eta_1 \sim N(a_2, \sigma_2)$, $\eta_2 \sim N(a_3, \sigma_3)$

В соответствии с определением имеем:

$$r(\varsigma,\xi)=\frac{M((\varsigma-M\varsigma)(\xi-M\xi))}{\sqrt{M(\varsigma-M\varsigma)^2 M(\xi-M\xi)^2}}=$$

$$= \frac{\mathrm{M}\left(\left(\xi + \eta_1 + \eta_2 - \mathrm{M}\xi - \mathrm{M}\eta_1 - \mathrm{M}\eta_2\right)(\xi - \mathrm{M}\xi)\right)}{\sqrt{\mathrm{D}\xi\,\mathrm{D}\left(\xi + \eta_1 + \eta_2\right)}} =$$

$$= \frac{1}{\sqrt{1 + \dfrac{\mathrm{D}\eta_1 + \mathrm{D}\eta_2}{\mathrm{D}\xi}}} = \frac{1}{\sqrt{1 + \dfrac{\sigma_2^2 + \sigma_3^2}{\sigma_1^2}}}.$$

Как видим, коэффициент корреляции зависит только от отношения дисперсий координат случайного вектора (ξ, η). Этот интересный результат следует запомнить. Приведем его интерпретацию в неформальных терминах. Пусть ξ – измеряемая величина, η – помеха. В какой мере эта помеха влияет на характеристику линейности зависимости случайных величин ξ (основная) и $\zeta = \xi + \eta$ (с учетом помехи)? Оказывается, что влияние помехи, как мы видели, измеряется отношением дисперсии помехи η к дисперсии величины ξ. Естественно, что в предельных случаях, когда $\mathrm{D}\eta = 0$ и $\mathrm{D}\eta = \infty$, коэффициент корреляции равен соответственно 1 и 0.

3.2.2. Корреляционное отношение

Чтобы ввести эту характеристику, представим дисперсию случайной величины η в следующем виде:

$$\mathrm{D}\eta = \mathrm{M}(\eta - \mathrm{M}\eta)^2 = \mathrm{M}(\eta - \varphi(\xi))^2 + \mathrm{M}(\varphi(\xi) - \mathrm{M}\eta)^2, \quad (2.9)$$

где $\varphi(x)$ – функция регрессии η на ξ.

Корреляционным отношением случайной величины η на случайную величину ξ называется отношение

$$Q^2_{\eta/\xi} = \frac{\mathrm{M}\left(\varphi(\xi) - \mathrm{M}\eta\right)^2}{\mathrm{M}\left(\eta - \mathrm{M}\eta\right)^2}. \quad (2.10)$$

Из определения корреляционного отношения (2.10) и выражения (2.9) замечаем, что

$$Q^2_{\eta/\xi} = 1 - \frac{\mathrm{M}\left(\eta - \varphi(\xi)\right)^2}{\mathrm{D}\eta}. \quad (2.11)$$

Свойства корреляционного отношения, легко выводимые из определения, выясняют его вероятностный смысл.

Свойство 1. Справедливо неравенство

$$0 \le Q^2_{\eta/\xi} \le 1.$$

Свойство 2. Чтобы случайные величины были независимы, необходимо и достаточно равенства нулю корреляционного отношения:

$$Q^2_{\eta/\xi} = 0.$$

Свойство 3. Чтобы между случайными величинами имела место собственная функциональная зависимость $\eta = \varphi(\xi)$, т.е.

$$P(\eta = \varphi(\xi)) = 1,$$

необходимо и достаточно, чтобы корреляционное отношение было равно единице:

$$Q^2_{\eta/\xi} = 1.$$

Свойство 4. Если регрессия η на ξ линейна, то корреляционное отношение равно квадрату коэффициента корреляции:

$$Q^2_{\eta/\xi} = r^2(\xi, \eta).$$

Таким образом, корреляционное отношение есть характеристика близости стохастической зависимости между случайными величинами к функциональной зависимости $\eta = \varphi(\xi)$, где $\varphi(x)$ есть функция регрессии η на ξ.

3.2.3. Остаток. Остаточная дисперсия

Пусть известно уравнение линейной среднеквадратичной регрессии η на $\xi_1, \xi_2, \ldots, \xi_n$:

$$y = \sum_{k=1}^{n} \alpha_k^o x_k. \qquad (2.12)$$

Рассмотрим соответствующую этому уравнению регрессии случайную величину η_r :

$$\eta_r = \sum_{k=1}^{n} \alpha_k^o \xi_k.$$

Случайная величина $\zeta = \eta - \eta_r$, определяемая формулой

$$\zeta = \eta - \sum_{k=1}^{n} \alpha_k^{0} \xi_k, \tag{2.13}$$

называется ***остатком*** случайной величины η относительно случайного вектора $(\xi_1, \xi_2, \ldots, \xi_n)$, а дисперсия этой случайной величины σ_ς^2 – ***остаточной дисперсией***:

$$\sigma_\varsigma^2 = \mathrm{D}\left(\eta - \sum_{k=1}^{n} \alpha_k^{0} \xi_k\right). \tag{2.14}$$

Как видно из формулы (1.21), в случае $n = 1$ остаточная дисперсия η относительно ξ такова:

$$\sigma_\varsigma^2 = \sigma_\eta^2 \left(1 - r^2(\xi, \eta)\right). \tag{2.15}$$

В общем n-мерном случае справедлива формула:

$$\sigma_\varsigma^2 = \sigma_n^2 \left(1 - r^2(\eta, \eta_r)\right). \tag{2.16}$$

Отметим очевидные свойства остаточной дисперсии, которые следуют из формулы (2.14).

Свойство 1. Наибольшее значение остаточная дисперсия принимает в случае, если η не коррелирована ни с одной из величин $\xi_1, \xi_2, \ldots, \xi_n$. В этом случае $r^2(\eta, \eta_r) = 0$ и остаточная дисперсия равна дисперсии η: $\sigma_\varsigma^2 = \mathrm{D}\eta$.

Свойство 2. Наименьшее значение остаточной дисперсии достигается при условии, что $|\,r(\eta, \eta_r)\,| = 1$, и оно равно нулю. Это имеет место в случае, если справедливо равенство

$$\eta = \sum_{k=1}^{n} \alpha_k^{0} \xi_k. \tag{2.17}$$

Свойство 3. Для двумерного случайного вектора остаточные дисперсии его компонент $\sigma_\varsigma^2(\xi)$ и $\sigma_\varsigma^2(\eta)$ ξ относительно η и η относительно ξ определяются формулами:

$$\sigma_\varsigma^2(\xi) = \mu_{02}\,(1 - r^2), \tag{2.18}$$

$$\sigma_\varsigma^2(\eta) = \mu_{20}\,(1 - r^2). \tag{2.19}$$

Ясно, что чем меньше остаточная дисперсия, тем точнее приближение случайной величины ξ с помощью соответствующего уравнения регрессии. Иногда в приложениях такое приближение характеризуют отношением $\sigma_\varsigma^2/\sigma_\eta^2$, где $0 \le \sigma_\varsigma^2/\sigma_\eta^2 \le 1$.

3.2.4. Сводный коэффициент корреляции

Сводным коэффициентом корреляции называется коэффициент корреляции случайной величины η и ее наилучшего в смысле среднеквадратичного линейного приближения η_r. Из формулы (2.16) следует:

$$r^2(\eta,\, \eta_r) = 1 - \sigma_\varsigma^2/\sigma_\eta^2. \tag{2.20}$$

Приведем свойства сводного коэффициента корреляции, вытекающие непосредственно из его определения.

Свойство 1. $-1 \le r(\eta,\, \eta_r) \le 1.$

Свойство 2. Равенство $r(\eta,\, \eta_r) = 0$ означает, что $r(\eta,\, \xi_k) = 0$ для всех k.

Свойство 3. Равенство $r(\eta,\, \eta_r) = 1$ означает, что имеет место собственная линейная зависимость

$$\eta = \sum_{k=1}^{n} \alpha_k^0 \xi_k\,.$$

Свойство 4. Если случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ не коррелированы, то справедливо равенство

$$r^2(\eta,\, \eta_r) = \sum_{k=1}^{n} r^2(\eta, \xi_k), \tag{2.21}$$

в котором легко узнать аналог ***теоремы Пифагора***.

П р и м е р. Найти характеристики связи для случайных величин ξ и η, если $\xi = \xi_1 + \xi_2$, $\eta = \xi + \zeta$ и известно, что ξ_1, ξ_2, ζ попарно не коррелированы и распределены нормально:

$$\xi_1 \sim N(a_1, \sigma_1),\ \xi_2 \sim N(a_2, \sigma_2),\quad \zeta \sim N(a_3, \sigma_3).$$

Коэффициент корреляции $r(\xi, \eta)$ был найден в п. 2.2.1. Для нашего случая имеем:

$$r(\xi_1, \eta) = \frac{1}{\sqrt{1 + D\eta / D\xi_1}} = \frac{\sigma_1}{\sqrt{\sigma_1 + \sigma_2 + \sigma_3}}.$$

$$r(\xi_2, \eta) = \frac{1}{\sqrt{1 + D\eta / D\xi_2}} = \frac{\sigma_2}{\sqrt{\sigma_1 + \sigma_2 + \sigma_3}}.$$

Определяя сводный коэффициент корреляции, получим:

$$r(\xi, \eta) = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}}.$$

Чтобы определить корреляционное отношение, найдем вначале функцию регрессии $\varphi(x, y)$:

$$\varphi(x, y) = M(\eta / \xi_1 = x, \xi_2 = y) = M(\xi + \zeta / \xi = x + y) = x + y + a_3.$$

В соответствии с тем, что регрессия, как видим, линейна, корреляционное отношение равно квадрату коэффициента корреляции:

$$Q^2_{\eta/\xi} = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}.$$

Остатком в соответствии с формулой (2.13) является случайная величина ζ, а ее дисперсия и будет остаточной дисперсией. Остаточная дисперсия определяется как дисперсия разности:

$$D(\eta - \xi) = D\zeta = \sigma_3^2.$$

3.3. Представление случайной величины в ортогональном базисе

Рассмотрим множество случайных величин таких, которые можно представить в виде линейной комбинации случайных величин ξ_1, ξ_2, …, ξ_n. Предположим, что для них существуют все моменты до второго порядка включительно и при этом все математические ожидания равны нулю, а дисперсии отличны от нуля.

Введем матрицу ковариаций, т.е. матрицу вторых смешанных моментов:

$$\mathrm{C} = (\mathrm{c}_{i\,j}), \quad 1 \le i, j \le n, \; \mathrm{c}_{i\,j} = \mathrm{M}(\xi_i \xi_j), \; \mathrm{c}_{i\,i} = \mathrm{D}\xi_i. \qquad (2.22)$$

Если случайные величины попарно не коррелированы, то C – диагональная матрица. Рассмотрим простейший случай: $n = 2$. Тогда матрица ковариаций C и корреляционная матрица R имеют диагональный вид:

$$\mathrm{C} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}; \quad \mathrm{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \qquad (2.23)$$

Элементом наилучшего в смысле среднего квадрата приближения случайной величины η (в предположении, что $\mathrm{M}\eta = 0$, $\mathrm{D}\eta = \sigma^2$) в ортогональном базисе ξ_1, ξ_2 будет линейная комбинация

$$\xi = \alpha_{10}\,\xi_1 + \alpha_{20}\,\xi_2, \qquad (2.24)$$

коэффициенты которой, как мы видели, определяются так:

$$\alpha_1^0 = \frac{c_1}{\sigma_1^2} = \frac{\sigma}{\sigma_1} r(\eta, \xi_1), \qquad \alpha_2^0 = \frac{c_2}{\sigma_2^2} = \frac{\sigma}{\sigma_2} r(\eta, \xi_2), \qquad (2.25)$$

где $c_1 = \mathrm{cov}(\eta, \xi_1)$, $c_2 = \mathrm{cov}(\eta, \xi_2)$. (Вспомним в линейной алгебре аналогичные формулы для коэффициентов разложения вектора по ортогональному базису на плоскости.)

Формулы (2.24) и (2.25) легко обобщаются на случай n-мерного пространства. Получаем разложение случайной величины η в системе произвольного числа n некоррелированных случайных величин. Здесь система $\xi_1, \xi_2, \ldots, \xi_n$ выступает в качестве ортогонального базиса. В n-мерном случае имеем:

$$\eta = \alpha_1{}^0\,\xi_1 + \alpha_2{}^0\,\xi_2 + \ldots + \alpha_n{}^0\,\xi_n, \qquad (2.26)$$

где $\alpha_k^0 = \dfrac{c_k}{\sigma_k^2}$, $k = 1, 2, \ldots, n$.

Особенно просто, как известно, выполняется разложение относительно ортонормированного базиса. Полагаем

$$\mathrm{cov}(\xi_i, \xi_j) = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases} \qquad (2.27)$$

Это означает, что система $\xi_1, \xi_2, \ldots, \xi_n$ попарно не коррелирована, и базисные величины имеют единичные дисперсии $\sigma_k^2 = 1$. В этом случае коэффициенты разложения (2.26) равны соответствующим коэффициентам корреляции:

$$\eta = r_{10}\xi_1 + r_{20}\xi_2 + \ldots + r_{n0}\xi_n. \qquad (2.28)$$

Геометрический смысл коэффициентов разложения (2.26) и (2.28) состоит в том, что они представляют собою проекции элемента η на соответствующий элемент ξ_k.

П р и м е р . Известны все моменты (до второго порядка включительно) системы случайных величин η, ξ_1, ξ_2, ξ_3.

Математические ожидания

5,112 ; 1,023; 2,005; 4,985.

Стандартные отклонения

0,490; 0,087; 0,170; 0,283.

Корреляционная матрица

$$R = \begin{pmatrix} 1{,}000 & 0{,}606 & 0{,}599 & -0{,}452 \\ 0{,}606 & 1{,}000 & 0{,}000 & 0{,}000 \\ 0{,}599 & 0{,}000 & 1{,}000 & 0{,}000 \\ -0{,}452 & 0{,}000 & 0{,}000 & 1{,}000 \end{pmatrix}.$$

В качестве ортогонального базиса можно выбрать попарно некоррелированные случайные величины ξ_1, ξ_2, ξ_3. Переходим к нормированным и центрированным величинам, которые образуют ортонормированный базис:

$$\xi_{10} = (\xi_1 - 1{,}023)/0{,}087;$$

$$\xi_{20} = (\xi_2 - 1{,}023)/0{,}170;$$

$$\xi_{30} = (\xi_3 - 4{,}985)/0{,}283.$$

В линейном пространстве случайных величин, натянутом на этот базис, находим элемент наилучшего среднеквадратичного приближения случайной величины η. Для этого составляем уравнение регрессии:

$(y - 5{,}112)/0.490 = 0{,}606(x_1 - 1{,}023)/0{,}087 +$

$$+ 0{,}599(x_2 - 1{,}023)/0{,}170 - 0{,}452(x_3 - 4{,}985)/0{,}283,$$

или короче

$$y = 3{,}422x_1 + 1{,}730x_2 - 0{,}784x_3 + 2{,}053.$$

Следовательно, элементом наилучшего среднеквадратического приближения случайной величины η в базисе ξ_1, ξ_2, ξ_3 будет случайная величина $\eta *$:

$$\eta * = 3{,}422\, \xi_1 + 1{,}730\, \xi_2 - 0{,}784\, \xi_3 + 2{,}053.$$

Коэффициенты корреляции

$$r_1 = 0{,}606; \qquad r_2 = 0{,}599; \qquad r_3 = -0{,}452$$

характеризуют степень влияния соответствующего фактора на характеристику η.

Точность описания этой характеристики с помощью параметров ξ_1, ξ_2, ξ_3 характеризуется остаточной дисперсией. Вычисляя остаточную дисперсию по формуле (3.18), получим:

$$\sigma_\eta^2 = 0{,}006.$$

Сводный коэффициент корреляции характеризует точность описания в относительных единицах. Вычисляя сводный коэффициент корреляции по формуле (2.21) (теорема Пифагора), получаем:

$$r(\eta, \eta *) = 0{,}965.$$

Близкое к единице значение сводного коэффициента корреляции говорит о высокой точности описания характеристики η с помощью выбранной линейной функции.

3.4. Представление случайной величины в произвольном базисе

Как мы уже заметили, выполнения условий

$$M\xi = 0, \quad D\xi = 1 \tag{2.29}$$

легко добиться, вводя замену

$$\xi' = (\xi - M\xi)/\sqrt{D\xi}\,. \tag{2.30}$$

Сложнее дело обстоит с выполнением условия (2.27). Здесь речь идет о переходе от произвольной системы к системе попарно некоррелированных случайных величин, т.е. к ортогональному базису. Этот процесс ортогонализации аналогичен известной процедуре Грамма – Шмидта, которая в линейной алгебре приводит к ортогональной системе векторов[9, 11, 13], а в математическом анализе – к ортогональной системе функций. Покажем эту процедуру, полагая для простоты $n = 2$.

Рассмотрим систему двух случайных величин ξ_1, ξ_2, для которых выполнены условия (2.29). Предположим, что ранг ее матрицы ковариаций $C = (c_{ij})$, $1 \le i, j \le 2$, равен двум, т. е. определитель этой матрицы не равен нулю. В противном случае, если $det C = 0$, из теоремы 1 п. 6 следует, что случайные величины ξ_1, ξ_2 линейно зависимы и потому $|r(\xi_1, \xi_2)| = 1$. Это тривиальный случай, базис составляет одна величина.

Для перехода к ортогональному базису ξ_{10}, ξ_{20} полагаем:

$$\xi_{10} = \xi_1,$$
$$\xi_{20} = \xi_2 + t\,\xi_{10}. \tag{2.31}$$

Выбираем t из условия некоррелированности ξ_{10}, ξ_{20}, получим:

$$M(\xi_{10}\,\xi_{20}) = r(\xi_1, \xi_2) + t = 0. \tag{2.32}$$

Поэтому в качестве ортогонального базиса можно принять

$$\xi_{10} = \xi_1, \quad \xi_{20} = \xi_2 - r\,\xi_1,$$

где $r = r(\xi_1, \xi_2)$ – коэффициент корреляции исходной системы величин. Однако, добиваясь выполнения условия (2.29) для новой системы, получаем:

$$\begin{cases} \xi_{10} = \xi_1, \\ \xi_{20} = \dfrac{1}{\sqrt{1+r^2}}\left(\xi_2 - r\xi_1\right). \end{cases} \tag{2.33}$$

Далее процедура выбора линии регрессии в базисе ξ_{10}, ξ_{20} происходит описанным выше образом. Обратный переход к старым координатам ξ_1, ξ_2 не представляет трудностей.

П р и м е р. В условиях предыдущего примера снимается предположение о некоррелированности случайных величин ξ_1, ξ_2, ξ_3. Это означает, что факторы, предположительно влияющие на характеристику η, не являются независимыми. Поэтому ξ_1, ξ_2, ξ_3 не могут представлять ортогональный базис.

Известна корреляционная матрица случайных величин η, ξ_1, ξ_2, ξ_3:

$$R = \begin{pmatrix} 1 & 0{,}606 & 0{,}599 & -0{,}452 \\ 0{,}606 & 1 & -0{,}351 & 0{,}000 \\ 0{,}599 & -0{,}351 & 1 & 0{,}000 \\ -0{,}452 & 0{,}000 & 0{,}000 & 1 \end{pmatrix}.$$

Как видим из корреляционной матрицы, между ξ_1 и ξ_2 имеет место некоторая корреляционная зависимость. Поэтому целесообразно перейти к ортонормированному базису и затем применить стандартную процедуру. Ради этого полагаем:

$$\begin{cases} \xi_{10} = \dfrac{\xi_1 - m_1}{\sigma_1}; \\ \xi_{20} = \dfrac{1}{\sqrt{1 + r^2(\xi_1, \xi_2)}} \left(\dfrac{\xi_2 - m_2}{\sigma_2} - r \dfrac{\xi_1 - m_1}{\sigma_1} \right); \\ \xi_{30} = \dfrac{\xi_3 - m_3}{\sigma_3}. \end{cases}$$

По этим формулам получаем выражение нового (ортогонального) базиса через старый:

$$\xi_{10} = 11{,}49\,\xi_1 - 11{,}75;$$
$$\xi_{20} = 31{,}65\,\xi_2 - 3{,}805\,\xi_1 - 52{,}62;$$
$$\xi_{30} = 5{,}88\,\xi_3 - 29{,}32.$$

ЗАДАНИЕ К ТИПОВОМУ РАСЧЕТУ

Задача №1

Дискретный двумерный случайный вектор (ξ, η) задан таблицей распределения.

ξ / η	2	3	4	5
1	–	0,03	0,09	p
2	0,02	0,06	0,13	0,1
3	0,07	0,11	0,08	0,04
4	0,14	0,05	–	–

Задание:

1. Найти неизвестную вероятность p.
2. Построить частные распределения координат случайного вектора.
3. Найти математические ожидания, дисперсии, и средние квадратические отклонения частных распределений координат.
4. Найти ковариацию и коэффициент корреляции. Сделать вывод.
5. Построить условные распределения координат случайного вектора.
6. Найти условные математические ожидания и дисперсии координат.
7. Составить линейное уравнение наилучшего приближения к линии регрессии.
8. Построить ломаные линии регрессии и прямые приближенной регрессии.

ВАРИАНТЫ ЗАДАНИЙ

№1

$\xi \backslash \eta$	2	3	4	5
1	0	0,03	0,09	p
2	0,02	0,06	0,13	0,10
3	0,07	0,11	0,08	0,04
4	0,14	0,05	0	0

№2

$\xi \backslash \eta$	2	4	5	8
3	0	0	0,07	0,20
5	0,02	0,09	0,12	p
8	0,06	0,09	0,08	0,03
10	0,11	0,07	0	0

№3

$\xi \backslash \eta$	– 4	– 2	1	4
– 2	0,15	0,05	0	0
– 1	0,06	0,14	0,07	0,04
1	0,02	0,05	0,12	P
2	0	0,01	0,04	0,16

№4

$\xi \backslash \eta$	3	6	9	12
1	0	0	0,08	0,13
3	0	0,05	0,13	0,06
5	0,08	0,12	0,07	0,03
7	0,15	0,06	0,04	p

№5

$\xi \backslash \eta$	4	5	6	7
2	p	0,07	0	0
5	0,04	0,15	0,05	0,04
8	0	0,06	0,12	0,09
11	0	0	0,04	0,17

№6

$\xi \backslash \eta$	– 8	– 6	– 4	– 2
1	0	0	P	0,20
5	0	0,05	0,21	0,03
9	0,02	0,19	0,04	0
13	0,16	0,03	0	0

№7

$\xi \backslash \eta$	3	4	5	6
2	0,11	0,06	0	p
5	0,08	0,13	0,08	0,04
8	0	0,07	0,15	0,07
11	0	0	0,06	0,14

№8

$\xi \backslash \eta$	2	3	5	8
1	0	0	0,06	0,16
3	0,01	0,09	0,12	0,08
5	0,06	0,13	0,07	0
7	0,14	0,08	p	0

№9

$\xi \setminus \eta$	1	2	4	7
– 7	0,17	0,08	0,03	0
– 5	0,07	0,12	0,06	0
– 3	0,02	0,04	0,07	0,05
– 1	0	0	0,07	0,14

№10

$\xi \setminus \eta$	2	4	6	8
1	0	0	0,07	0,12
4	0	0,04	p	0,08
7	0,09	0,15	0,06	0,02
10	0,13	0,07	0,01	0

№11

$\xi \setminus \eta$	1	4	8	11
2	0,18	0,06	P	0
4	0,06	0,16	0,06	0
6	0,01	0,07	0,14	0,04
8	0	0,02	0,05	0,11

№12

$\xi \setminus \eta$	– 8	– 6	– 4	– 2
1	0	0,02	0,05	0,12
3	0,03	0,07	0,14	0,09
5	0,06	0,15	0,08	0
7	0,13	p	0	0

№13

$\xi \setminus \eta$	2	4	5	8
1	0,12	0,06	0	0
3	0,07	0,13	0,08	0
5	0,02	P	0,15	0,08
7	0	0,01	0,07	0,14

№14

$\xi \setminus \eta$	– 8	– 6	– 4	– 2
1	0	0	P	0,20
5	0	0,05	0,21	0,03
9	0,02	0,19	0,04	0
13	0,16	0,03	0	0

№15

$\xi \backslash \eta$	2	3	4	5
1	0	0,03	0,09	p
2	0,02	0,06	0,13	0,10
3	0,07	0,11	0,08	0,04
4	0,14	0,05	0	0

Задача № 2

Известны распределения трёх независимых случайных величин ξ, η_1, η_2. Случайная величина ζ равна сумме: $\zeta = \xi + \eta_1 + \eta_2$, где ζ – измеряемая величина, ξ – результат измерения, η_1, η_2 – помехи.

Задание:

1. Вывести формулу для коэффициента корреляции $r(\zeta, \xi)$;
2. Составить линейное уравнение регрессии;
3. Вычислить остаточную дисперсию;
4. Построить график зависимости коэффициента корреляции $r(\zeta, \xi)$ от дисперсии аддитивной помехи $\eta = \eta_1 + \eta_2$;
5. Указать наиболее благоприятные для точности измерения значения параметров распределений ξ, η_1, η_2 в заданных диапазонах.

Варианты задания

Номер	Распределения			Параметры распределений		
варианта	ξ	η_1	η_2	ξ	η_1	η_2
1.	Э(λ)	$N(a,\sigma)$	$N(a,\sigma)$	$\lambda = 0{,}2 \pm 0{,}1$	$\sigma = 2 \pm 1$	$\sigma = 2 \pm 1$
2.	$N(a,\sigma)$	$N(a,\sigma)$	$N(a,\sigma)$	$\sigma = 2 \pm 1$	$\sigma = 2 \pm 1$	$\sigma = 2 \pm 1$
3.	$N(a,\sigma)$	$N(a,\sigma)$	Э(λ)	$\sigma = 2 \pm 1$	$\sigma = 2 \pm 1$	$\lambda = 0{,}2 \pm 0{,}1$
4.	$N(a,\sigma)$	Э(λ)	$N(a,\sigma)$	$\sigma = 2 \pm 1$	$\lambda = 0{,}2 \pm 0{,}1$	$\sigma = 2 \pm 1$
5.	Э(λ)	Э(λ)	$N(a,\sigma)$	$\lambda = 0{,}2 \pm 0{,}1$	$\lambda = 0{,}2 \pm 0{,}1$	$\sigma = 2 \pm 1$
6.	$N(a,\sigma)$	$N(a,\sigma)$	$N(a,\sigma)$	$\sigma = 1{,}5 \pm 0{,}5$	$\sigma = 1{,}5 \pm 0{,}5$	$\sigma = 1{,}5 \pm 0{,}5$
7.	Э(λ)	$N(a,\sigma)$	$N(a,\sigma)$	$\lambda = 0{,}15 \pm 0{,}05$	$\sigma = 1{,}5 \pm 0{,}5$	$\sigma = 1{,}5 \pm 0{,}5$
8.	$N(a,\sigma)$	Э(λ)	$N(a,\sigma)$	$\sigma = 1{,}5 \pm 0{,}5$	$\lambda = 0{,}15 \pm 0{,}05$	$\sigma = 1{,}5 \pm 0{,}5$
9.	$N(a,\sigma)$	$N(a,\sigma)$	Э(λ)	$\sigma = 3 \pm 1$	$\sigma = 3 \pm 1$	$\lambda = 0{,}2 \pm 0{,}1$
10.	Э(λ)	Э(λ)	$N(a,\sigma)$	$\lambda = 0{,}15 \pm 0{,}05$	$\lambda = 0{,}15 \pm 0{,}05$	$\sigma = 2 \pm 1$

11.	$Э(\lambda)$	$N(a,\sigma)$	$N(a,\sigma)$	$\lambda = 0{,}15 \pm 0{,}05$	$\sigma = 2 \pm 1$	$\sigma = 2 \pm 1$
12.	$N(a,\sigma)$	$N(a,\sigma)$	$N(a,\sigma)$	$\sigma = 3 \pm 1$	$\sigma = 3 \pm 1$	$\sigma = 3 \pm 1$
13.	$N(a,\sigma)$	$N(a,\sigma)$	$Э(\lambda)$	$\sigma = 3 \pm 1$	$\sigma = 3 \pm 1$	$\lambda = 0{,}15 \pm 0{,}05$
14.	$N(a,\sigma)$	$Э(\lambda)$	$N(a,\sigma)$	$\sigma = 3 \pm 1$	$\lambda = 0{,}2 \pm 0{,}1$	$\sigma = 3 \pm 1$
15.	$Э(\lambda)$	$Э(\lambda)$	$N(a,\sigma)$	$\lambda = 0{,}2 \pm 0{,}1$	$\lambda = 0{,}2 \pm 0{,}1$	$\sigma = 3 \pm 1$
16.	$N(a,\sigma)$	$N(a,\sigma)$	$N(a,\sigma)$	$\sigma = 5 \pm 1$	$\sigma = 3 \pm 1$	$\sigma = 4 \pm 1$
17.	$N(a,\sigma)$	$N(a,\sigma)$	$Э(\lambda)$	$\sigma = 4 \pm 1$	$\sigma = 5 \pm 1$	$\lambda = 0{,}4 \pm 0{,}1$
18.	$N(a,\sigma)$	$Э(\lambda)$	$N(a,\sigma)$	$\sigma = 3 \pm 1$	$\lambda = 0{,}4 \pm 0{,}1$	$\sigma = 5 \pm 1$
19.	$Э(\lambda)$	$Э(\lambda)$	$N(a,\sigma)$	$\lambda = 0{,}4 \pm 0{,}1$	$\lambda = 0{,}4 \pm 0{,}1$	$\sigma = 4 \pm 1$