

Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

Кафедра прикладной математики и информатики

Математическая статистика

Отчет по лабораторной работе №7

Выполнил студент гр. 5030102/20202

Тишковец С.Е.

Преподаватель

Баженов А.Н.

Санкт-Петербург

2025

Оглавление

1. Постановка задачи	3
2. Теоретическая информация	3
2.1. Квартиль и интервальные оценки	3
2.2. Индекс Жаккара.....	4
2.3. Метод решения	4
3. Результаты исследования	5
3.1. Графики	5
3.2. Оценка оптимальных параметров сдвига	5
4. Выводы	6

1. Постановка задачи

Сгенерировать 2 выборки X_1 и X_2 мощностью $n = 1000$.

Средние и ширины выборок должны отличаться, например:

$$X_1 = N(0, 0.95), X_2 = N(1, 1.05)$$

где $N(m, \sigma)$ — нормальное распределение.

Для выборок X_1 и X_2 найти внутренние и внешние оценки:

$$\text{Inn } X_i = [Q_{1/4}, Q_{3/4}]$$

$$\text{Out } X_i = [\min X_i, \max X_i]$$

Здесь $Q_{1/4}, Q_{3/4}$ — первый и третий квартили

Определить параметр сдвига a : $X_1 + a = X_2$

2. Теоретическая информация

2.1. Квартиль и интервальные оценки

Квартиль — это значение, разделяющее упорядоченные данные на четыре равные части.

- Первый квартиль ($Q_{1/4}$) — значение, ниже которого находится 25% данных.
- Третий квартиль ($Q_{3/4}$) — значение, ниже которого находится 75% данных.

Внутренняя оценка выборки ($\text{Inn } X_i$) определяется как интервал между первым и третьим квартилем:

$$\text{Inn } X_i = [Q_{1/4}, Q_{3/4}]$$

Этот интервал отражает «основную массу» данных и устойчив к выбросам.

Внешняя оценка выборки ($\text{Out } X_i$) определяется через минимальное и максимальное значения выборки:

$$\text{Out } X_i = [\min X_i, \max X_i]$$

что охватывает всю вариацию данных, включая возможные выбросы.

2.2. Индекс Жаккара

Индекс Жаккара широко используется для оценки степени схожести двух множеств. В случае работы с интервалами он определяется как отношение длины пересечения интервалов к длине их объединения:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Пересечение двух интервалов $[a_1, a_2]$ и $[b_1, b_2]$ вычисляется по формулам:

- левая граница пересечения = $\max(a_1, b_1)$
- правая граница пересечения = $\min(a_2, b_2)$

Если левая граница пересечения больше или равна правой, пересечение считается пустым.

Объединение интервалов определяется так:

- левая граница объединения = $\min(a_1, b_1)$
- правая граница объединения = $\max(a_2, b_2)$

Индекс Жаккара принимает значение от 0 (полное отсутствие пересечения) до 1 (полное совпадение интервалов). Использование индекса Жаккара позволяет количественно оценить степень перекрытия интервалов между выборками при различных значениях сдвига.

2.3. Метод решения

Варьировать параметр сдвига a и вычислять 2 меры совместности

$$J_{Inn} = \frac{Inn X_1 \wedge Inn X_2}{Inn X_1 \vee Inn X_2}$$

$$J_{Out} = \frac{Out X_1 \wedge Out X_2}{Out X_1 \vee Out X_2}$$

Здесь J - индекс Жаккара

\wedge, \vee — минимум и максимум по включению.

Поскольку выборки X_1 и X_2 имеют разные средние значения, предполагается существование параметра a , такого что:

$$X_1 + a \approx X_2$$

В реальных условиях a не известен заранее. Чтобы найти его, мы варьируем a в некотором диапазоне значений и для каждого a рассчитываем индексы $J_{Inn}(a)$ и $J_{Out}(a)$, которые отражают степень совпадения соответствующих интервалов. Наилучшее значение a выбирается как то, при котором индекс Жаккара достигает максимума:

$$a_{Inn} = \arg \max J_{Inn}(a)$$

$$a_{Out} = \arg \max J_{Out}(a)$$

Таким образом, задача сводится к оптимизации функции схожести между интервалами двух выборок относительно параметра сдвига a .

3. Результаты исследования

3.1. Графики

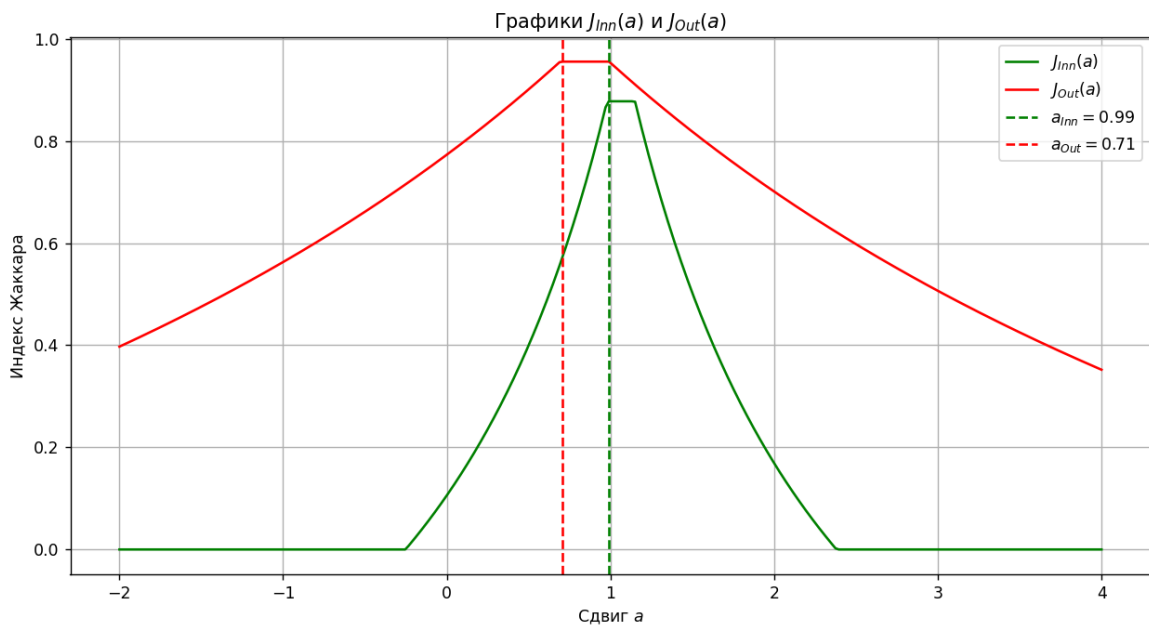


Рис. 1. Графики $J_{Inn}(a)$ и $J_{Out}(a)$

3.2. Оценка оптимальных параметров сдвига

$$a_{Inn} = \arg \max J_{Inn}(a) \approx 0.98$$

$$a_{Out} = \arg \max J_{Out}(a) \approx 0.72$$

Это свидетельствует о том, что выборки X_1 и X_2 действительно связаны с систематическим сдвигом около 1, что соответствует исходным условиям генерации данных.

4. Выводы

В работе показано, что индекс Жаккара позволяет эффективно количественно оценивать степень перекрытия интервалов между выборками при варьировании сдвига. Максимальные значения индексов соответствуют сдвигам, близким к реальной разнице средних значений.

Оптимальный сдвиг по внутренним оценкам a_{Inn} оказался более точным по сравнению с внешними a_{Out} , что связано с меньшей чувствительностью внутренних оценок к выбросам. Внешние оценки, охватывая всю выборку, подвержены влиянию экстремальных значений.

Таким образом, метод на основе внутренних интервалов более надёжен для оценки сдвига между распределениями. Индекс Жаккара показал высокую эффективность и может быть рекомендован для практического применения в задачах анализа данных.