Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Отчёт по лабораторной работе №7

по дисциплине
«Математическая статистика»

Выполнила студентка
группы 5030101/20202 Чинь Тхи Тху Хоай

Проверил
Преподаватель Баженов Александр Николаевич

Санкт-Петербург
2025

Содержание

1 Постановка задачи

Сгенерировать 2 выборки X_1 и X_2 мощностью $n = 1000$. Средние и ширины выборок должны отличаться, например:

$$X_1 = N(0,\ 0.95), \quad X_2 = N(1,\ 1.05) \tag{1}$$

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

Для выборок X_1 и X_2 найти внутренние и внешние оценки.

$$Inn\ X_i = [Q_{1/4},\ Q_{3/4}], \tag{2}$$

$$Out\ X_i = [\min X_i,\ \max X_i]. \tag{3}$$

Здесь $Q_{1/4}$, $Q_{3/4}$ — первый и третий квартили. Определить параметр сдвига a

$$X_1 + a = X_2$$

2 Теория

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 ($Inn\ X$) определяется как интервал между первым и третьим квартилем:

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

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

Внешняя оценка выборки ($Out\ X$) определяется через минимальное и максимальное значения выборки:

$$Out\ X = [\min(X), \max(X)]$$

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:

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

где:

- $|A \cap B|$ — длина пересечения интервалов A и B,
- $|A \cup B|$ — длина объединения интервалов A и B.

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

$$\text{левая граница пересечения} = \max(a_1, b_1),$$

$$\text{правая граница пересечения} = \min(a_2, b_2).$$

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

Объединение интервалов определяется так:

$$\text{левая граница объединения} = \min(a_1, b_1),$$

$$\text{правая граница объединения} = \max(a_2, b_2).$$

Таким образом:

$$|A \cap B| = \max(0, \min(a_2, b_2) - \max(a_1, b_1)),$$

$$|A \cup B| = \max(a_2, b_2) - \min(a_1, b_1).$$

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

$$J_{Inn} = \frac{Inn\ X_1 \wedge Inn\ X_2}{Inn\ X_1 \vee InnX_2}, \tag{4}$$

$$J_{Out} = \frac{Out\ X_1 \wedge Out\ X_2}{Out\ X_1 \vee OutX_2}, \tag{5}$$

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:

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

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

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

3 Программная реализация

Лабораторная работа выполнена на языке Python 3.12.6 в среде разработки Visual Studio Code. Использовались дополнительные библиотеки:

1. matplotlib
2. math
3. numpy

В приложении находится ссылка на GitHub репозиторий с исходным кодом.

4 Результаты

4.1 Построение графиков $J_{Inn}(a)$ и $J_{Out}(a)$

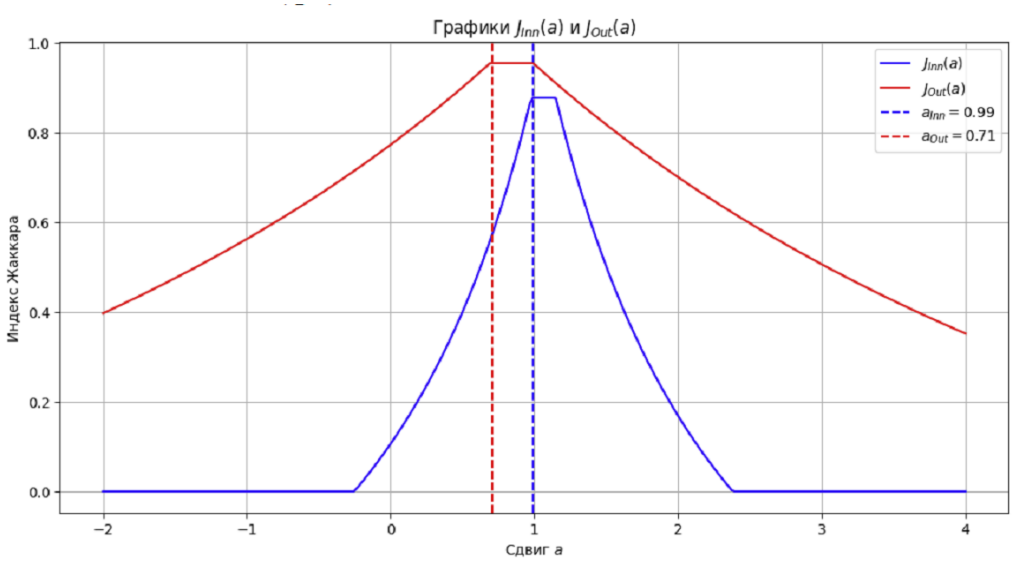

Рис. 1: Графики $J_{Inn}(a)$ и $J_{Out}(a)$ в зависимости от сдвига a.

4.2 Оценка оптимальных параметров сдвига

Из анализа графиков были определены значения сдвига, максимизирующие индексы Жаккара:

- Оптимальный сдвиг по внутренней оценке:

$$a_{Inn} = \arg\max_a J_{Inn}(a) \approx 0.99$$

- Оптимальный сдвиг по внешней оценке:

$$a_{Out} = \arg\max_a J_{Out}(a) \approx 0.71$$

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

5 Обсуждение

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.

6 Приложение

Код программы GitHub URL:

https://github.com/Akira1707/Math-Statistic/tree/main/Lab7