Санкт-Петербургский Политехнический университет Петра Великого

Отчет по лабораторной работе №7

Определение систематического сдвига в данных

Студент: Швачко Никита Андреевич
Преподаватель: Баженов Александр Николаевич
Группа: 5030102/20202

Санкт-Петербург 2025

1 Постановка задачи

Цель лабораторной работы — определить систематический сдвиг между двумя выборками с помощью индекса Жаккара для твинов (двойных интервалов). Для этого были сгенерированы две выборки X_1 и X_2, обладающие различными средними и стандартными отклонениями:

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

Задача заключалась в том, чтобы варьировать параметр сдвига a таким образом, чтобы найти оптимальное значение, при котором индексы Жаккара для твинов достигают максимума. Для каждой выборки были рассчитаны два типа твинов:

- Внутренний твин: $\left[Q_{1/4}, Q_{3/4}\right]$ — первый и третий квартили
- Внешний твин: $[\min X_i, \max X_i]$ — минимальное и максимальное значение выборки

Твин представляет собой двойной интервал вида $X = [a, b] = [[a, a], [b, b]]$, где a и b — концы интервала. В результате работы были найдены оценки сдвига a, при которых индексы Жаккара $J_{\text{Inn}}(a)$ и $J_{\text{Out}}(a)$ достигают максимума.

2 Описание используемых методов

Для решения задачи был использован следующий подход:

1. **Генерация выборок**: С использованием библиотеки NumPy были сгенерированы две выборки X_1 и X_2 по нормальному распределению с заданными параметрами.
2. **Расчет твинов**: Для каждой выборки были рассчитаны два типа твинов:
   - Внутренний твин (25% и 75% квантили)
   - Внешний твин (минимум и максимум)

   Твин представляет собой двойной интервал, где каждый конец является интервалом вида $[a, a]$.
3. **Индекс Жаккара**: Для определения степени схожести выборок использовался индекс Жаккара, который вычисляется как:

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

   где $\wedge$ и $\vee$ — операции минимума и максимума по включению соответственно.
4. **Определение сдвига**: Параметр сдвига a изменялся в диапазоне от -2 до 4, и для каждого значения сдвига рассчитывались индексы Жаккара для обоих типов твинов. Искались значения a, при которых эти индексы достигают максимума:

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

5. **Визуализация**: Для представления результатов были построены графики зависимостей индексов Жаккара от значения сдвига, а также гистограммы распределений выборок с отмеченными твинами и их центрами.

3 Результаты эксперимента

В результате эксперимента были получены следующие данные:

- **Статистические характеристики выборок**:
  - Выборка X_1: $\mu \approx 0.018$, $\sigma \approx 0.930$
  - Выборка X_2: $\mu \approx 1.074$, $\sigma \approx 1.047$
  - Теоретическая разница средних: 1.000
- **Твины для выборки X2**:
  - Внутренний твин: $[[0.363, 0.363], [1.765, 1.765]]$
  - Длина внутреннего твина: 1.402
  - Центр внутреннего твина: 1.064
  - Внешний твин: $[[-2.087, -2.087], [4.353, 4.353]]$
  - Длина внешнего твина: 6.440
  - Центр внешнего твина: 1.133
- **Оценки сдвига**:
  - Оценка сдвига для внутреннего твина: $a_{\text{Inn}} \approx 0.990$
  - Оценка сдвига для внешнего твина: $a_{\text{Out}} \approx 0.709$
  - Отклонение от теоретического значения для внутреннего твина: 0.010
  - Отклонение от теоретического значения для внешнего твина: 0.291
- **Твины при оптимальном сдвиге**:
  - Внутренние твины:
    - X1 (со сдвигом 0.990): $[[0.375, 0.375], [1.606, 1.606]]$
    - X2: $[[0.363, 0.363], [1.765, 1.765]]$
    - Индекс Жаккара: 0.878
  - Внешние твины:
    - X1 (со сдвигом 0.709): $[[-2.370, -2.370], [4.369, 4.369]]$
    - X2: $[[-2.087, -2.087], [4.353, 4.353]]$
    - Индекс Жаккара: 0.956
- **Графики**: На графиках отображены зависимости индексов Жаккара от значения сдвига a и распределения выборок с отмеченными твинами и их центрами. Максимальные значения индексов Жаккара соответствуют значениям сдвига a_{Inn} и a_{Out} соответственно.

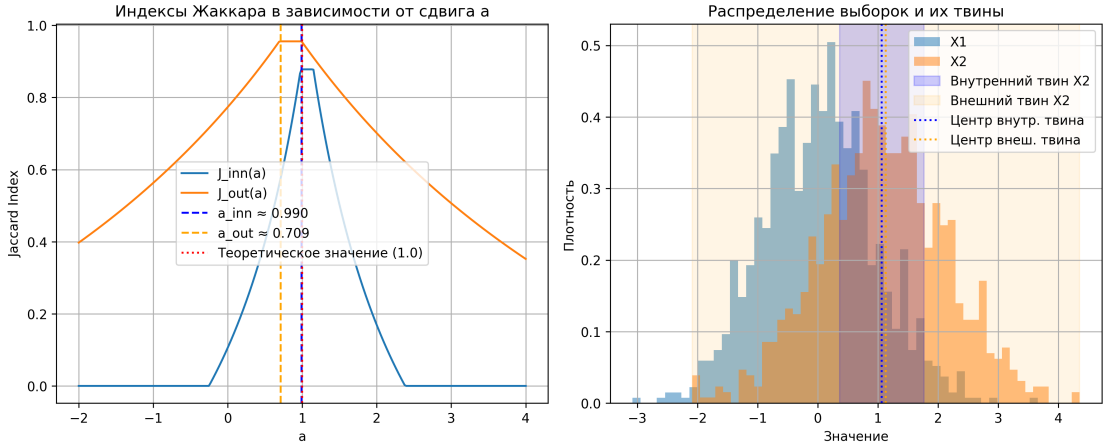

Рис. 1: Зависимость индексов Жаккара от параметра сдвига a и распределение выборок с твинами

4 Выводы

В ходе лабораторной работы были исследованы изменения индексов Жаккара при варьировании сдвига между выборками, что позволило точно оценить параметры сдвига a. Полученные результаты показывают, что:

- Метод внутреннего твина дал точную оценку сдвига ($a_{\text{Inn}} \approx 0.990$), которая практически совпадает с теоретическим значением (отклонение всего 0.010).
- Метод внешнего твина оказался менее точным ($a_{\text{Out}} \approx 0.709$), что может быть связано с большей чувствительностью к выбросам.
- Использование твинов позволяет получить более полное представление о структуре данных, так как каждый твин представляет собой двойной интервал, учитывающий как внутреннюю, так и внешнюю оценку.
- Центры твинов могут служить дополнительными характеристиками для анализа распределения данных.
- Индекс Жаккара для внешних твинов (0.956) оказался выше, чем для внутренних (0.878), что говорит о большей степени перекрытия внешних интервалов при оптимальном сдвиге.