Лекция 10

ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СИСТЕМЫ ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН. n-МЕРНЫЙ СЛУЧАЙНЫЙ ВЕКТОР

ЦЕЛЬ ЛЕКЦИИ: определить числовые характеристики системы двух случайных величин: начальные и центральные моменты, ковариацию, коэффициент корреляции и регрессию; описать двумерное нормальное распределение; сформулировать закон распределения и найти числовые характеристики n*-мерного случайного вектора; вывести плотность распределения многомерного гауссова распределения.*

Начальные и центральные моменты

Обычно рассматривают в качестве числовых характеристик системы случайных величин начальные и центральные моменты различных порядков.

Начальным моментом порядка k,s системы двух случайных величин (X,Y) называется математическое ожидание произведения X^k на Y^s:

$$\alpha_{k,s} = M[X^k Y^s]. \tag{5.17}$$

Центральным моментом порядка k,s системы двух случайных величин (X,Y) называется математическое ожидание произведения $\dot{X}^k$ на $\dot{Y}^s$:

$$\mu_{k,s} = M[\dot{X}^k \dot{Y}^s], \tag{5.18}$$

где $\dot{X} = X - m_X$, $\dot{Y} = Y - m_Y$ – центрированные случайные величины.

Для системы дискретных случайных величин (X,Y)

$$\alpha_{k,s} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^k y_j^s p_{ij}\ ;\ \mu_{k,s} = \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i - m_X)^k (y_j - m_Y)^s p_{ij}\ ,$$

а для системы непрерывных случайных величин (X,Y)

$$\alpha_{k,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^s f(x,y)dxdy;$$

$$\mu_{k,s} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-m_X)^k (y-m_Y)^s f(x,y)dxdy.$$

Порядок моментов определяется суммой индексов $k+s$.

Начальные моменты первого порядка – это математические ожидания случайных величин X и Y:

$$\alpha_{1,0} = M[X^1Y^0] = M[X] = m_X; \; \alpha_{0,1} = M[X^0Y^1] = M[Y] = m_Y.$$

Отметим, что точка (m_X, m_Y) представляет собой характеристику положения случайной точки (X,Y), и разброс возможных значений системы случайных величин происходит вокруг этой точки.

Центральные моменты первого порядка равны нулю: $\mu_{1,0} = \mu_{0,1} = 0$.

Начальные моменты второго порядка:

$$\alpha_{2,0} = M[X^2Y^0] = M[X^2] = \alpha_2[X];$$

$$\alpha_{0,2} = M[X^0Y^2] = M[Y^2] = \alpha_2[Y];$$

$$\alpha_{1,1} = M[X^1Y^1] = R_{XY}. \qquad (5.19)$$

Начальный момент $\alpha_{1,1}$ называется смешанным начальным моментом второго порядка и обозначается как R_{XY}.

Центральные моменты второго порядка:

$$\mu_{2,0} = M[\dot{X}^2\dot{Y}^0] = M[\dot{X}^2] = D[X] = D_X;$$

$$\mu_{0,2} = M[\dot{X}^0\dot{Y}^2] = M[\dot{Y}^2] = D[Y] = D_Y;$$

$$\mu_{1,1} = M[\dot{X}^1\dot{Y}^1] = K_{XY}. \qquad (5.20)$$

Первые два центральных момента – это дисперсии случайных величин X и Y. Момент $\mu_{1,1}$ называется смешанным центральным моментом второго порядка или ковариацией (корреляционным моментом) и обычно обозначается как K_{XY}.

Ковариация

Для системы двух случайных величин (X, Y) ковариация выражается формулой

$$K_{XY} = M[\dot{X}\dot{Y}] = M[(X - m_X)(Y - m_Y)], \tag{5.21}$$

при этом $K_{XY} = K_{YX}$.

Дисперсию можно рассматривать как частный случай ковариации, т. е.

$$D_X = M[\dot{X}^2] = M[\dot{X}\dot{X}] = K_{XX}, \ D_Y = M[\dot{Y}\dot{Y}] = K_{YY}.$$

<u>Для независимых случайных величин ковариация всегда равна нулю</u>. Докажем это утверждение.

$$K_{XY} = \mu_{1,1} = M[\dot{X}\dot{Y}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_X)(y - m_Y) f(x, y)dxdy,$$

но для независимых случайных величин по теореме умножения плотностей имеем

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

Следовательно,

$$K_{XY} = \underbrace{\int_{-\infty}^{\infty}(x - m_X) f_1(x)dx}_{\mu_{1,0}=0} \underbrace{\int_{-\infty}^{\infty}(y - m_Y) f_2(y)dy}_{\mu_{0,1}=0} = 0.$$

Таким образом, доказано, что ковариация двух независимых случайных величин равна нулю.

Ковариация характеризует степень зависимости случайных величин и их рассеивание вокруг точки (m_X, m_Y). Ее можно выразить через начальные моменты:

$$K_{XY} = \alpha_{1,1} - \alpha_{1,0}\alpha_{0,1} = R_{XY} - m_X m_Y. \tag{5.22}$$

Ковариация двух случайных величин равна математическому ожиданию их произведения минус произведение математических ожиданий этих величин.

Размерность ковариации, также как и дисперсии, равна квадрату размерности случайной величины.

Степень зависимости случайных величин X и Y удобнее характеризовать посредством безразмерной величины – коэффициента корреляции

$$r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}, \tag{5.23}$$

который характеризует степень линейной зависимости, проявляющейся в том, что при возрастании одной из случайных величин другая также проявляет тенденцию возрастать или, наоборот, убывать.

Если $r_{XY} > 0$, то говорят, что случайные величины X и Y связаны положительной корреляцией; при $r_{XY} < 0$ – отрицательная корреляция между случайными величинами. Диапазон изменения r_{XY}

$$-1 \le r_{XY} \le 1. \tag{5.24}$$

Модуль коэффициента корреляции $|r_{XY}|$ характеризует степень "тесноты" линейной зависимости или уклонение корреляционной связи от линейной функциональной зависимости случайных величин X и Y.

При независимости случайных величин $r_{XY} = 0$, а при линейно функциональной зависимости $Y = aX + b$, $(a \ne 0)$:

$$r_{XY} = +1 \text{ при } a > 0; \; r_{XY} = -1 \text{ при } a < 0.$$

Если коэффициент корреляции равен нулю, то говорят, что случайные величины X и Y не коррелированы, но это не означает, что они независимы. При $r_{XY} = 0$ можно лишь утверждать, что между случайными величинами отсутствует линейная связь.

Регрессия

Условным математическим ожиданием одной из случайных величин, входящих в систему (X, Y), называется ее математическое ожидание, вычисленное при условии, что другая случайная величина приняла определенное значение, т. е. полученное на основе условного закона распределения.

Для дискретных случайных величин

$$M[Y/x_i] = m_{Y/x_i} = \sum_{j=1}^{m} y_j p_{y_j/x_i} \; ;$$

$$M[X/y_i] = m_{X/y_i} = \sum_{i=1}^{n} x_i p_{x_i/y_j} \; ,$$

где $p_{y_j/x_i} = P\{Y = y_j / X = x_i\}$; $p_{x_i/y_j} = P\{X = x_i / Y = y_j\}$ – условные вероятности случайных величин Y и X соответственно.

Для непрерывных случайных величин

$$M[Y/x\} = m_{Y/x} = \int_{-\infty}^{\infty} y f_2(y/x)dy \; ;$$

$$M[X/y\} = m_{X/y} = \int_{-\infty}^{\infty} x f_1(x/y)dx \; ,$$

где $f_2(y/x)$ и $f_1(x/y)$ – условные плотности распределения случайных величин: Y при $X = x$ и X при $Y = y$ соответственно.

Условное математическое ожидание случайной величины Y при заданном $X = x$: $M[Y/x] = m_{Y/x}$ называется регрессией Y на x; аналогично $M[X/y] = m_{X/y}$ – регрессией X на y. Графики этих зависимостей как функции x или y называют линиями регрессии, или "кривыми регрессии" Y на x и X на y соответственно (см. рис. 5.10).

Для независимых случайных величин линии регрессии Y на x и X на y параллельны осям абсцисс, так как математическое ожидание каждой из случайных величин не зависит от того, какое значение приняла другая случайная величина (см. рис. 5.10).

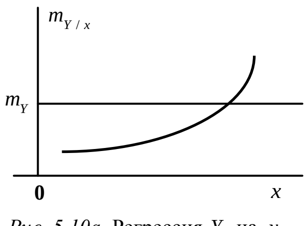

Рис. 5.10а. Регрессия Y на x

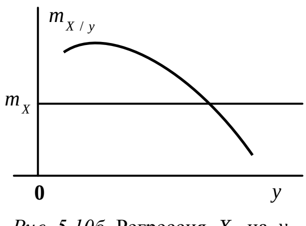

Рис. 5.10б. Регрессия X на y

Двумерное нормальное распределение

Система двух непрерывных случайных величин (X, Y) распределена по нормальному закону, если ее совместная плотность имеет вид

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-r_{XY}^2}} \exp\left\{-\frac{1}{2(1-r_{XY}^2)}\left[\frac{(x-m_X)^2}{\sigma_X^2} - \right.\right.$$

$$\left.\left. - \frac{2r_{XY}(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right]\right\}. \qquad (5.25)$$

Это двумерное нормальное распределение, или нормальный закон распределения на плоскости, который полностью определяется заданием его числовых характеристик: $m_X, m_Y, D_X, D_Y, r_{XY}$.

Условные законы распределения:

$$f_1(x/y) = \frac{f(x, y)}{f_2(y)} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-r_{XY}^2}} \exp\left[-\frac{1}{2(1-r_{XY}^2)}\left(\frac{x-m_X}{\sigma_X} - r_{XY}\frac{y-m_Y}{\sigma_Y}\right)^2\right];$$

$$f_2(y/x) = \frac{f(x, y)}{f_1(x)} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-r_{XY}^2}} \exp\left[-\frac{1}{2(1-r_{XY}^2)}\left(\frac{y-m_Y}{\sigma_Y} - r_{XY}\frac{x-m_X}{\sigma_X}\right)^2\right].$$

Каждый из этих условных законов распределения является нормальным с условным математическим ожиданием и условной дисперсией, определяемой по формулам:

$$m_{X/y} = m_X + r_{XY}\frac{\sigma_X}{\sigma_Y}(y-m_Y),\ D_{X/y} = \sigma_X^2(1-r_{XY}^2);$$

$$m_{Y/x} = m_Y + r_{XY}\frac{\sigma_Y}{\sigma_X}(x-m_X),\ D_{Y/x} = \sigma_Y^2(1-r_{XY}^2).$$

Отсюда следует, что для системы нормально распределенных случайных величин X и Y линии регрессии $m_{X/y}$ и $m_{Y/x}$ представляют собой прямые линии, т. е. регрессия для нормально распределенной системы (X, Y) всегда линейна. Для полного описания такой системы нужно знать пять параметров: координаты центра рассеивания (m_X, m_Y) и ковариационную матрицу, состоящую их четырех элементов:

$$\|K_{XY}\| = \begin{Vmatrix} D_X & K_{XY} \\ K_{YX} & D_Y \end{Vmatrix}, \text{ при этом } K_{XY} = K_{YX}.$$

При $r_{XY} = 0$ (случайные величины X и Y не коррелированы) совместная плотность распределения системы (X, Y) имеет вид

$$f(x, y) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{(x - m_X)^2}{2\sigma_X^2}\right] \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left[-\frac{(y - m_Y)^2}{2\sigma_Y^2}\right] =$$
$$= f_1(x) \cdot f_2(y),$$

т. е. если две нормальные случайные величины X и Y <u>не коррелированы</u>, то они и <u>независимы</u>.

Закон распределения и числовые характеристики n-мерного случайного вектора

Закон распределения системы n случайных величин – n-мерного случайного вектора $\vec{X} = (X_1, X_2, \dots, X_n)$ с составляющими $X_1, X_2, \dots, X_n$ – в общем случае может быть задан в виде функции распределения:

$$F(x_1, x_2, \dots, x_n) = P\{X_1 < x_1, X_2 < x_2, \dots, X_n < x_n\} =$$
$$= P\{\prod_{i=1}^{n} \{X_i < x_i\}\}. \tag{5.26}$$

Свойства функции распределения n-мерного случайного вектора аналогичны свойствам функции распределения одной или двух случайных величин:

1. $F(x_1, x_2, \dots, x_n)$ есть неубывающая функция каждого из своих аргументов.

2. Если хотя бы один из аргументов $(x_1, x_2, \ldots, x_n)$ обращается в $-\infty$, то функция распределения равна нулю.
3. Функция распределения любой подсистемы $(X_1, X_2, \ldots, X_k)$ системы $(X_1, X_2, \ldots, X_k, \ldots, X_n)$ определяется, если положить в функции распределения $F(x_1, x_2, \ldots, x_n)$ аргументы, соответствующие остальным случайным величинам, равными $+\infty$:

$$F_{1,2,\ldots,k}(x_1, x_2, \ldots, x_k) = F(x_1, x_2, \ldots, x_k, +\infty, \ldots, +\infty).$$

Чтобы определить функцию распределения $F_k(x_k)$ любой из случайных величин, входящих в систему, нужно положить в $F(x_1, x_2, \ldots, x_n)$ все аргументы, кроме x_k, равными $+\infty$:

$$F_k(x_k) = F(+\infty, +\infty, \ldots, x_k, +\infty, \ldots, +\infty).$$

4. Функция распределения $F(x_1, x_2, \ldots, x_n)$ непрерывна слева по каждому из своих аргументов.
5. Если случайные величины $X_1, X_2, \ldots, X_n$ независимы, то

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2)\cdots F_n(x_n).$$

Для системы непрерывных случайных величин функция распределения $F(x_1, x_2, \ldots, x_n)$ непрерывна и дифференцируема по каждому из аргументов, а также существует n-я смешанная частная производная

$$\frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x \cdot \partial x_2 \cdots \partial x_n},$$

которая является совместной плотностью распределения системы непрерывных случайных величин $(X_1, X_2, \ldots, X_n)$, т. е.

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x \cdot \partial x_2 \cdots \partial x_n}. \tag{5.27}$$

Свойства совместной функции распределения:

1. $f(x_1, x_2, \ldots, x_n) \geq 0$.

2. $\int\limits_{-\infty}^{\infty}(n)\int\limits_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)dx_1 dx_2 \cdots dx_n = 1$.

3. Если случайные величины $X_1, X_2, \ldots, X_n$ независимы, то

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n).$$

Закон распределения системы n зависимых случайных величин, являющийся функцией многих аргументов, весьма неудобен в практическом применении. Поэтому в практических (инженерных) приложениях теории вероятностей рассматриваются в основном числовые характеристики n-мерного случайного вектора:

1. n математических ожиданий:

$$m_1 = M[X_1];\ m_2 = M[X_2]; \ldots; m_n = M[X_n];$$

2. n дисперсий:

$$D_1 = D[X_1];\ D_2 = D[X_2]; \ldots; D_n = D[X_n];$$

3. $n(n-1)$ ковариаций:

$$K_{ij} = M[\dot{X}_i \dot{Y}_j] \text{ при } i \neq j.$$

Учитывая, что дисперсия случайной величины X_i есть ковариация K_{ii}, то все ковариации K_{ij} ($i \neq j$) совместно с дисперсиями $D_i = K_{ii}$ образуют матрицу ковариаций (ковариационную или корреляционную матрицу) – таблицу, состоящую из n строк и n столбцов:

$$\left\| K_{ij} \right\| = \left\| \begin{matrix} K_{11} & K_{12} & \ldots & K_{1n} \\ K_{21} & K_{22} & \ldots & K_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ K_{n1} & K_{n2} & \ldots & K_{nn} \end{matrix} \right\|. \quad (5.28)$$

Так как $K_{ij} = K_{ji}$, то матрица ковариаций симметрична относительно главной диагонали, на которой стоят дисперсии случайных величин. Если случайные величины попарно не коррелированные, т. е. $K_{ij} = 0$ при $i \neq j$, то матрица (5.25) становится диагональной:

$$\left\| K_{ij} \right\| = \left\| \begin{matrix} K_{11} & 0 & \dots & 0 \\ 0 & K_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & K_{nn} \end{matrix} \right\|.$$

Вместо матрицы ковариаций можно записать матрицу коэффициентов корреляции:

$$\left\| r_{ij} \right\| = \left\| \begin{matrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{matrix} \right\|, \tag{5.29}$$

где $r_{ij} = \dfrac{K_{ij}}{\sigma_i \sigma_j} = \dfrac{K_{ij}}{\sqrt{D_i D_j}} = \dfrac{K_{ij}}{\sqrt{K_{ii} K_{jj}}}$; $K_{ii} = D_{ii}$, $r_{ii} = \dfrac{D_i}{\sigma_i \sigma_i} = 1$.

Отсюда единицы по главной диагонали в матрице (5.29). Если же случайные величины попарно не коррелированные, т. е. $r_{ij} = 0$ при $i \neq j$, то матрица коэффициентов корреляции будет единичной матрицей:

$$\left\| r_{ij} \right\| = \left\| \begin{matrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{matrix} \right\|.$$

Иногда рассматривают условное математическое ожидание одной из случайных величин, например X_1, при условии, что все остальные случайные величины $X_2, X_3, \dots, X_n$ приняли определенные значения: $x_2, x_3, \dots, x_n$.

$$M[X_1 / x_2,\ldots, x_n] = m_{X_1 / x_2,\ldots,x_n} = \int\limits_{-\infty}^{\infty} x_1 \frac{f(x_1, x_2,\ldots, x_n)}{f_{2,\ldots,n}(x_2,\ldots, x_n)} dx_1 .$$

Это условное математическое ожидание называется регрессией X_1 на $x_2, x_3, \ldots, x_n$. Геометрически регрессия интерпретируется как поверхность в n-мерном пространстве $(x_1, x_2, \ldots, x_n)$ и называется поверхностью регрессии X_1 на $x_2, x_3, \ldots, x_n$.

Регрессия будет линейной, если поверхность регрессии описывается линейной функцией, т. е.

$$m_{X_1 / x_2,\ldots,x_n} = \gamma_{10} + \sum_{i=2}^{n} \gamma_{1i} x_i ,$$

где γ_{10}, γ_{1i} $(i = \overline{2, n})$ – постоянные коэффициенты.

В двумерном случае линия регрессии прямая, в трехмерном – плоскость; в общем случае – гиперплоскость в пространстве n измерений.

Для системы случайных величин $X_1, X_2, X_3, \ldots, X_n$, *имеющей нормальное распределение, регрессия всегда линейна*.

Многомерное нормальное распределение

Совместная плотность распределения вероятности системы произвольного числа n нормальных случайных величин – случайного вектора $\vec{X} = (X_1, X_2, \ldots, X_n)$ – имеет вид

$$f(x_1, x_2,\ldots, x_n) = \\ = \frac{1}{(2\pi)^{n/2}\sqrt{\Delta}} \exp[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{ij}^{(-1)} (x_i - m_{X_i})(x_j - m_{X_j})] , \quad (5.30)$$

где Δ – определитель ковариационной матрицы $\|K_{ij}\|$ системы случайных величин $(X_1, X_2, \ldots, X_n)$; $K_{ij}^{(-1)} = A_{ij} / \Delta$ – элементы обратной ковариационной матрицы, A_{ij} – алгебраическое дополнение элемента K_{ij} матрицы ковариаций.

Таким образом, параметрами n-мерного нормального распределения являются:

- вектор математических ожиданий $\vec{m} = (m_1, m_2, \ldots, m_n)$;
- ковариационная матрица $\|K_{ij}\|$ размером $n \times n$.

Если нормально распределенные случайные величины $X_1, X_2, \ldots, X_n$ не коррелированы, то корреляционная матрица становится диагональной

$$\|K_{ij}\| = \begin{Vmatrix} D_1 & 0 & \ldots & 0 \\ 0 & D_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & D_n \end{Vmatrix},$$

ее определитель $\Delta = D_1 \cdot D_2 \cdot \ldots \cdot D_n = \prod_{i=1}^{n} D_i$, а обратная корреляционная матрица будет иметь вид

$$\|K_{ij}^{(-1)}\| = \begin{Vmatrix} 1/D_1 & 0 & \ldots & 0 \\ 0 & 1/D_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1/D_n \end{Vmatrix}.$$

Таким образом, совместную плотность распределения можно привести к виду

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\Delta}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - m_{X_i}}{\sqrt{D_i}} \right)^2 \right] =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \left(\frac{x_i - m_i}{\sigma_i} \right)^2 \right] = \prod_{i=1}^{n} f_i(x_i).$$

Для нормально распределенной системы случайных величин из попарной некоррелированности отдельных величин, входящих в систему, следует их независимость.