Тема 11. Корреляция. Случайные векторы

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:

$$\mathrm{cov}(X,Y) = \mathrm{M}(X-\mathrm{M}X)(Y-\mathrm{M}Y) = \mathrm{M}XY - \mathrm{M}X\cdot\mathrm{M}Y. \qquad (1)$$

$$\rho(X,Y) = \mathrm{cov}(X,Y) \Big/ \sqrt{\mathrm{D}X\cdot\mathrm{D}Y}. \qquad (2)$$

$\mathrm{M}XY$ для дискретного случайного вектора (X,Y) вычисляется по формуле

$$\mathrm{M}XY = \sum_i\sum_j x_i y_j\,\mathbf{P}(X=x_i, Y=y_j) = \sum_i\sum_j x_i y_j\,p_{ij}. \qquad (3)$$

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

$$\mathrm{M}XY = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\,f_{X,Y}(x,y)\,dxdy. \qquad (4)$$

В силу известного из линейной алгебры *неравенства Коши — Буняковского*

$$|\mathrm{M}XY| \le \sqrt{\mathrm{M}X^2\cdot\mathrm{M}Y^2} \qquad (5)$$

для коэффициента корреляции выполняется неравенство

$$|\rho(X,Y)| \le 1. \qquad (6)$$

Для доказательства достаточно подставить в формулу (5) случайные величины $X-\mathrm{M}X$ и $Y-\mathrm{M}Y$.

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

Свойства ковариации

1) $\mathrm{cov}(a+bX,\ c+dY) = bd\,\mathrm{cov}(X,Y)$, где a, b, c, d — любые константы.

2) $\rho(a+bX,\ c+dY) = \rho(X,Y)$, если $b>0$ и $d>0$.

3) $\mathrm{D}(X_1+\ldots+X_n) = \sum_{i=1}^{n}\mathrm{D}X_i + 2\sum_{1\le i<j\le n}\mathrm{cov}(X_i,X_j)$.

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».

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

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

Пример 3. Пусть $\omega = (i_1, i_2, \dots, i_n)$ — случайная перестановка чисел от 1 до n, $|\Omega| = n!$. Рассмотрим $\boldsymbol{X}(\omega) = \omega$. Тогда $\mathbf{P}(X_1 = i_1, \dots, X_n = i_n) = 1/n!$.

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

$$\{\omega : X_1(\omega) \le x_1, \dots, X_n(\omega) \le x_n\} = \bigcap_{i=1}^{n} \{\omega : X_i(\omega) \le x_i\}. \tag{7}$$

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 (7), также является событием, и для него определена вероятность $\mathbf{P}(X_1 \le x_1, \dots, X_n \le x_n)$.

Определение. Функция n переменных

$$F_{\boldsymbol{X}}(x_1, \dots, x_n) = \mathbf{P}(X_1 \le x_1, \dots, X_n \le x_n) \tag{8}$$

называется *функцией распределения* n-мерного случайного вектора $\boldsymbol{X}$.

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

$$\mathbf{P}(\boldsymbol{X} \in \mathrm{A}) = \int_{\mathrm{A}} \dots \int f_{\boldsymbol{X}}(x_1, \dots, x_n)\, dx_1 \dots dx_n. \tag{9}$$

Пример 4. Обобщим пример 1 из темы 5, в котором было определено биномиальное распределение. Пусть в урне находятся l_1 занумерованных шаров 1-го цвета, l_2 занумерованных шаров 2-го цвета, … , l_k занумерованных шаров цвета k-го цвета, причём $l_1 + \dots + l_k = m$. 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, $j = 1, \dots, k$.

Найдём распределение дискретного случайного вектора $\boldsymbol{X} = (X_1, \dots, X_n)$ т. е. подсчитаем $\mathbf{P}(X_1 = i_1, \dots, X_k = i_k)$ для произвольных целых неотрицательных чисел $i_1, \dots, i_k$, где $i_1 + \dots + i_k = n$.

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. Итого — $C_n^{i_1} l_1^{i_1}$ 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.

Итого — $C_{n-i_1}^{i_2} l_2^{i_2}$ вариантов. Продолжая аналогично и перемножив количества вариантов для всех k цветов, получим:

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

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:

$$\mathbf{P}(X_1 = i_1,\ \ldots,\ X_k = i_k) = \frac{n!}{i_1!\, i_2! \ldots i_k!} p_1^{i_1} p_2^{i_2} \ldots p_k^{i_k}. \qquad (10)$$

Определение. Формула (10) задаёт *полиномиальное распределение* (*multinomial distribution*). Биномиальное распределение является его частным случаем при $k = 2$.

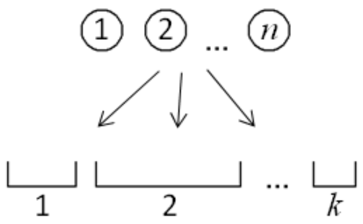

Рис. 1

В частности, если $p_1 = \ldots = p_k = 1/k$, то получаем распределение количеств попаданий в ящики при размещении наудачу n занумерованных шаров по k ящикам (см. рис. 1):

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

В заключение обобщим понятия математического ожидания и дисперсии на векторный случай. *Математическим ожиданием* случайного вектора $\boldsymbol{X}$ называется n-мерный числовой вектор $\mathrm{M}\boldsymbol{X} = (\mathrm{M}X_1, \ldots, \mathrm{M}X_n)$. 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 $\mathrm{D}X_i = \mathrm{cov}(X_i,\ X_i),\ i = 1, \ldots, n$.

Домашнее задание

Если **первая** буква вашего **имени** находится в диапазоне:

А, Б, В, Г, Д, Е, Ё, то «своими» являются задачи 11.1 и 11.5;
Ж, З, И, Й, К, Л, М, то «своими» являются задачи 11.2 и 11.6;
Н, О, П, Р, то «своими» являются задачи 11.3 и 11.7;
С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ы, Ь, Э, Ю, Я, то «своими» являются задачи 11.4 и 11.8.

11.1) В урне лежат 3 шара с номерами 1, 2, 3. Наудачу: а) без возвращения; б) с возвращением извлекают 2 шара. Пусть X — номер первого шара, Y — номер второго шара. Вычислить $\operatorname{cov}(X, Y)$.

11.2) Точка (X, Y) выбирается наудачу в квадрате $[0, 1]^2$. Вычислить: а) $\operatorname{cov}(X, Y)$; б) $\operatorname{cov}(X,\ X+Y)$.

11.3) Монету бросают: а) 3 раза; б) 2 раза, отмечая результат каждого бросания знаком + или – 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. Вычислить $\mathrm{M}XY$ и $\operatorname{cov}(X, Y)$.

11.4) Точка $\omega = (x,\ y)$ взята наудачу в квадрате $[0, 1]^2$. Положим $X(\omega) = \min\{x,\ y\}$, $Y(\omega) = \max\{x,\ y\}$. Вычислить: а) $\mathrm{M}XY$; б) $\operatorname{cov}(X, Y)$.

11.5) Из урны с l белыми и $m - l$ чёрными шарами наугад <u>без возвращения</u> извлекаются n шаров. Пусть I_k — индикатор того, что k-й шар окажется белым. Найти $\rho(I_i,\ I_k)$ при $i \neq k$.

11.6) Пусть $\omega = (i_1,\ \ldots,\ i_n)$ — случайная перестановка чисел от 1 до n, $X_k(\omega) = i_k$, $k = 1, \ldots, n$. Вычислить $\operatorname{cov}(X_j,\ X_k)$ при $j \neq k$. Записать ответ без знака суммы в наиболее простом виде.

11.7) По k ящикам наудачу раскладываются n занумерованных шаров. Пусть X_j обозначает число шаров в j-м ящике, $j = 1, \ldots, k$. Вычислить $\operatorname{cov}(X_j,\ X_l)$ при $j \neq l$. Максимально упростить ответ. (*Указание*. $X_j = I_{1j} + \ldots + I_{nj}$, где I_{ij} — индикатор того, что i-й шар ($i = 1, \ldots, n$) попал в j-й ящик.)

11.8) В урне находятся a белых, b черных и c серых шаров. Пусть X и Y — количества белых и черных шаров при n-кратном выборе с возвращением. Найти коэффициент корреляции между X и Y.

11.9)* Игральную кость бросают n раз. Обозначим через X_i число очков, выпавших при i-м бросании. Рассмотрим случайные величины $Y_i = X_i / (X_1 + \ldots + X_n)$, $i = 1, \ldots, n$. Найти $\rho(Y_i,\ Y_j)$ при $i \neq j$.