

Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Кафедра «Прикладная математика»

**Отчёт по лабораторной работе №1
по дисциплине «Анализ данных с интервальной
неопределённостью»**

Выполнил студент:
Куксенко Кирилл Сергеевич
группа: 5040102/20201

Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2023 г.

Содержание

1	Постановка задачи	2
2	Теория	2
2.1	Индекс Жаккара	2
2.2	Нахождение оптимального значения R	2
3	Реализация	3
4	Результаты	3
5	Обсуждение	12

Список иллюстраций

1	Интервальная выборка X_1	3
2	Интервальная выборка X_2	4
3	Частота пересечений подинтервалов с интервалами выборки X_1	4
4	Частота пересечений подинтервалов с интервалами выборки X_2	5
5	Зависимость индекса Жаккара от значения R	6
6	Объединённая выборка $X_1 \cup R_{opt}X_2$	6
7	Частота пересечений подинтервалов с интервалами выборки $X_1 \cup R_{opt}X_2$	7
8	Зависимость частоты пересечения моды с интервалами $X_1 \cup RX_2$	8
9	Внутренняя и внешняя оценки R	9
10	Интервальная выборка X'_1	10
11	Интервальная выборка X'_2	10
12	Зависимость индекса Жаккара от значения R	11
13	Объединённая выборка $X'_1 \cup R'_{opt}X'_2$	11

1 Постановка задачи

Имеется две вещественные выборки $\overline{X_1}, \overline{X_2}$. Необходимо построить из них две интервальные выборки X_1, X_2 и найти такой вещественный коэффициент R , что выборка $X_1 \cup R X_2$ будет наиболее совместной в смысле индекса Жаккара.

2 Теория

2.1 Индекс Жаккара

Индекс Жаккара определяет степень совместности двух интервалов x, y .

$$JK(x, y) = \frac{wid(x \wedge y)}{wid(x \vee y)} \quad (1)$$

Здесь \wedge, \vee представляют собой операции взятия минимума и максимума по включению в полной арифметике Каухера. Формула 1 легко может быть обобщена на случай интервальной выборки $X = \{x_i\}_{i=1}^n$.

$$JK(X) = \frac{wid(\wedge_{i=1,n} x_i)}{wid(\vee_{i=1,n} x_i)} \quad (2)$$

Видно, что $JK(X) \in [-1, 1]$. Для удобства перенормируем значение $JK(X)$ так, чтобы оно было в интервале $[0, 1]$.

$$JK(X) = \frac{1}{2} + \frac{1}{2} JK(X) \quad (3)$$

2.2 Нахождение оптимального значения R

Для нахождения оптимального R необходимо сначала найти верхнюю и нижнюю границы $\underline{R}, \overline{R}$.

$$\underline{R} = \frac{\min_{i=1,n} \underline{x_{1i}}}{\max_{i=1,n} \underline{x_{2i}}} \quad (4)$$

$$\overline{R} = \frac{\max_{i=1,n} \overline{x_{1i}}}{\min_{i=1,n} \overline{x_{2i}}} \quad (5)$$

Затем оптимальное значение R может быть найдено методом половинного деления.

3 Реализация

Весь код написан на языке Python (версии 3.7.3). [Ссылка на GitHub с исходным кодом](#).

4 Результаты

Данные были взяты из файлов *data/dataset1/+0_5V/+0_5V_0.txt* и *data/dataset/-0_5V/-0_5V_42.txt*. Обынтерваливание было произведено следующим образом.

$$\mathbf{x}_i = [(x_i - \delta_i) - \varepsilon, (x_i - \delta_i) + \varepsilon], \varepsilon = \frac{100}{2^{14}} \quad (6)$$

где x_i - точечное значение, δ_i - точечная погрешность. Набор δ_i получен из соответствующих файлов в *data/dataset1/ZeroLine.txt*

Сначала посмотрим на исходные интервальные выборки X_1, X_2 .

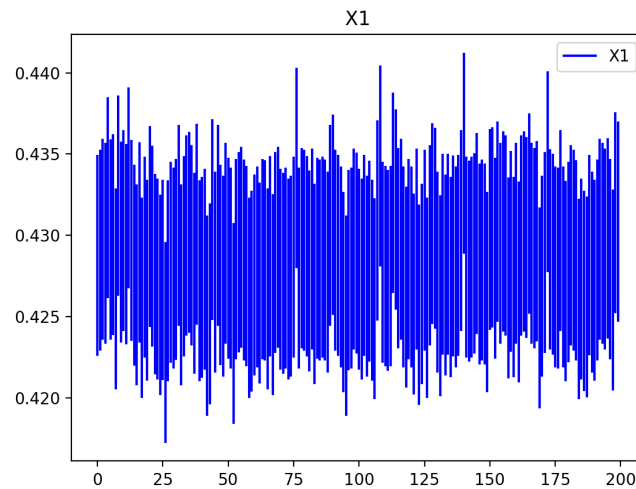


Рис. 1: Интервальная выборка X_1

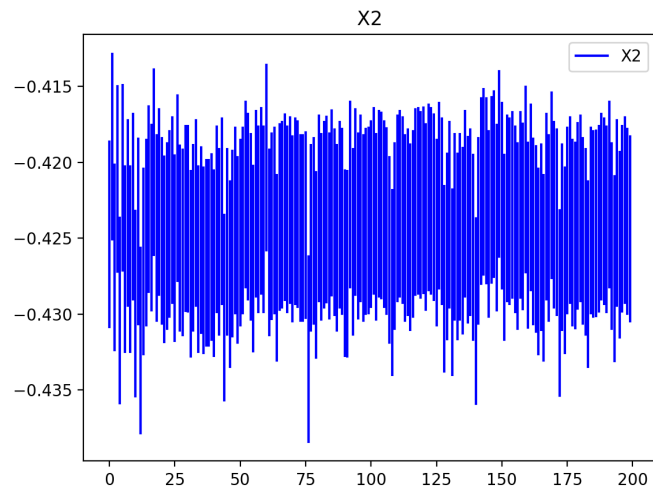


Рис. 2: Интервальная выборка X_2

Также построим график частоты пересечений подинтервалов для построения моды с исходными интервалами выборок. Сначала для X_1 .

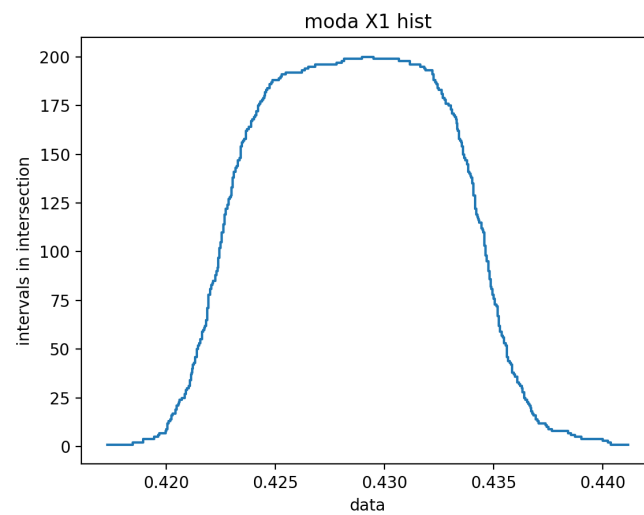


Рис. 3: Частота пересечений подинтервалов с интервалами выборки X_1

Затем для X_2 .

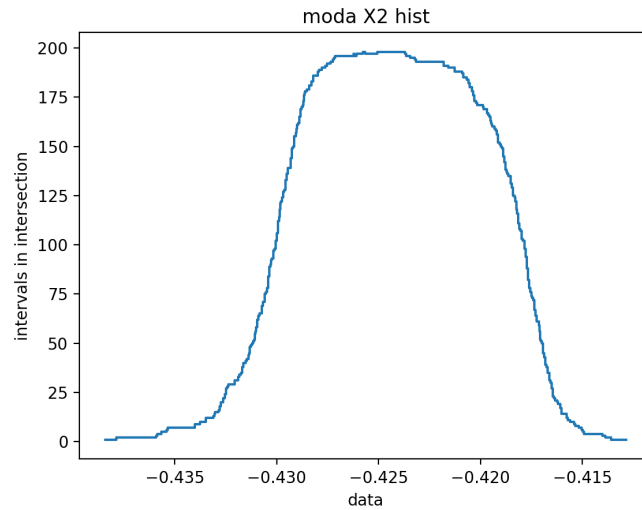


Рис. 4: Частота пересечений подинтервалов с интервалами выборки X_2

Мода для выборки X_1 равна интервалу $\mu_{X_1} = [0.4290, 0.4295]$, для выборки X_2 мода равно интервалу $\mu_{X_2} = [-0.4258, -0.4257]$.

Посчитаем индекс Жаккара обеих выборок. $JK(X_1) = 0.511$, $JK(X_2) = 0.477$. Найдем оптимальное значение R (для наглядности на графике 5 изображён более широкий интервал значений R).

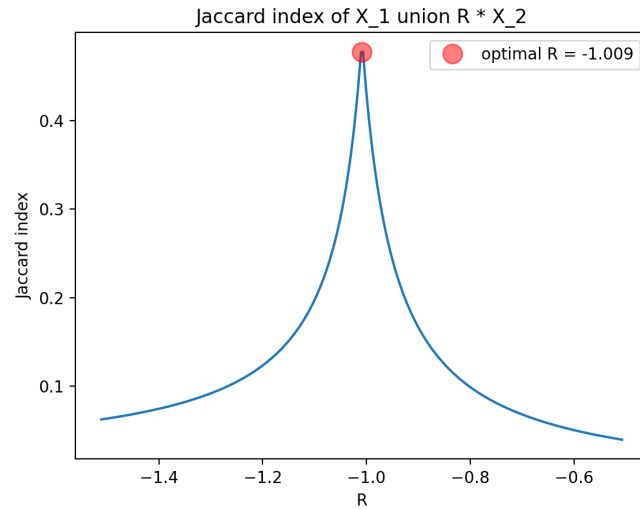


Рис. 5: Зависимость индекса Жаккара от значения R

Оптимальное значение R оказалось равно $R_{opt} = -1.009$. Построим объединённую выборку $X = X_1 \cup R_{opt}X_2$.

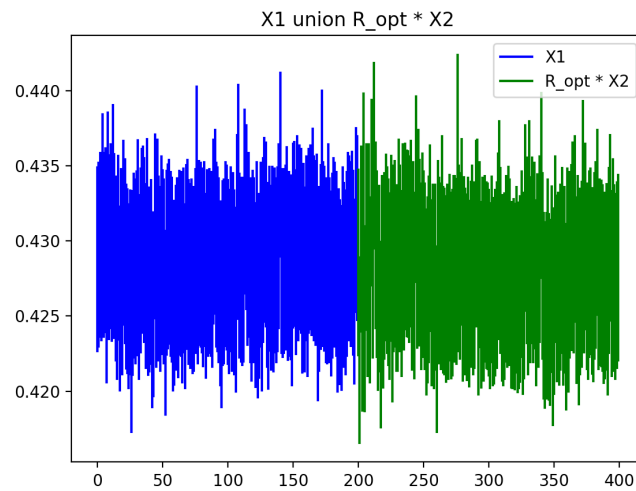


Рис. 6: Объединённая выборка $X_1 \cup R_{opt}X_2$

Индекс Жаккара полученной выборки равен $JK(X) = 0.477$.

Построим график частоты пересечений подинтервалов с объединённой выборкой $X_1 \cup R_{opt}X_2$.

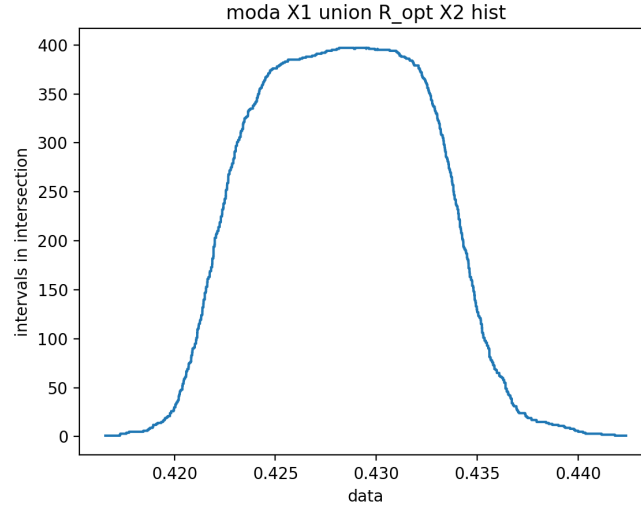


Рис. 7: Частота пересечений подинтервалов с интервалами выборки $X_1 \cup R_{opt}X_2$

Мода для объединённой выборки $X_1 \cup R_{opt}X_2$ равна интервалу $\mu_{X_1 \cup R_{opt}X_2} = [0.4281, 0.4289]$.

Посмотрим на зависимость частоты пересечений моды $\mu(R)$ с интервалами для объединённой выборки $X_1 \cup RX_2$ в зависимости от значений R .

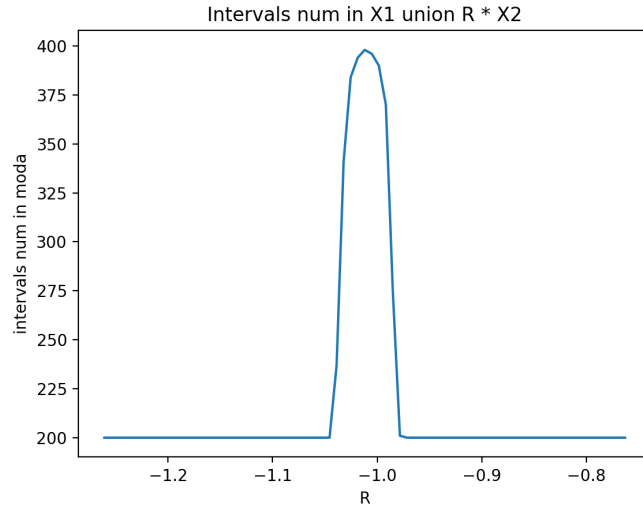


Рис. 8: Зависимость частоты пересечения моды с интервалами $X_1 \cup RX_2$

Найдём внутреннюю и внешнюю интервальные оценки $\mathbf{R} = [R_{in}, R_{out}]$. Для этого введём уровень доверия $\alpha = 0.95$ и найдём крайние значения R , удовлетворяющие $JK(R) > JK(R_{opt}) * \alpha$ для внутренней оценки и $\mu(R) > \mu(R_{opt}) * \alpha$ для внешней. Результаты представлены на рис. 9 (график $\mu(R)$ нормирован так, чтобы $\max_R \mu(R)$ и $\max_R JK(R)$ были равны).

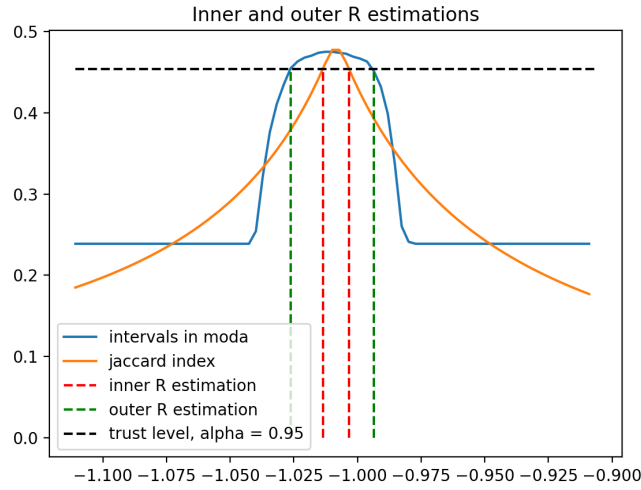


Рис. 9: Внутренняя и внешняя оценки R

В итоге получили следующие оценки: $R_{in} = [-1.0134, -1.0034]$, $R_{out} = [-1.0262, -0.9935]$. Тогда $\mathbf{R} = [[-1.0134, -1.0034], [-1.0262, -0.9935]]$.

Сравним полученные результаты с теми, что будут без учёта δ_i . $X'_k = \{[x_i - \varepsilon, x_i + \varepsilon]\}_{i=1}^n, k = 1, 2$.

X'_1 имеют вид.

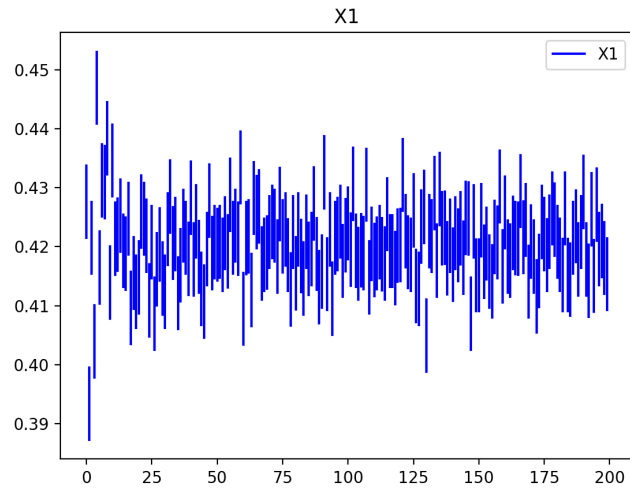


Рис. 10: Интервальная выборка X'_1

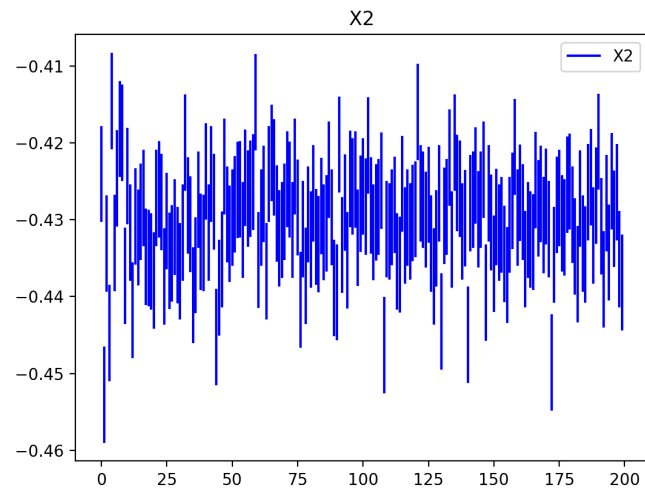


Рис. 11: Интервальная выборка X'_2

Вычислим индекс Жаккара $JK(X'_1) = 0.1855$, $JK(X'_2) = 0.2421$.
Зависимость индекса Жаккара от значения параметра R имеет вид.

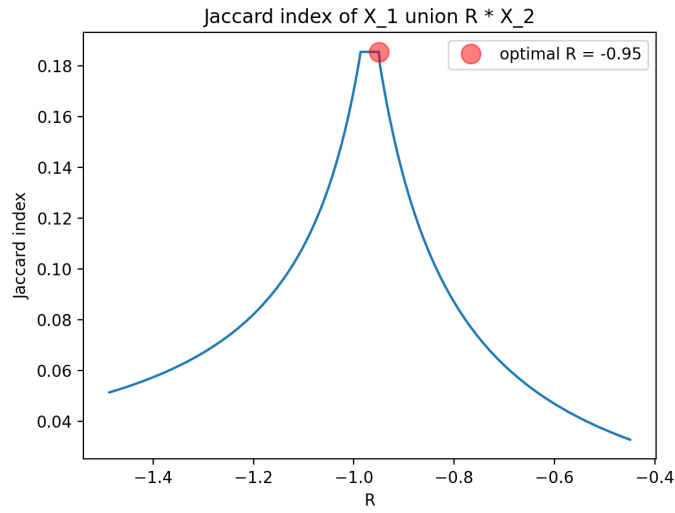


Рис. 12: Зависимость индекса Жаккара от значения R

Видно, что оптимальное значение параметра R равно $R'_{opt} = 0.9496$, что значительно отличается от первого случая. Тогда объединённая выборка $X'_1 \cup R'_{opt}X'_2$ имеет вид.

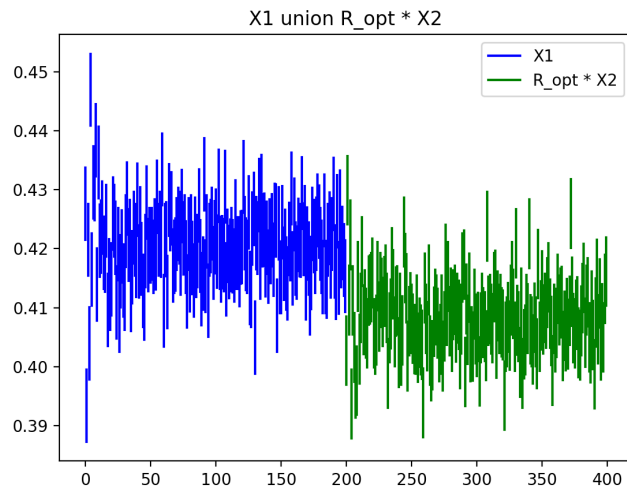


Рис. 13: Объединённая выборка $X'_1 \cup R'_{opt}X'_2$

5 Обсуждение

Из полученных результатов можно заметить следующее. Как видно на рисунке 5 график значений индекса Жаккара в зависимости от параметра R имеет один локальный минимум. Также видно, что индекс Жаккара объединённой выборки $X = X_1 \cup RX_2$ для любого значения R не превосходит значения индексов Жаккара для каждой выборки X_1, X_2 по отдельности, что вполне ожидаемо. Несмотря на это $JK(X)$ не сильно отличается от значений $JK(X_1), JK(X_2)$, скорее всего это связано с тем, что интервалы из X_1 и RX_2 имеют примерно одинаковую длину, что видно на рисунке 6.

На рисунках 5, 8, 9 видно, что индекс Жаккара имеет более "острый" график в окрестности максимума нежели график максимального пересечения моды. Как следствие, для одинакового достаточно большого уровня доверия $\alpha \approx 1.0$ индекс Жаккара даёт более точную оценку оптимального значения R .