Тема 5. Многомерные распределения

5.1. Пара случайных величин, совместные и частные распределения

Напомним, что распределением дискретной случайной величины X называется набор пар (x_i, p_i), где x_i — это всевозможные значения, принимаемые X, $p_i = \mathbf{P}(X = x_i)$ — соответствующие вероятности. Обобщим понятие распределения на дискретный двумерный случайный вектор (X, Y).

Определение. *Совместным распределением* двух дискретных случайных величин X и Y (или *распределением дискретного случайного вектора* (X, Y)) называется набор всевозможных пар их значений (x_i, y_j), $i, j = 1, 2, \ldots,$ и соответствующих вероятностей $r_{ij} = \mathbf{P}(X = x_i, Y = y_j)$, где $\sum r_{ij} = 1$.

Можно представлять, что над каждой точкой плоскости с координатами (x_i, y_j) возвышается столбик высоты r_{ij}, если $r_{ij} > 0$ (рис. 1).

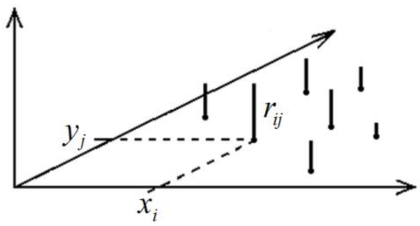

Рис. 1

Совместное распределение дискретных случайных величин с конечным числом значений удобно задавать перекрёстной таблицей следующего вида:

$X \setminus Y$	y_1	$\ldots$	y_m
x_1	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	r_{ij}	$\ldots$
x_n	$\ldots$	$\ldots$	$\ldots$

Пример 1. Симметричная монета бросается дважды. Здесь $\omega = (i_1, i_2)$, $i_k = 0$ или 1. Тогда $|\Omega|$ = 4. Положим $X(\omega) = i_1 + i_2$, $Y(\omega) = i_1 i_2$. Прежде чем строить перекрёстную таблицу, составим таблицу значений каждой из величин X и Y на всех возможных элементарных событиях ω:

ω	$X(\omega)$	$Y(\omega)$
00	0	0
01	1	0
10	1	0
11	2	1

Подсчитывая количество ω, на которых принимается каждое из трёх возможных значений $(0, 0)$, $(1, 0)$ и $(2, 1)$ случайного вектора (X, Y), получим следующую перекрёстную таблицу:

$X \setminus Y$	0	1
0	1/4	0
1	1/2	0
2	0	1/4

Определение. *Частные* (*маргинальные*[1]) *распределения* случайных величин X и Y вычисляются на основе совместного распределения, соответственно, по формулам

$$p_i = \mathbf{P}(X = x_i) = \sum_j r_{ij}, \qquad q_j = \mathbf{P}(Y = y_j) = \sum_i r_{ij}. \tag{1}$$

Формулы (1) следуют из аксиомы счётной аддитивности вероятностной меры. Если совместное распределение пары случайных величин X и Y задано таблицей, то для вычисления частного распределения X (Y), следует просуммировать вероятности r_{ij} по строкам (столбцам) этой таблицы.

Вопрос 1. Какое (частное) распределение имеет случайная величина: а) X; б) Y из примера 1?

В общем случае совместное распределение случайных величин X и Y задаётся с помощью *совместной функции распределения*

$$F_{X,Y}(x, y) = \mathbf{P}(X \le x, Y \le y), \tag{2}$$

где аргументы x и y принимают произвольные действительные значения.

На основе функции $F_{X,Y}(x, y)$ частные функции распределения случайных величин X и Y согласно свойству непрерывности вероятностной меры вычисляются, соответственно, по формулам

$$F_X(x) = \lim_{y\to+\infty} F_{X,Y}(x, y); \qquad F_Y(y) = \lim_{x\to+\infty} F_{X,Y}(x, y). \tag{3}$$

Пример 2. Пусть X и Y — координаты точки, взятой наудачу в квадрате $[0, 1]^2$. Тогда при $0 < x < 1$ и $0 < y < 1$ совместная функция распределения, очевидно, задаётся формулой $F_{X,Y}(x, y) = xy$.

Вопрос 2. Какой формулой выражается эта функция при: а) $0 < x < 1$, $y > 1$; б) $x > 1$, $0 < y < 1$?

Напомним, что для произвольной случайной величины X и любых чисел $a < b$ верна формула $\mathbf{P}(X \in (a, b]) = F_X(b) - F_X(a)$. Эту формулу нетрудно обобщить на двумерный случай.

Вопрос 3. Как выразить через значения совместной функции распределения $F_{X,Y}(x, y)$ вероятность попадания произвольного случайного вектора (X, Y) в прямоугольник $(a, b] \times (c, d]$?

Умея вычислять вероятности попадания произвольного случайного вектора (X, Y) в любые прямоугольники на плоскости со сторонами, параллельными координатным осям, можно найти вероятности попадания в более сложные множества, например, в треугольник $\{x > 0,\ y > 0,\ x + y < 1\}$. Этот треугольник можно представить как счётное объединение непересекающихся квадратов (рис. 2) и воспользоваться аксиомой счётной аддитивности вероятностной меры из темы 3.

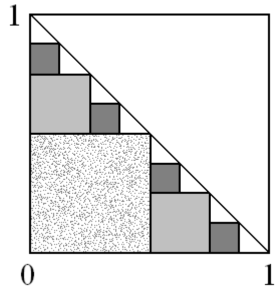

Рис. 2

[1] Marginal — (англ.) краевой, на полях.

В непрерывном случае также имеется другой способ вычисления вероятностей попадания случайного вектора (X, Y) в произвольные множества на плоскости. Этот способ базируется на понятии совместной плотности распределения.

Определение. Если для функции $f_{X,Y}(x, y) \geq 0$ и произвольного множества A, имеющего площадь, выполняется представление

$$\mathbf{P}((X,Y) \in \mathrm{A}) = \iint\limits_{\mathrm{A}} f_{X,Y}(x,y)\,dx\,dy, \tag{4}$$

то говорят, что случайные величины X и Y имеют *совместную плотность распределения* $f_{X,Y}(x, y)$. Иначе говорят, что случайный вектор (X, Y) имеет *плотность распределения* $f_{X,Y}(x, y)$. Слово «распределения» часто опускают ради краткости.

Что означает интеграл в правой части формулы (4)? Для произвольной функции $g(x, y)$ интеграл $\iint\limits_{\mathrm{A}} g(x,y)\,dx\,dy$ имеет следующий геометрический смысл — это объём цилиндра над множеством A, ограниченного сверху «шапочкой», высекаемой цилиндром из поверхности, заданной формулой $z = g(x, y)$ (рис. 3).

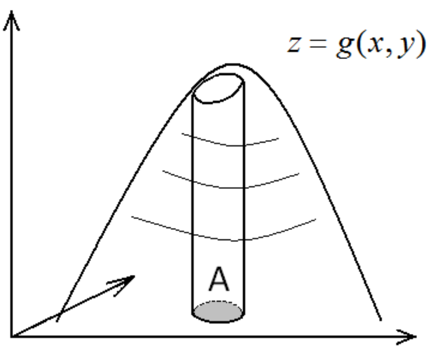

Рис. 3

Формально интеграл определяется следующим образом:

$$\iint\limits_{\mathrm{A}} g(x,y)\,dx\,dy = \lim_{n\to\infty} \sum_{(x_i, y_j)\in \mathrm{A}} g(x_i, y_j)\,d_n, \tag{5}$$

где (x_i, y_j) — середины квадратных ячеек со сторонами длины $1/n$, попавшие внутрь множества A (фигуры, ограниченной замкнутым контуром на рис. 4), $d_n = 1/n^2$ — площадь отдельной ячейки.

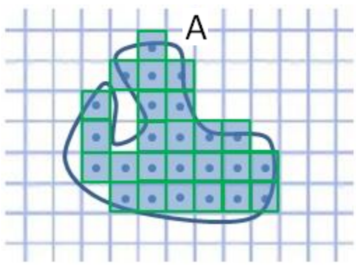

Рис. 4

На рис. 5 изображены две плотности: простая одновершинная и сложная многовершинная.

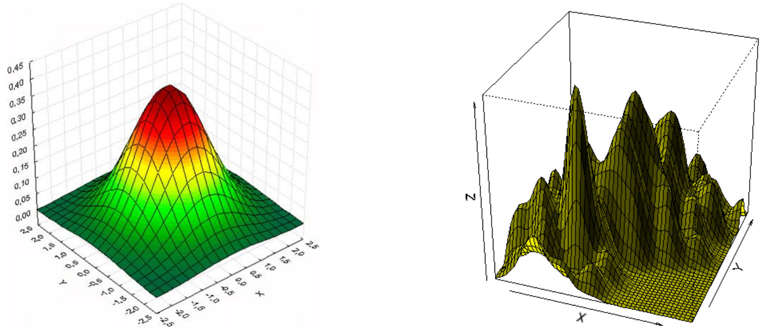

Рис. 5

Если плотность случайного вектора (X, Y) существует, то её можно вычислить по формуле

$$f_{X,Y}(x,y) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} F_{X,Y}(x,y)\right). \qquad (6)$$

В частности, случайный вектор (X, Y) из примера 2 с функцией распределения $F_{X,Y}(x,y) = xy$ на квадрате $[0,1]^2$ согласно формуле (6) имеет плотность $f_{X,Y}(x,y) = I_{[0,1]^2}$ — индикатор $[0,1]^2$.

Обратно, функцию распределения $F_{X,Y}(x,y)$ можно восстановить по плотности $f_{X,Y}(x,y)$. Действительно, согласно формуле (4) для множества $\mathsf{A} = (-\infty, x] \times (-\infty, y]$ имеем:

$$F_{X,Y}(x,y) = \mathbf{P}(X \le x, Y \le y) = \iint\limits_{\mathsf{A}} f_{X,Y}(x,y)\,dx\,dy = \int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f_{X,Y}(x,y)\,dx\,dy.$$

В свою очередь, частные плотности компонент случайного вектора (X, Y) вычисляются согласно формулам

$$f_X(x) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx, \qquad (7)$$

которые являются непрерывными аналогами формул (1).

Пример 3. Пусть совместная функция распределения случайных величин X и Y задана формулой

$$F_{X,Y}(x,y) = (1 - e^{-x} - e^{-y} + e^{-x-y}) I_{\{x>0,\, y>0\}}.$$

Вопрос 4. Какой формулой задаётся плотность $f_{X,Y}(x,y)$?

Вопрос 5. Какой формулой задаётся: а) функция распределения $F_X(x)$; б) плотность $f_X(x)$?

5.2. Ковариация и коэффициент корреляции

Важнейшими характеристиками силы связи (зависимости) случайных величин X и Y служат ковариация $\mathrm{cov}(X,Y)$ и коэффициент корреляции $\rho(X,Y)$. Приведём определяющие их формулы:

$$\mathrm{cov}(X,Y) = \mathsf{M}((X - \mathsf{M}X)(Y - \mathsf{M}Y)) = \mathsf{M}XY - \mathsf{M}X \cdot \mathsf{M}Y. \qquad (8)$$

$$\rho(X,Y) = \mathrm{cov}(X,Y) \Big/ \sqrt{\mathsf{D}X \cdot \mathsf{D}Y}. \qquad (9)$$

Вопрос 6. Как доказать второе равенство в формуле (8)?

Для дискретного случайного вектора (X, Y) математическое ожидание $\mathsf{M}XY$ вычисляется по формуле

$$\mathsf{M}XY = \sum_i \sum_j x_i y_j \mathbf{P}(X = x_i, Y = y_j) = \sum_i \sum_j x_i y_j r_{ij}. \quad (10)$$

Для случайного вектора (X, Y), имеющего плотность $f_{X,Y}(x, y)$,

$$\mathsf{M}XY = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, f_{X,Y}(x, y)\, dxdy. \quad (11)$$

В силу *неравенства Коши — Буняковского — Шварца*

$$|\mathsf{M}XY| \le \sqrt{\mathsf{M}X^2 \cdot \mathsf{M}Y^2} \quad (12)$$

для коэффициента корреляции выполняется неравенство

$$|\rho(X,Y)| \le 1.$$

Для доказательства достаточно подставить в формулу (12) случайные величины $X - \mathsf{M}X$ и $Y - \mathsf{M}Y$.

Пусть $Y = a + bX$, где a и b — константы. Если $b > 0$, т. е. если X и Y положительно линейно связаны, то $\rho(X, Y) = 1$ (см. задачу 5.5). Если $b < 0$, т. е. если X и Y отрицательно линейно связаны, то $\rho(X, Y) = -1$. Таким образом, для линейно связанных случайных величин X и Y в неравенстве $-1 \le \rho(X, Y) \le 1$ достигаются границы. При значениях $\rho(X, Y) \approx 0$ линейная связь между случайными величинами слабая или вообще отсутствует.

Свойства ковариации

1) $\operatorname{cov}(a + bX,\ c + dY) = bd \operatorname{cov}(X, Y)$, где a, b, c, d — любые константы.
2) $\rho(a + bX,\ c + dY) = \rho(X, Y)$, если $b > 0$ и $d > 0$.
3) $\mathsf{D}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathsf{D}X_i + 2 \sum_{1 \le i < j \le n} \operatorname{cov}(X_i, X_j)$.

Свойство 2 выражает инвариантность коэффициента корреляции к преобразованиям сдвига-масштаба шкал измерений показателей X и Y. Свойство 3 является наиболее важным. Присутствие ковариации в этой формуле в значительной степени объясняет появление понятия «ковариация».

5.3. Случайные векторы, совместное распределение компонент, матрица ковариаций

Понятия совместного распределения, функции распределения и плотности без труда обобщаются с двумерного случая на n-мерный.

Определение. Вектор $\boldsymbol{X} = (X_1, \ldots, X_n)$, компонентами которого являются случайные величины, называется *n-мерным случайным вектором*.

Определение. *Распределением* дискретного случайного вектора $\boldsymbol{X}$ называется набор всевозможных значений $x_1, \ldots, x_n$ его компонент и соответствующих вероятностей $\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n)$.

Как задаётся распределение случайного вектора в общем случае? Для произвольных действительных чисел $x_1, \ldots, x_n$ рассмотрим множество

$$\{\omega : X_1(\omega) \le x_1, \ldots, X_n(\omega) \le x_n\} = \bigcap_{i=1}^{n} \{\omega : X_i(\omega) \le x_i\}. \tag{13}$$

Ввиду того, что X_i — случайные величины, множества $\{\omega : X_i(\omega) \le x_i\}$ являются событиями. Поэтому их пересечение, стоящее в правой части формулы (13), также является событием ввиду замкнутости класса событий относительно операции пересечения (см. раздел 3.1). Следовательно, вероятность $\mathbf{P}(X_1 \le x_1, \ldots, X_n \le x_n)$ определена для произвольных действительных чисел $x_1, \ldots, x_n$.

Определение. Функция n переменных $F_{\boldsymbol{X}}(x_1, \ldots, x_n) = \mathbf{P}(X_1 \le x_1, \ldots, X_n \le x_n)$ называется *функцией распределения* n-мерного случайного вектора $\boldsymbol{X}$.

Определение. *Плотность* $f_{\boldsymbol{X}}(x_1, \ldots, x_n) \ge 0$ n-мерного случайного вектора $\boldsymbol{X}$ определяется как такая функция, что для произвольного n-мерного множества А, имеющего n-мерный объём, выполняется представление

$$\mathbf{P}(\boldsymbol{X} \in \mathrm{A}) = \int \ldots \int_{\mathrm{A}} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n.$$

Пример 4. Обобщим пример 1 из темы 2, в котором было определено биномиальное распределение. Пусть в урне находятся l_1 занумерованных шаров 1-го цвета, l_2 занумерованных шаров 2-го цвета, $\ldots$, l_k занумерованных шаров цвета k-го цвета, причём $l_1 + \ldots + l_k = m$. Из урны наудачу с возвращением выбираются n шаров. Тогда $|\Omega| = m^n$. Определим случайную величину X_j как число шаров j-го цвета среди n выбранных шаров, $j = 1, \ldots, k$.

Найдём распределение дискретного случайного вектора $\boldsymbol{X} = (X_1, \ldots, X_k)$, т. е. подсчитаем $\mathbf{P}(X_1 = i_1, \ldots, X_k = i_k)$ для целых чисел $i_1 \ge 0, \ldots, i_k \ge 0$, удовлетворяющих условию $i_1 + \ldots + i_k = n$.

Выбрать среди n мест подмножество из i_1 мест для шаров 1-го цвета можно $C_n^{i_1}$ способами. Для каждого из этих мест имеется l_1 вариантов выбора номера шара 1-го цвета. Всего — $C_n^{i_1} l_1^{i_1}$ вариантов. Далее, выбрать среди $(n - i_1)$ оставшихся мест подмножество из i_2 мест для шаров 2-го цвета можно $C_{n-i_1}^{i_2}$ способами. Для каждого из этих мест имеется l_2 вариантов выбора номера шара 2-го цвета. Всего — $C_{n-i_1}^{i_2} l_2^{i_2}$ вариантов. Продолжая аналогично и перемножая количества вариантов для всех k цветов, получим:

$$\mathbf{P}(X_1 = i_1, \ldots, X_k = i_k) = \frac{C_n^{i_1} l_1^{i_1} C_{n-i_1}^{i_2} l_2^{i_2} C_{n-i_1-i_2}^{i_3} l_3^{i_3} \cdot \ldots \cdot 1 \cdot l_k^{i_k}}{m^n} = C_n^{i_1} \frac{l_1^{i_1}}{m^{i_1}} C_{n-i_1}^{i_2} \frac{l_2^{i_2}}{m^{i_2}} C_{n-i_1-i_2}^{i_3} \frac{l_3^{i_3}}{m^{i_3}} \cdot \ldots \cdot 1 \cdot \frac{l_k^{i_k}}{m^{i_k}}.$$

Выражая числа сочетаний через факториалы и используя обозначения $p_j = l_j / m$, находим:

$$\mathbf{P}(X_1 = i_1, \ldots, X_k = i_k) = \frac{n!}{i_1! \cancel{(n - i_1)!}} \cdot \frac{\cancel{(n - i_1)!}}{i_2!(n - i_1 - i_2)!} \cdot \ldots \cdot 1 \cdot p_1^{i_1} p_2^{i_2} \ldots p_k^{i_k}.$$

Перекрёстно сокращая факториалы (*штриховые линии в формуле сверху*), окончательно выводим:

$$\mathbf{P}(X_1 = i_1,\ \ldots,\ X_k = i_k) = \frac{n!}{i_1!\, i_2! \ldots i_k!}\, p_1^{i_1} p_2^{i_2} \ldots p_k^{i_k}. \tag{14}$$

Определение. Формула (14) задаёт *полиномиальное распределение* (*multinomial distribution*).

Биномиальное распределение является частным случаем при $k = 2.$ Если $p_1 = p_2 = \ldots = p_k = 1/k,$ то получаем совместное распределение количеств попаданий X_j в ящики при размещении наудачу n занумерованных шаров по k ящикам:

$$\mathbf{P}(X_1 = i_1,\ \ldots,\ X_k = i_k) = \frac{n!}{i_1!\, i_2! \ldots i_k!}\, k^{-n}.$$

В заключение обобщим понятия математического ожидания и дисперсии на векторный случай. *Математическим ожиданием* случайного вектора $\boldsymbol{X}$ называется n-мерный числовой вектор $\mathsf{M}\boldsymbol{X} = (\mathsf{M}X_1, \ldots, \mathsf{M}X_n)$. Аналогом дисперсии является квадратная *ковариационная матрица* $\mathrm{cov}(\boldsymbol{X})$ размерности $n \times n$, элементами которой служат $\mathrm{cov}(X_i,\ X_j)$. Ковариационная матрица симметрична. На её главной диагонали располагаются дисперсии компонент $\mathsf{D}X_i = \mathrm{cov}(X_i,\ X_i),\ \ i = 1, \ldots, n.$

Задачи для решения на занятии

1) Симметричную монету бросают 3 раза, отмечая результат каждого бросания знаком плюс или минус в зависимости от того, что выпало — герб или решка соответственно. Пусть X — число выпавших гербов, Y — число перемен знака в образовавшейся последовательности плюсов и минусов. Нарисовать таблицу совместного распределения случайных величин X и Y, а также частное распределение случайной величины: а) X; б) Y.

2) В урне лежат 3 шара с номерами 1, 2, 3. Наудачу извлекаются два шара. Пусть X — номер первого шара, Y — номер второго шара. Вычислить $\mathrm{cov}(X, Y)$, если шары извлекаются:
а) с возвращением; б) без возвращения. (*Указание.* Используйте формулу (10).)

3) Точка (X, Y) выбирается наудачу в квадрате $[0, 1]^2$. Вычислить $\mathrm{cov}(X, Y)$.
(*Указание.* Используйте формулу (11) и определение (5).)

Домашнее задание

Если пятая буква вашей фамилии находится в диапазоне:
«**А – Е**», то «своими» являются задачи 5.1 и 5.5;
«**Ж – М**», то «своими» являются задачи 5.2 и 5.6;
«**Н – Р**», то «своими» являются задачи 5.3 и 5.7;
«**С – Я**», то «своими» являются задачи 5.4 и 5.8.

5.1) В условиях приведённого выше примера 1 вычислить $\mathrm{cov}(X, Y)$.

5.2) Бросаются две игральных кости (кубика). Здесь $\omega = (i,\ j)$, где $1 \le i \le 6$, $1 \le j \le 6$, $|\Omega| = 6^2 = 36$. Найти совместное распределение случайных величин $X = \min\{i,\ j\},\ Y = \max\{i,\ j\}$.

5.3) Точка $\omega = (x,\ y)$ взята наудачу в квадрате $[0, 1]^2$. Положим $X = \min\{x,\ y\},\ Y = \max\{x,\ y\}$. Найти:
а) функцию распределения и плотность случайного вектора (X, Y); б) $\mathbf{P}(X + Y \le 1)$.
(*Указание.* В пункте а) см. задачу 3.4 и используйте формулу (6). В пункте б) примените формулу (4).)

5.4) Пусть $\mathbf{P}(X=i, Y=j)=(1-\alpha)(1-\beta)\alpha^i\beta^j$; $i,\ j=0, 1, 2, \ldots$; $0<\alpha<1,\ 0<\beta<1$.
Найти без знака суммы: а) $\mathbf{P}(X=i)$, $\mathbf{P}(Y=j)$; б) $F_{X,Y}(k,l)$ для произвольных целых $k\geq 0$ и $l\geq 0$.

5.5) Пусть $Y=a+bX$, где a и b — константы, $\mathbf{D}X\neq 0$. Доказать, что $\rho(X,Y)=1$ при $b>0$.

5.6) В условиях задачи 5.4 положим $Z=\max\{X,Y\}$. Найти (без знака суммы) для произвольного целого числа $k\geq 0$: а) $\mathbf{P}(Z\leq k)$; б) $\mathbf{P}(Z=k)$.

5.7) Функция распределения случайного вектора (X,Y) на квадрате $[0, 1]^2$ задаётся формулой $F_{X,Y}(x,y)=\min\{x,y\}$. Найти $\mathbf{P}((X,Y)\in[0, 1]^2)$ и частные функции распределения $F_X(x)$, $F_Y(y)$.

5.8) В условиях задачи 5.7 найти $\mathbf{P}(X<Y)$, $\mathbf{P}(X>Y)$, $\mathbf{P}(X=Y)$. (*Указание.* Используйте ответ на вопрос 3: $\mathbf{P}((X,Y)\in(a,b]\times(c,d])=F_{X,Y}(b,d)-F_{X,Y}(b,c)-F_{X,Y}(a,d)+F_{X,Y}(a,c)$.)