Ковариация и корреляция

Маргарян Ашот Араратович

04 июля 2025 г

1. Ковариация двух случайных величин

Пусть X и Y — две случайные величины. Тогда:

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \tag{1}$$

Раскрыв выражение, можно получить:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \tag{2}$$

И мы последнее слагаемое определяем как **ковариацию**:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$$

Упражнение: Из (1) получить (2).

2. Матрица ковариации для n случайных величин

Пусть есть случайный вектор $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$. Тогда матрица ковариации Σ имеет вид:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

Матрица симметрична: $\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_j,X_i)$.

3. Коэффициент линейной корреляции Пирсона

Коэффициент Пирсона между набором случайных величин X и Y определяется следующим образом:

$$r_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

где σ_X и σ_Y — стандартные отклонения.

Подходит для:

- Количественных данных
- Линейных зависимостей
- Нормально распределённых переменных (желательно)

Измеряет: степень линейной связи между переменными ($r \in [-1; 1]$)

Пример расчёта

Имеются данные:

$$\begin{array}{cccccc} X & 1 & 2 & 3 & 4 & 5 \\ Y & 2 & 4 & 5 & 4 & 5 \end{array}$$

Рассчитаем:

$$\bar{X} = 3, \quad \bar{Y} = 4$$

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{4}\cdot(4+0+0+0+2) = 1.5$$

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2} = \sqrt{2.5}, \quad \sigma_Y = \sqrt{1.5}$$

$$r_{XY} = \frac{1.5}{\sqrt{2.5\cdot 1.5}} \approx \frac{1.5}{1.936} \approx 0.775$$

Рекомендации к использованию

- Используется, если предполагается линейная связь
- Чувствителен к выбросам
- Не отражает нелинейную зависимость

4. Проверка значимости корреляции (t-критерий)

Гипотезы:

$$H_0 : \rho = 0 \quad \text{(нет линейной связи)}, \quad H_1 : \rho \neq 0$$

Статистика t рассчитывается по формуле:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Применим к примеру выше:

$r \approx 0.775, \quad n = 5$

$$t = \frac{0.775\cdot\sqrt{3}}{\sqrt{1-(0.775)^2}} = \frac{0.775\cdot 1.732}{\sqrt{1-0.6006}} = \frac{1.342}{\sqrt{0.3994}} \approx \frac{1.342}{0.632} \approx 2.123$$

Проверим по критическим значениям распределения Стьюдента при $df = 3$ и уровне значимости $\alpha = 0.05$:

$$t_{\text{кр}} \approx 3.182$$

Так как $t_{\text{набл}} = 2.123 < t_{\text{кр}} = 3.182$, статистически значимая связь **не обнаружена** на уровне 5%.

Вывод: хотя корреляция высока, из-за малого размера выборки нельзя утверждать о её статистической значимости.

5. Коэффициент ранговой корреляции Спирмена

Пусть X и Y - набор случайных величин, а $R(X)$ и $R(Y)$ - соответствующие ранги элементов X и Y. Тогда коэффициент критерия Спирмена r_s рассчитывается по формуле:

$$r_s = \frac{\mathrm{Cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

где $d_i = R(x_i) - R(y_i)$ — разность рангов, присвоенных значениям x_i и y_i.

Легко заметить, что коэффициент корреляции Спирмена - это линейный коэффициент критерия Пирсона для рангов элемнтов.

Подходит для:

- Порядковых и количественных данных
- Малых выборок
- Нелинейных, но монотонных зависимостей

Измеряет: степень монотонной зависимости между переменными. $r_s \in [-1; 1]$
Интерпретация:

- $r_s = 1$: строго возрастающая зависимость
- $r_s = -1$: строго убывающая зависимость
- $r_s = 0$: отсутствие монотонной зависимости

Пример расчёта

Пусть:

X	10	20	30	40	50
Y	3	2	4	1	5

Ранги:

$$R(X) = [1, 2, 3, 4, 5], \quad R(Y) = [3, 2, 4, 1, 5]$$

$$d_i = R(X) - R(Y) = [-2, 0, -1, 3, 0] \Rightarrow \sum d_i^2 = 14$$

$$r_s = 1 - \frac{6 \cdot 14}{5(5^2-1)} = 1 - \frac{84}{120} = 0.3$$

Рекомендации к использованию

- Используется при нарушении нормальности
- Подходит для устойчивой оценки зависимости при наличии выбросов

Проверка значимости с помощью t-критерия

Формула аналогична:

$$t = \frac{r_s \cdot \sqrt{n-2}}{\sqrt{1 - r_s^2}}$$

$$t = \frac{0.3 \cdot \sqrt{3}}{\sqrt{1 - 0.09}} \approx \frac{0.5196}{\sqrt{0.91}} \approx \frac{0.5196}{0.9539} \approx 0.545$$

Критическое значение $t_{\text{кр}}(3, 0.05) \approx 3.182$
Вывод: Связь не является статистически значимой на уровне 5%.

6. Коэффициент корреляции Мэтьюса (MCC)

Определяется как:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Где:

- TP — true positives (истинно положительные)
- TN — true negatives (истинно отрицательные)
- FP — false positives (ложно положительные)
- FN — false negatives (ложно отрицательные)

Подходит для: бинарных классификаторов (0/1), особенно при несбалансированных выборках.
Измеряет: симметричную корреляцию между предсказанием и фактом. $MCC \in [-1; 1]$
Интерпретация:

- +1: идеальное соответствие
- 0: случайное соответствие
- −1: полное несоответствие

Пример расчета MCC

	Факт 1	**Факт 0**
Прогноз 1	TP = 40	FP = 10
Прогноз 0	FN = 5	TN = 45

$$MCC = \frac{40 \cdot 45 - 10 \cdot 5}{\sqrt{(50)(45)(55)(50)}} = \frac{1800 - 50}{\sqrt{6187500}} \approx \frac{1750}{2487.47} \approx 0.703$$

Рекомендации к использованию

- Лучше F1 при несбалансированных классах
- Симметричен: не зависит от выбора положительного класса

7. Проверка значимости МСС через χ^2-критерий

Формула:

$$\chi^2 = n \cdot (\text{МСС})^2$$

Для предыдущего примера: $n = 100,\ \text{МСС} = 0.703$

$$\chi^2 = 100 \cdot 0.703^2 = 100 \cdot 0.4942 = 49.42$$

Сравним с критическим значением $\chi^2(1)$ при $\alpha = 0.05$: $\chi^2_{\text{кр}} \approx 3.84$
Вывод: Наблюдаемая корреляция значима на уровне 0.05

8. Шкала Чеддока (для интерпретации $|r|$)

$\|r\|$	Интерпретация
0.00–0.10	Отсутствие связи
0.10–0.30	Слабая связь
0.30–0.50	Умеренная связь
0.50–0.70	Заметная связь
0.70–0.90	Высокая связь
0.90–1.00	Очень высокая связь

Примечание: Шкала Чеддока — эмпирическая и может отличаться по контексту задачи и предметной области.

9. Таблица сопряжённости

Таблица сопряжённости (или таблица сопряжённого распределения, *contingency table*) — это двухмерная таблица, в которой отражено совместное распределение двух категориальных признаков. Строки соответствуют значениям одного признака, столбцы — значениям другого, а ячейки содержат количество наблюдений, попавших в соответствующую комбинацию категорий.

Пример таблицы сопряжённости для признаков A (4 категории) и B (5 категорий):

	B_1	B_2	B_3	B_4	B_5
A_1	5	10	15	7	3
A_2	12	6	8	9	5
A_3	9	4	11	2	3
A_4	6	7	13	5	2

Такая таблица используется как основа для вычисления статистик зависимости между категориальными признаками: критерия хи-квадрат, коэффициента Крамера и др.

10. Коэффициент Крамера (Cramér's V)

Определение и формула

Коэффициент Крамера V измеряет степень связи между двумя категориальными переменными и основан на значении критерия согласия χ^2.

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

где:

- χ^2 — статистика хи-квадрат, рассчитанная по таблице сопряжённости (уровень значимости α обычно берем за 0.05, а количество степеней свободы

$$df = (r-1) \cdot (c-1)$$

;
- n — общее число наблюдений;
- $k = \min(r, c)$ — минимальное из количества строк r и столбцов c.

Типы данных и интерпретация

Коэффициент Крамера подходит для:

- номинальных и порядковых признаков,
- анализа таблиц $r \times c$ произвольной размерности.

Интерпретация значения $V \in [0, 1]$:

- $V = 0$ — переменные статистически независимы,
- $V \to 1$ — сильная ассоциация между переменными.

Пример расчёта

Рассмотрим таблицу сопряжённости, где $r = 4$, $c = 5$:

	B_1	B_2	B_3	B_4	B_5
A_1	5	10	15	7	3
A_2	12	6	8	9	5
A_3	9	4	11	2	3
A_4	6	7	13	5	2

Общее количество наблюдений:

$$n = 5 + 10 + \cdots + 2 = 162$$

Пусть по таблице рассчитано значение статистики $\chi^2 = 18.45$. Тогда:

$$V = \sqrt{\frac{18.45}{162 \cdot (4-1)}} = \sqrt{\frac{18.45}{486}} \approx \sqrt{0.03796} \approx 0.195$$

Интерпретация результата

Коэффициент $V \approx 0.195$ указывает на слабую зависимость между переменными A и B.

Рекомендации по использованию

- Используйте Cramér's V при анализе зависимости между двумя категориальными переменными, особенно при таблицах больше, чем 2×2.
- Значение V следует интерпретировать вместе с χ^2-проверкой на значимость (особенно при малых выборках).
- Не используйте Cramér's V для количественных переменных — он предназначен только для категориальных данных.

Итог: что мы изучили

В ходе этой главы мы:

- познакомились с понятием **ковариации** как меры совместного изменения переменных;
- исследовали различные виды **корреляции** — Пирсона, Спирмена, Кендалла, Мэтьтьюса, Крамера;
- научились различать виды данных (количественные, порядковые, категориальные) и выбирать соответствующий коэффициент;
- рассмотрели способы проверки статистической значимости корреляции (t-критерий, χ^2);
- ввели **таблицу сопряженности** как основу анализа категориальных признаков.

Ковариация сама по себе может быть трудна для интерпретации, так как зависит от масштаба. Коэффициенты корреляции нормируют её в диапазон от -1 до 1, позволяя оценивать не только силу, но и направление связи.

Для практических задач аналитики важно не только вычислить корреляцию, но и корректно её интерпретировать в контексте признаков, шкал измерения и поставленной задачи.