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

Сгенерировать 2 выборки X_1 и X_2 мощностью $\mathrm{n} = 1000$.

Средние и ширины выборок должны отличаться, например:

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

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

Для выборок X_1 и X_2 найти внутренние и внешние оценки:

$$\mathrm{Inn}\, X_i = \left[Q_{1/4}, Q_{3/4}\right]$$

$$\mathrm{Out}\, X_i = [minX_i, maxX_i]$$

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

Определить параметр сдвига a: $X_1 + \mathrm{a} = X_2$

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 ($\mathrm{Inn}\, X_i$) определяется как интервал между первым и третьим квартилем:

$$\mathrm{Inn}\, X_i = \left[Q_{1/4}, Q_{3/4}\right]$$

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

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

$$\mathrm{Out}\, X_i = [minX_i, maxX_i]$$

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:

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

Пересечение двух интервалов $[\mathrm{a}_1,\ \mathrm{a}_2]$ 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

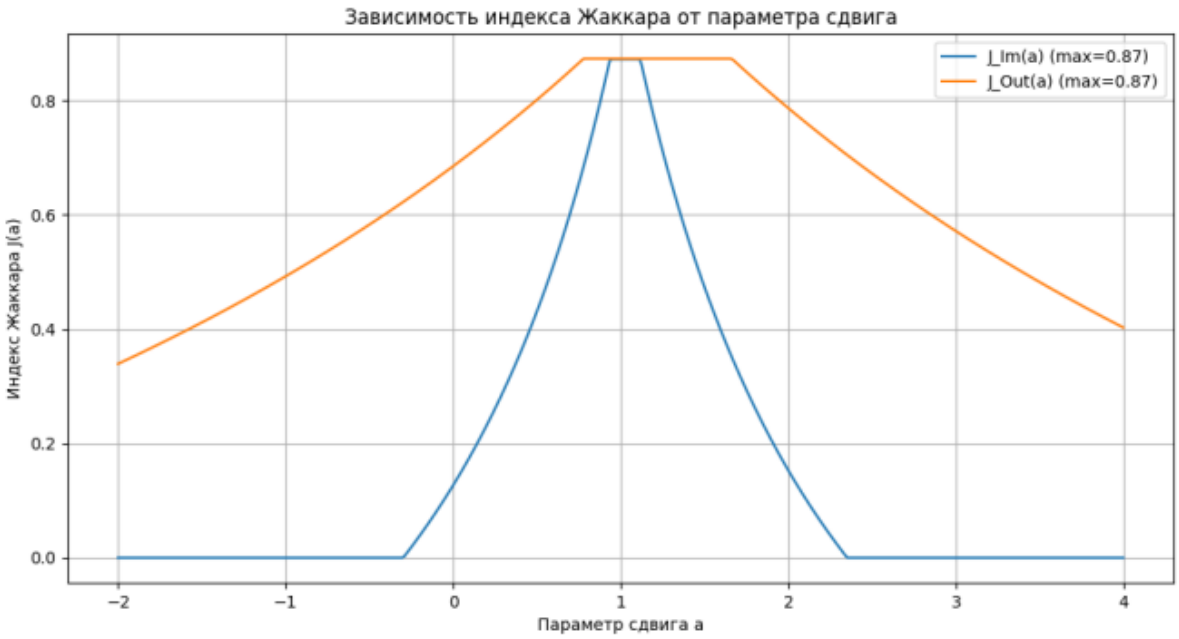

Рис. 1. Графики $J_{Inn}(a)$ и $J_{Out}(a)$

3.2. Анализ графика

График $J(a)$ отражает зависимость индекса Жаккара от параметра сдвига a для внутренних ($J_{Inn}(a)$)и внешних ($J_{Out}(a)$)оценок.

• Для внутренних оценок ($J_{Inn}(a)$) график имеет колоколообразную форму с максимумом, близким к теоретическому значению a = 1.

• Для внешних оценок ($J_{Out}(a)$) график более широкий, с максимумом, смещённым вправо относительно a = 1.

• Пик $J_{Inn}(a)$более узкий, чем $J_{Out}(a)$, что указывает на более высокую чувствительность внутренних оценок к изменению a.

3.3. Внутренняя оценка

Для внутренних оценок, основанных на интерквартильных интервалах, были получены следующие результаты:

- Максимальное значение индекса Жаккара: $J_{Inn} = 0.8721$.
- Оптимальные значения параметра сдвига a: $a_{Inn} \in [0.9459,1.1142]$.
- Среднее значение параметра сдвига: $\overline{a_{Inn}} = 1.0301$

3.4. Внешняя оценка

Для внешних оценок, основанных на минимаксных интервалах, результаты оказались следующими:

- Максимальное значение индекса Жаккара: $J_{\mathrm{Out}} = 0.8731$.
- Оптимальные значения параметра сдвига a: $a_{\mathrm{Out}} \in [0.7859,1.6542]$.
- Среднее значение параметра сдвига: $\overline{a_{\mathrm{Out}}} = 1.2106$

3.5. Анализ результатов

Теоретически, $a = 1$, так как выборки были сгенерированы с математическими ожиданиями, равными 0 и 1.

1. Внутренние оценки дают значение a_{Inn}, близкое к теоретическому ($\overline{a_{Inn}} = 1.0301$), с узким интервалом оптимальных значений ([0.9459, 1.1142]), что указывает на высокую точность и стабильность метода.

2. Внешние оценки менее точны ($\overline{a_{\mathrm{Out}}} = 1.2106$), с более широким интервалом оптимальных значений ([0.7859,1.6542]), что может быть связано с чувствительностью к выбросам или асимметрией выборок.

3. Индексы Жаккара для внутренних ($J_{Inn} = 0.8721$) и внешних ($J_{\mathrm{Out}} = 0.8731$) оценок близки, но более широкий интервал для a_{Out} свидетельствует о меньшей устойчивости внешних оценок.

4. Выводы

Результаты показывают, что внутренние оценки, основанные на интерквартильных интервалах, обеспечивают более точную и устойчивую оценку параметра сдвига a, близкую к теоретическому значению $a = 1$.

Внешние оценки, использующие минимаксные интервалы, демонстрируют меньшую точность и большую чувствительность к выбросам, что проявляется в более широком интервале оптимальных значений a и смещении среднего значения ($\overline{a_{\mathrm{Out}}} = 1.2106$). Это подтверждает, что для больших выборок ($n = 1000$) внутренние оценки предпочтительнее для задач определения параметра сдвига.