Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический иститут

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

Список иллюстраций

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 $X_1 \cup RX_2$ будет наиболее совместной в смысле индекса Жаккара.

2 Теория

2.1 Индекс Жаккара

Индекс Жаккара определяет степень совместности двух интервалов x, y.

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

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 $X = \{x_i\}_{i=1}^n$.

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

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

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

2.2 Нахождение оптимального значения R

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

$$\underline{R} = \frac{\min_{i=1,n} \underline{x_{1i}}}{\max_{i=1,n} \overline{x_{2i}}} \tag{4}$$

$$\overline{R} = \frac{\max_{i=1,n} \overline{x_{1i}}}{\min_{i=1,n} \underline{x_{2i}}} \tag{5}$$

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_85.txt* и *data/dataset/-0_5V/-0_5V_6.txt*. Обынтерваливание было произведено следующим образом.

$$x = [x_0 - \varepsilon, x_0 + \varepsilon], \varepsilon \in U(0.01, 0.05) \quad (6)$$

где x_0 - точечное значение, $U(0.01, 0.05)$ - равномерное распределение.

Сначала посмотрим на исходные интервальные выборки X_1, X_2.

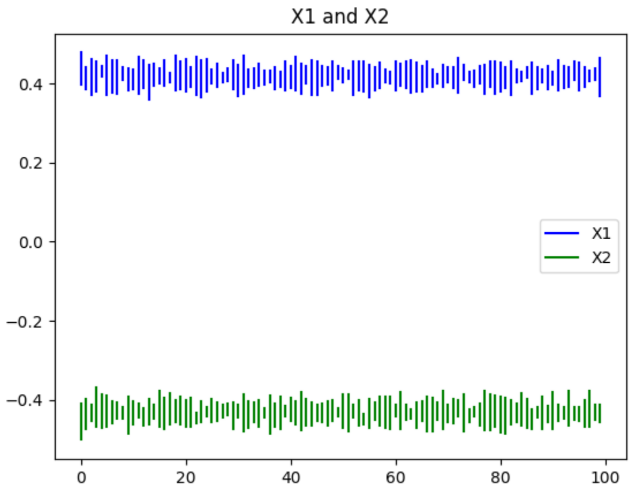

Рис. 1: Исходные интервальные выборки

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

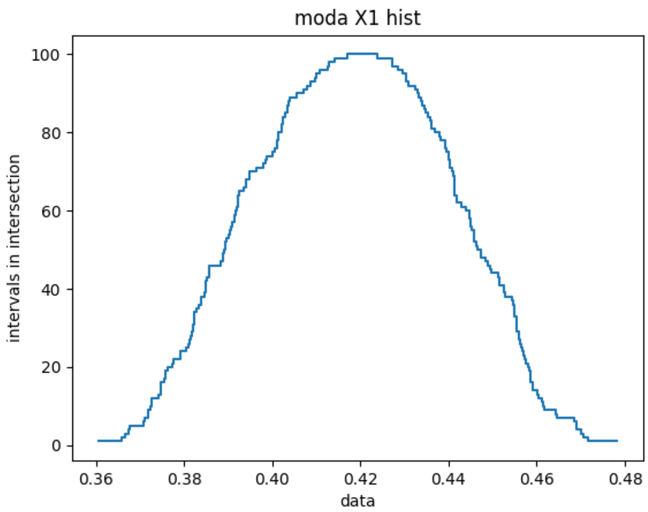

Рис. 2: Частота пересечений подинтервалов с интервалами выборки X_1

Затем для X_2.

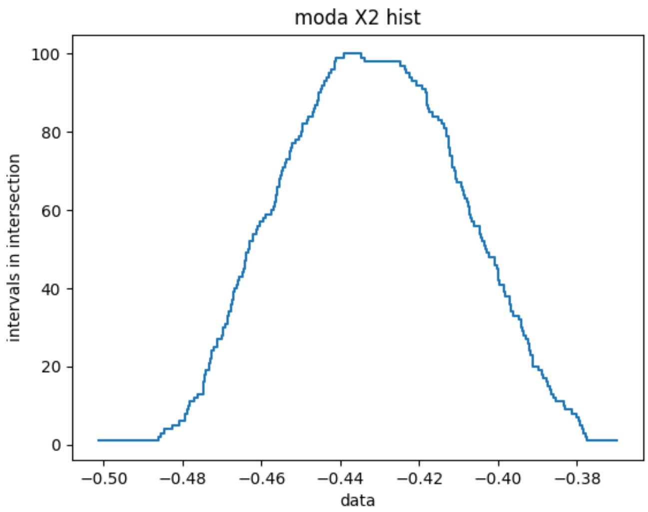

Рис. 3: Частота пересечений подинтервалов с интервалами выборки X_2

Посчитаем индекс Жаккара обеих выборок. $JK(X_1) = 0.528, JK(X_2) = 0.516$. Верхняя и нижняя границы $\underline{R} = -0.975, \overline{R} = -0.953$. Найдем оптимальное значение R (для наглядности на графике 4 изображён более широкий интервал значений R).

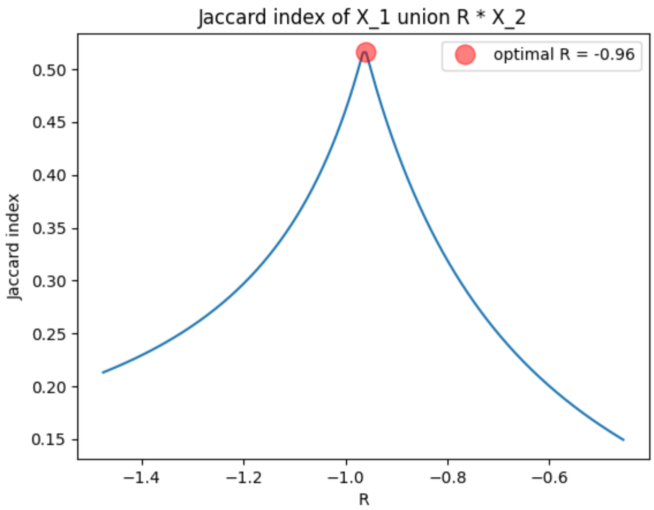

Рис. 4: Зависимость индекса Жаккара от значения R

Оптимальное значение R оказалось равно $R_{opt} = -0.960$ Построим объединённую выборку $X = X_1 \cup R_{opt} X_2$.

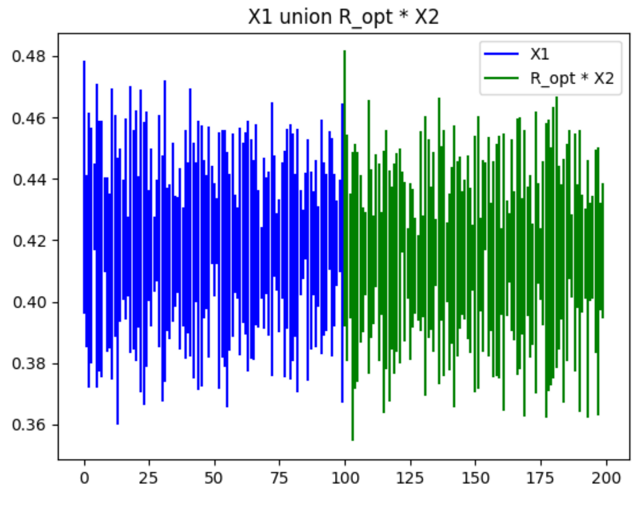

Рис. 5: Объединённая выборка $X_1 \cup R_{opt} X_2$

Индекс Жаккара полученной выборки равен $JK(X) = 0.516$.

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

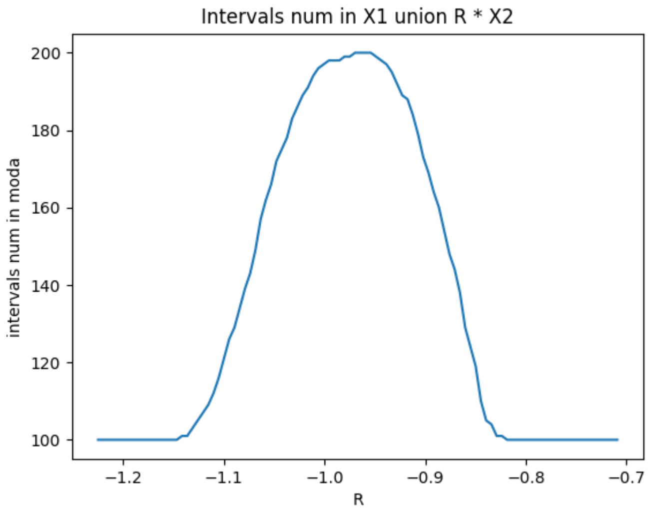

Рис. 6: Зависимость частоты пересечения моды с интервалами $X_1 \cup RX_2$

Найдём внутреннюю и внешнюю интервальные оценки $\mathbf{R} = [R_{in}, R_{out}]$. Для этого введём уровень доверия $\alpha = 0.95$ и найдем крайние значений R, удовлетворяющие $JK(R) > JK(R_{opt}) * \alpha$ для внутренней оценки и $\mu(R) > \mu(R_{opt}) * \alpha$ для внешний. Результаты представлены на рис. 7 (график $\mu(R)$ нормирован так, чтобы $\max_R \mu(R)$ и $\max_R JK(R)$ были равны).

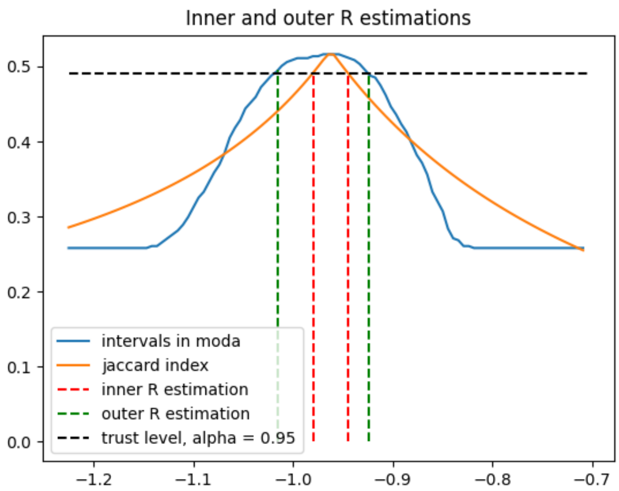

Рис. 7: Внутренняя и внешняя оценки R

В итоге получили следующие оценки: $R_{in} = [-0.978, -0.945]$, $R_{out} = [-1.015, -0.924]$. Тогда $\mathbf{R} = [[-0.978, -0.945], [-1.015, -0.924]]$.

5 Обсуждение

Из полученных результатов можно заметить следующее. Как видно на рисунке 4 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 5.

На рисунках 4, **??**, 7 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.