

Теория вероятностей и математическая статистика

Линейные комбинации случайных величин. Ковариация и корреляция.

Глеб Карпов

Программы БИ, УЦП, ТАДвБ ВШБ ВШЭ.

Линейная комбинация случайных величин

Математическое ожидание комбинации

- Часто важно анализировать поведение суммы двух или более случайных величин
- В основном мы хотим понять, как ведет себя переменная $T = aX \pm bY$, каковы ее характеристики.
- Линейное свойство математического ожидания: $E[aX \pm bY] = aE[X] \pm bE[Y]$.

$$\begin{aligned} E[T] &= \sum_{i=1}^n \sum_{j=1}^m (ax_i \pm by_j) P(X = x_i, Y = y_j) = \\ &= \sum_{i=1}^n \sum_{j=1}^m ax_i P(X = x_i, Y = y_j) \pm \sum_{i=1}^n \sum_{j=1}^m by_j P(X = x_i, Y = y_j) = \\ &= a \sum_{i=1}^n x_i \underbrace{\sum_{j=1}^m P(X = x_i, Y = y_j)}_{\text{marginal for } X} \pm b \sum_{j=1}^m y_j \underbrace{\sum_{i=1}^n P(X = x_i, Y = y_j)}_{\text{marginal for } Y} = \\ &= a \sum_{i=1}^n x_i P(X = x_i) \pm b \sum_{j=1}^m y_j P(Y = y_j) = aE[X] \pm bE[Y] \end{aligned}$$

Линейная комбинация случайных величин

Но дисперсия - это уже совсем другая история...

- Нас интересует $Var[T] = Var[aX \pm bY]$.
- Применим сокращенную формулу для дисперсии:

$$\begin{aligned} Var[T] &= E[T^2] - (E[T])^2 = E[(aX \pm bY)^2] - (aE[X] \pm bE[Y])^2 \\ &= E[a^2X^2 \pm 2abXY + b^2Y^2] - ((aE[X])^2 \pm 2abE[X]E[Y] + (bE[Y])^2) \\ &= a^2E[X^2] \pm 2abE[XY] + b^2E[Y^2] - a^2(E[X])^2 \mp 2abE[X]E[Y] - b^2(E[Y])^2 \\ &= a^2Var[X] + b^2Var[Y] \pm 2ab(E[XY] - E[X]E[Y]) \end{aligned}$$

- В процессе использовали свойство линейности математического ожидания.
- Слагаемое $E[XY] - E[X]E[Y]$ называется **ковариацией** X и Y .
- Определение: $Cov(X, Y) = E[XY] - E[X]E[Y]$.

Ковариация и независимость

Утверждение: Если X и Y - независимые случайные величины, то $E[XY] = E[X]E[Y]$

i Proof

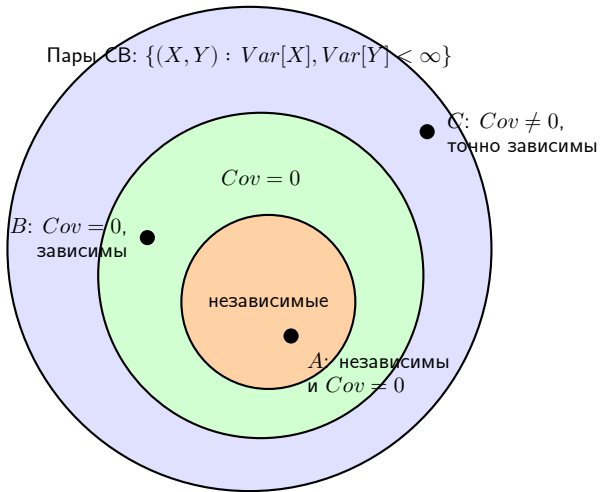
Используем определение независимости величин:

$$\forall x_i, \forall y_j : P_X(X = x_i) P_Y(Y = y_j) = P(\{X = x_i\} \cap \{Y = y_j\}) = P(X = x_i, Y = y_j)$$

$$\begin{aligned} E[XY] &= \sum_{i=1}^n \sum_{j=1}^m x_i y_j P(X = x_i, Y = y_j) = \sum_{i=1}^n \sum_{j=1}^m x_i y_j P_X(x_i) P_Y(y_j) \\ &= \sum_{i=1}^n x_i P_X(x_i) \sum_{j=1}^m y_j P_Y(y_j) = E[X]E[Y]. \end{aligned}$$

Будьте внимательны: обратное утверждение неверно!

Логика высказываний: независимость и ковариация



Случай C используем для детекции зависимости: $Cov \neq 0 \Rightarrow$ зависимы.

Корреляция

- Ковариация не является хорошей мерой зависимости, поскольку она зависит от масштабов X и Y .
- Идея: нормализовать ковариацию и избавиться от масштабов.
- Корреляция:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

- Она строго ограничена от -1 до 1 и:

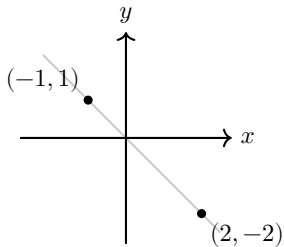
$$\text{Corr}(X, Y) = 1 \Leftrightarrow Y = kX + b, \quad k > 0$$

$$\text{Corr}(X, Y) = -1 \Leftrightarrow Y = -kX + b, \quad k > 0$$

- Важно: Если переменные независимы, то их ковариация (и корреляция) равна нулю. Обратное утверждение НЕВЕРНО. Ковариация может быть равной нулю, но при этом величины являются зависимыми.

Пример:

Случайный вектор (X, Y) принимает значения $(-1, 1)$ и $(2, -2)$ с вероятностями $\frac{2}{3}$ и $\frac{1}{3}$.



	$Y = 1$	$Y = -2$
$X = -1$	$\frac{2}{3}$	0
$X = 2$	0	$\frac{1}{3}$

1. Найдите маргинальные функции вероятности для X и Y .
2. Независимы ли X и Y ?
3. Найдите ковариацию (и корреляцию) между X и Y .

Решение задачи

1. Маргинальные функции вероятности

Для X :

$$P(X = -1) = \sum_j P(X = -1, Y = y_j) = \frac{2}{3} + 0 = \frac{2}{3}$$

$$P(X = 2) = \sum_j P(X = 2, Y = y_j) = 0 + \frac{1}{3} = \frac{1}{3}$$

Для Y :

$$P(Y = 1) = \sum_i P(X = x_i, Y = 1) = \frac{2}{3} + 0 = \frac{2}{3}$$

$$P(Y = -2) = \sum_i P(X = x_i, Y = -2) = 0 + \frac{1}{3} = \frac{1}{3}$$

Пример

Проверка независимости

Проверим условие независимости для пары:

$$P(X = -1, Y = 1) = \frac{2}{3} \quad \text{и} \quad P(X = -1)P(Y = 1) = \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9} \neq \frac{2}{3}$$

Ответ: X и Y не независимы.

Пример

Ковариация и корреляция

Сначала найдем математические ожидания:

$$E[X] = (-1) \cdot \frac{2}{3} + 2 \cdot \frac{1}{3} = 0$$

$$E[Y] = 1 \cdot \frac{2}{3} + (-2) \cdot \frac{1}{3} = 0$$

Теперь найдем $E[XY]$:

$$E[XY] = (-1) \cdot 1 \cdot \frac{2}{3} + 2 \cdot (-2) \cdot \frac{1}{3} = -\frac{2}{3} - \frac{4}{3} = -2$$

Пример

Ковариация:

$$Cov(X, Y) = E[XY] - E[X]E[Y] = -2 - 0 \cdot 0 = -2$$

Для корреляции найдем дисперсии:

$$E[X^2] = (-1)^2 \cdot \frac{2}{3} + 2^2 \cdot \frac{1}{3} = \frac{2}{3} + \frac{4}{3} = 2$$

$$Var[X] = E[X^2] - (E[X])^2 = 2 - 0 = 2$$

$$E[Y^2] = 1^2 \cdot \frac{2}{3} + (-2)^2 \cdot \frac{1}{3} = \frac{2}{3} + \frac{4}{3} = 2$$

$$Var[Y] = E[Y^2] - (E[Y])^2 = 2 - 0 = 2$$

Пример

Корреляция:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{-2}{\sqrt{2 \cdot 2}} = \frac{-2}{2} = -1$$

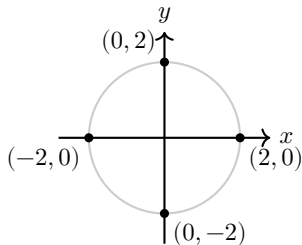
Корреляция равна -1 , что означает полную отрицательную линейную зависимость между X и Y (здесь $Y = -X$).

Пример зависимости при нулевой корреляции

Случайный вектор (X, Y) равномерно принимает одно из четырёх значений на окружности:

$$(2, 0), \quad (0, 2), \quad (-2, 0), \quad (0, -2),$$

каждое с вероятностью $\frac{1}{4}$.



	$Y = -2$	$Y = 0$	$Y = 2$
$X = -2$	0	$\frac{1}{4}$	0
$X = 0$	$\frac{1}{4}$	0	$\frac{1}{4}$
$X = 2$	0	$\frac{1}{4}$	0

Найдите коэффициент корреляции и проверьте случайные величины на зависимость.

Пример: решение

Маргинальные распределения:

x	-2	0	2
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

y	-2	0	2
$P(Y=y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$E[X] = (-2) \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 0, \quad E[Y] = (-2) \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 0.$$

Во всех исходах произведение $XY = 0$, поэтому $E[XY] = 0$ и

$$Cov(X, Y) = E[XY] - E[X]E[Y] = 0 \implies Corr(X, Y) = 0.$$

Зависимость — проверка через произведение маргинальных вероятностей, например для пары $(0, -2)$:

$$P(X=0, Y=-2) = \frac{1}{4}, \quad P(X=0)P(Y=-2) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \neq \frac{1}{4}.$$

Значит X и Y зависимы (хотя $Corr(X, Y) = 0$).