4. СИСТЕМА ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Многомерной случайной величиной (векторной случайной величиной, случайным вектором или случайной точкой) называют упорядоченный набор нескольких случайных величин:

$$\mathbf{X} = (X_1, X_2, ..., X_n).$$

Функцией распределения системы (X, Y) двух случайных величин называется вероятность того, что ее составляющие X и Y одновременно окажутся меньше заданных значений x и y:

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

Последнее соотношение имеет простой геометрический смысл (рис. 4.1): функция распределения $F(x, y)$ равна вероятности попадания случайной точки (X, Y) в бесконечный квадрант с вершиной в точке (x, y) и расположенный левее и ниже этой точки.

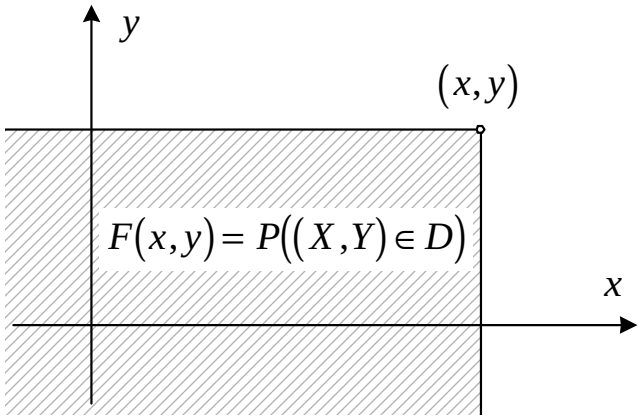

Рис. 4.1. Геометрический смысл функции распределения

Функция распределения системы случайных величин есть неубывающая функция каждого своего аргумента. Повсюду на минус бесконечности функция распределения равна нулю. Если все аргументы функции распределению обращаются в бесконечность, то функция распределения становится равной единице. Если один из аргументов обращается в бесконечность, то функция распределения системы переходит в функцию распределения случайной величины, соответствующей другому аргументу.

Вероятность попадания случайной точки в прямоугольник R, ограниченный прямыми $x=\alpha$, $x=\beta$, $y=\gamma$ и $y=\delta$ равна:

$$P((X, Y) \in R) = F(\alpha, \gamma) + F(\beta, \delta) - F(\alpha, \delta) - F(\beta, \gamma).$$

Рядом распределения системы дискретных случайных величин называют множество ее возможных значений и соответствующих этим значениям вероятностей. Если составляющие системы сами являются непрерывными случайными величинами, то распределение системы можно охарактеризовать плотностью вероятности:

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

График плотности вероятности системы двух случайных величин называют поверхностью распределения, а линии уровня этой поверхности называют линиями равной вероятности. По известной плотности функцию распределения системы двух случайных величин можно найти как интеграл:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv.$$

Условие нормировки плотности:

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

Вероятность попадания случайной точки в произвольную область:

$$P((X,Y)\in D)=\iint_{D} f(x,y)dxdy.$$

Плотности вероятностей составляющих системы:

$$f_1(x)=\int_{-\infty}^{\infty} f(x,y)dy,\ f_2(y)=\int_{-\infty}^{\infty} f(x,y)dx.$$

Для исчерпывающего описания системы необходимы как законы распределения ее составляющих, так и зависимость между составляющими. Эта зависимость может быть охарактеризована с помощью условных законов распределения.

Условным законом распределения случайной величины называют закон распределения, найденный в предположении, что на значения другой случайной величины наложено некоторое ограничение. Для непрерывно распределенных случайных величин по определению:

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

Поэтому условные плотности вероятностей равны:

$$f(x\mid y)=\frac{f(x,y)}{f_2(y)};\ f(y\mid x)=\frac{f(x,y)}{f_1(x)}.$$

Если закон распределения случайной величины не зависит от ограничений, накладываемых на другую случайную величину, то случайные величины называют независимыми. Для независимых случайных величин X и Y справедливо:

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

Поэтому плотность вероятности системы независимых случайных величин равна произведению плотностей вероятности ее составляющих:

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

Таким образом, найти плотность вероятности системы случайных величин по известным плотностям ее составляющих можно только в том случае, если входящие в систему случайные величины независимы. Для таких величин условие нормировки плотности дает

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x)f_2(y)dxdy=\left(\int_{-\infty}^{\infty} f_1(x)dx\right)\left(\int_{-\infty}^{\infty} f_2(y)dy\right)=1,$$

что возможно тогда и только тогда, когда единицам равны оба сомножителя.

Начальный момент порядка k, s системы двух дискретных случайных величин равен:

$$\alpha_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij},$$

где $p_{ij} = P\big((X,Y) = (x_i, y_j)\big)$, а суммирование выполняется по всем возможным значениями случайных величин. Для системы двух непрерывных случайных величин:

$$\alpha_{k,s} = \iint x^k y^s f(x,y)dxdy,$$

где интегрирование выполняется по всей плоскости *xOy*.

Центральные моменты порядка *k, s* равны

$$\mu_{k,s} = \sum_i \sum_j (x_i - \alpha_{1,0})^k (y_j - \alpha_{0,1})^s p_{ij},$$

$$\mu_{k,s} = \iint (x - \alpha_{1,0})^k (y - \alpha_{0,1})^s f(x,y)dxdy.$$

Часто порядком момента называют не пару чисел *k,s*, а их сумму *k*+*s*. В этом случае значения *k* и *s* обычно ясны из контекста. Выражения для моментов порядка *k*,0 и 0,*s* совпадают с выражениями для одномерной случайной величины. Как следствие, свойства таких моментов также аналогичны известным свойствам; в частности, $\mu_{1,0} = \mu_{0,1} = 0$.

Математические ожидания случайных величин *X* и *Y*, входящих в систему, равны начальным моментам порядка 1,0 и 0,1 соответственно:

$$M[X] = \alpha_{1,0},\ M[Y] = \alpha_{0,1}.$$

Дисперсии составляющих могут быть найдены как центральные моменты порядков 2,0 и 0,2:

$$D[X] = \mu_{2,0},\ D[Y] = \mu_{0,2}.$$

Помимо математических ожиданий и дисперсий составляющих, важной характеристикой системы оказывается корреляционный момент (момент связи, или ковариация) – второй смешанный центральный момент:

$$K[X,Y] = \mu_{1,1},$$

и связанный с ним безразмерный коэффициент корреляции:

$$\rho = \frac{K[X,Y]}{\sqrt{D[X]D[Y]}} = \frac{\mu_{1,1}}{\sqrt{\mu_{2,0}\mu_{0,2}}}.$$

Корреляционной зависимостью называют зависимость между значениями одних случайных величин и средними значениями других случайных величин. Коэффициент корреляции характеризует наличие линейной зависимости между случайными величинами. Коэффициент корреляции принимает значения из отрезка $[-1;1]$; случайные величины, для которых этот коэффициент равен нулю, называют некоррелированными. Из независимости случайных величин следует их некоррелированность. Например, для системы независимых непрерывных случайных величин $f(x,y) = f_1(x) f_2(y)$, поэтому

$$K[X,Y] = \mu_{1,1} = \iint \Big(x - \iint x f(x,y)dxdy\Big)\Big(y - \iint y f(x,y)dxdy\Big) f(x,y)dxdy =$$

$$= \iint \Big(x - \iint x f_1(x) f_2(y)dxdy\Big)\Big(y - \iint y f_1(x) f_2(y)dxdy\Big) f_1(x) f_2(y)dxdy =$$

$$= \iint \left(x - \left(\int_{-\infty}^{\infty} x f_1(x) dx \right) \left(\int_{-\infty}^{\infty} f_2(y) dy \right) \right) \left(y - \left(\int_{-\infty}^{\infty} y f_2(y) dy \right) \left(\int_{-\infty}^{\infty} f_1(x) dx \right) \right) f_1(x) f_2(y) dx dy =$$

$$= \left(\int_{-\infty}^{\infty} \left(x - \left(\int_{-\infty}^{\infty} x f_1(x) dx \right) \right) f_1(x) dx \right) \left(\int_{-\infty}^{\infty} \left(y - \left(\int_{-\infty}^{\infty} y f_2(y) dy \right) \right) f_2(y) dy \right) = \mu_{1,0} \mu_{0,1} = 0 .$$

Обратное утверждение о независимости некоррелированных случайных величин неверно: существуют зависимые, но некоррелированные случайные величины. Независимость случайных величин может быть установлена только на основе анализа всех смешанных моментов.

Пусть из генеральной совокупности извлечены две выборки $\{x_i\}$, $\{y_i\}$ одинакового объема N. Их можно считать выборкой $\{(x_i, y_i)\}$ значений системы двух случайных величин (X, Y). Соотношения для оценок математических ожиданий и дисперсий следуют из формул для моментов и повторяют соотношения для одномерной случайной величины:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \; ; \; \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \; ; \; s_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \; ; \; s_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 .$$

Для системы случайных величин естественно поставить вопрос о наличии и статистической значимости их взаимной зависимости. В рамках линейного корреляционного анализа эти вопросы решаются для линейной зависимости.

Для решения первого вопроса находят оценки ковариации и коэффициента корреляции:

$$K = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}), \; r = \frac{K}{\sqrt{s_x^2 s_y^2}} .$$

Полученные оценки позволяют сделать предварительное суждение о неизвестных значениях ковариации и коэффициента корреляции генеральной совокупности. Очевидно, что эти оценки могут оказаться (и, скорее всего, окажутся) ненулевыми даже в случае справедливости нулевой гипотезы $H_0 : K[X, Y] = 0$ о равенстве ковариации нулю. Решение вопроса о статистической значимости найденных оценок состоит в проверке указанной гипотезы. Для проверки находят эмпирическое значение статистики

$$w = \frac{1}{2} \ln \frac{1+r}{1-r} .$$

При нулевой гипотезе эта статистика подчинена нормальному закону с математическим ожиданием, равным нулю, и стандартным отклонением, равным

$$\sigma_w = \frac{1}{\sqrt{N-3}} .$$

Это позволяет найти вероятность критического события, состоящего в том, что в условиях эксперимента (для выборки объема N) при справедливой нулевой гипотезе $H_0 : K[X, Y] = 0$ значения K и r окажутся по абсолютной величине не меньше, чем в эксперименте:

$$P(A \mid H_0) = \sqrt{\frac{N-3}{2\pi}} \int_{|w|}^{\infty} e^{-\frac{(N-3)}{2}t^2} dt = 1 - 2\Phi\left(|w|\sqrt{N-3}\right),$$

где $\Phi(t)$ – функция Лапласа.

Если найденная вероятность оказывается меньше выбранного уровня значимости, то гипотеза о равенстве ковариации (и коэффициента корреляции) нулю отвергается в пользу двусторонней альтернативы, и найденная по выборке оценка коэффициента корреляции принимается как статистически значимая.

Ранее для случайных величин было указано, что коррелированность – необходимое, но не достаточное условие зависимости. Это утверждение сохраняет справедливость и для оценок. Для примера рассмотрим величины, связанные функциональной зависимостью $y = x^2 - 2$; очевидно, эти величины зависимы. Пусть в результате эксперимента получена выборка

$$\{(-2,2),(-1,-1),(0,-2),(1,-1),(2,2)\}.$$

Имеем: $\bar{x} = 0$, $\bar{y} = 0$; оценки ковариации и коэффициента корреляции:

$$K = (-4+1+0-1+4)/4 = 0,\ r = 0;$$

несмотря на наличие функциональной зависимости между составляющими выборку значениями (x, y) эти значения они оказались некоррелированными.

Следует заметить, что при $r \approx 0$ вероятность критического события будет близкой к единице вне зависимости от объема выборки (при $N > 3$) и найденную оценку нельзя будет принять как статистически значимую; тем не менее, линейный корреляционный анализ исчерпывающим образом не решает вопрос об исследовании зависимостей между случайными величинами. Подробнее этот вопрос рассматривается в рамках регрессионного анализа.