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:
Бендюков Александр
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 **14**

Список иллюстраций

1 Исходные данные выборка X_1 3

2 Гистограмма распределения δ_i для X_1 4

3 Исходные данные выборка X_2 4

4 Гистограмма распределения δ_i для X_2 5

5 Интервальная выборка X_1 5

6 Интервальная выборка X_2 6

7 Частота пересечений подинтервалов с интервалами выборки X_1 6

8 Частота пересечений подинтервалов с интервалами выборки X_2 7

9 Зависимость индекса Жаккара от значения R 8

10 Объединённая выборка $X_1 \cup R_{opt}X_2$ 8

11 Частота пересечений подинтервалов с интервалами выборки $X_1 \cup R_{opt}X_2$ 9

12 Зависимость частоты пересечения моды с интервалами $X_1 \cup RX_2$ 10

13 Внутренняя и внешняя оценки R 11

14 Интервальная выборка X_1' 12

15 Интервальная выборка X_2' 12

16 Зависимость индекса Жаккара от значения R 13

17 Зависимость числа интервалов в моде от R 13

18 Объединённая выборка $X_1' \cup R_{opt}' X_2'$ 14

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.

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

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 $\underline{R}, \overline{R}$.

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

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

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.

$$\mathbf{x}_i = [(x_i - \delta_i) - \varepsilon, (x_i - \delta_i) + \varepsilon], \varepsilon = \frac{1}{2^{14}} \tag{6}$$

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

Для начала рассмотрим исходные данные с учётом и без учёта δ_i.

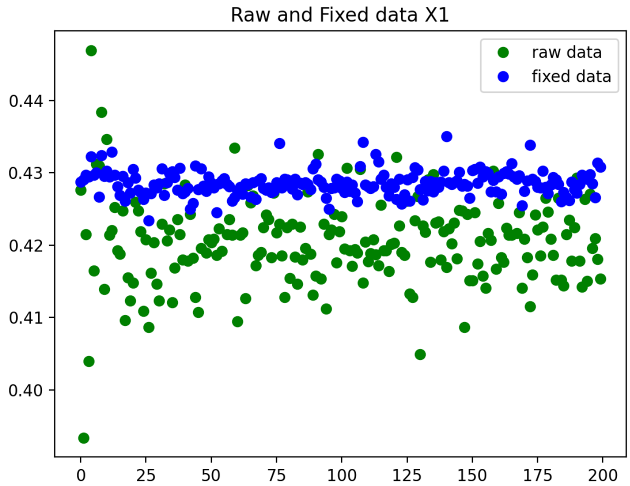

Рис. 1: Исходные данные выборка X_1

Гистограмма распеределения δ_i для X_1 имеет вид.

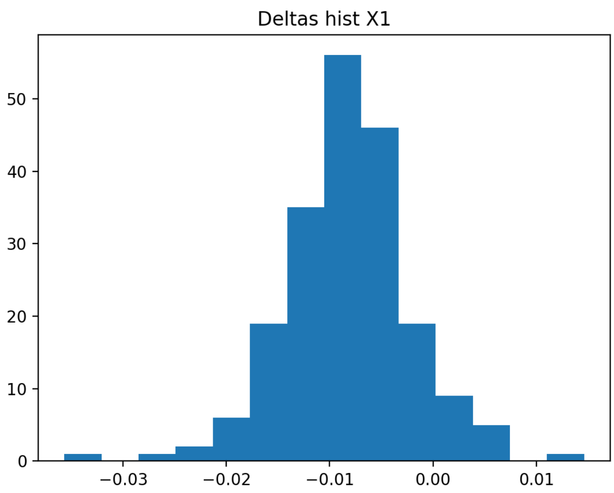

Рис. 2: Гистограмма распределения δ_i для X_1

Тоже самое для X_2

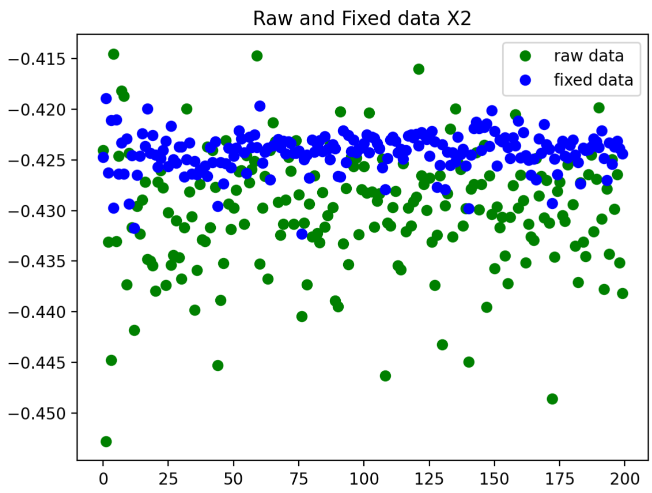

Рис. 3: Исходные данные выборка X_2

Гистограмма распередеделения δ_i для X_2 имеет вид.

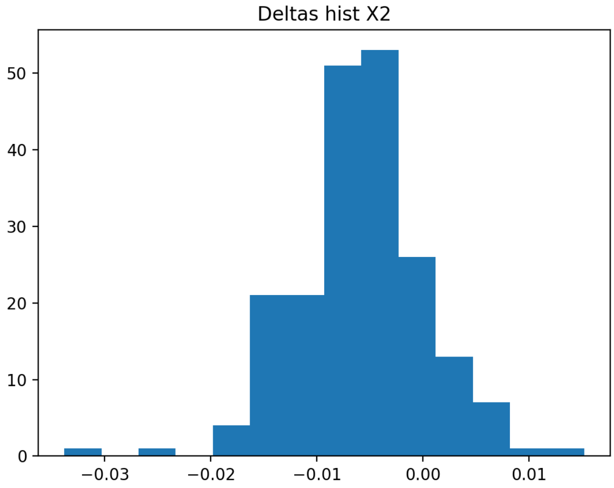

Рис. 4: Гистограмма распределения δ_i для X_2

На рис. 1, 3 видно, что учёт δ_i значительно уменьшил разброс исходных данных.

Теперь посмотрим на построенные интервальные выборки X_1, X_2.

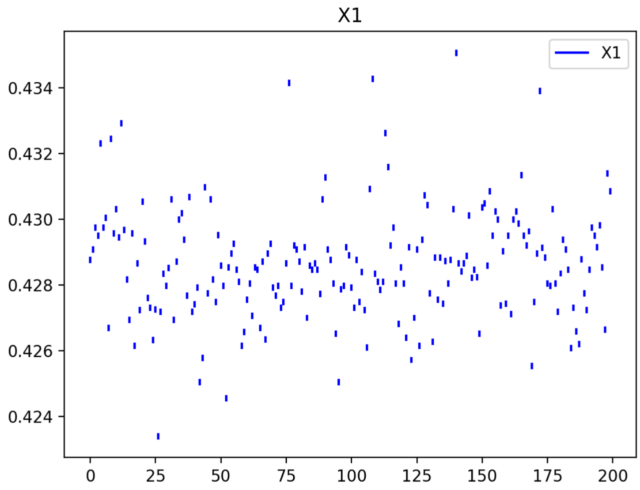

Рис. 5: Интервальная выборка X_1

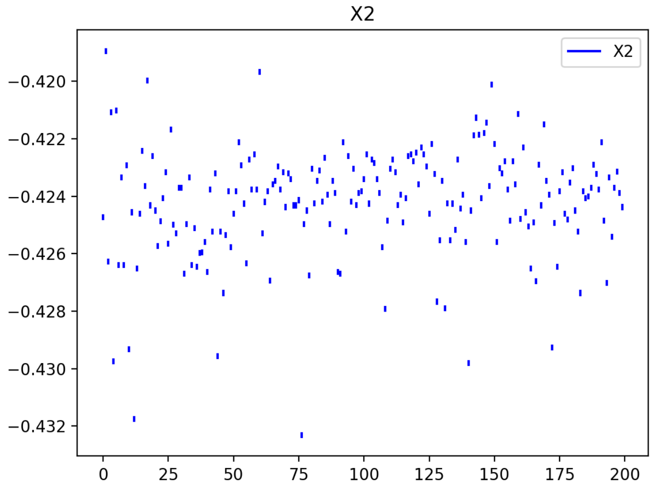

Рис. 6: Интервальная выборка X_2

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

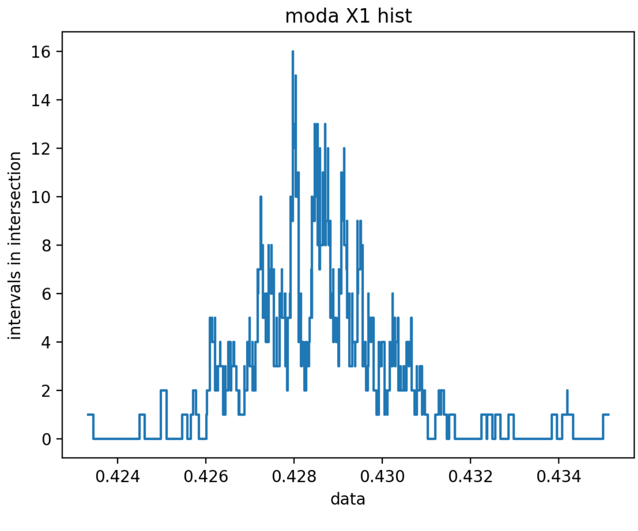

Рис. 7: Частота пересечений подинтервалов с интервалами выборки X_1

Затем для X_2.

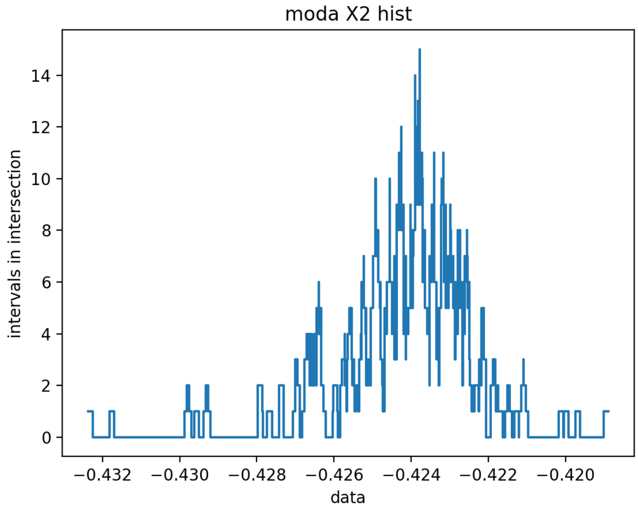

Рис. 8: Частота пересечений подинтервалов с интервалами выборки X_2

Мода для выборки X_1 равна интервалу $\mu_{X_1} = [0.427979, 0.427981]$, для выборки X_2 мода равно интервалу $\mu_{X_2} = [-0.423771, -0.423769]$.

Посчитаем индекс Жаккара обеих выборок. $JK(X_1) = 0.01036$, $JK(X_2) = 0.00905$. Найдем оптимальное значение R (для наглядности на графике 9 изображён более широкий интервал значений R).

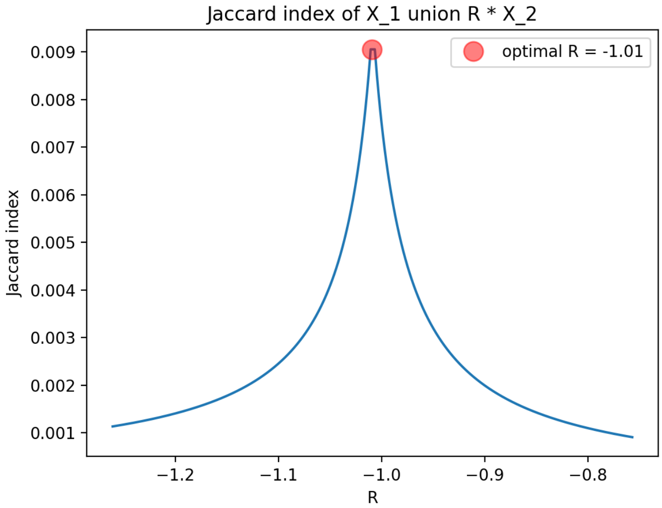

Рис. 9: Зависимость индекса Жаккара от значения R

Оптимальное значение R оказалось равно $R_{opt} = -1.0095$ Построим объединённую выборку $X = X_1 \cup R_{opt}X_2$.

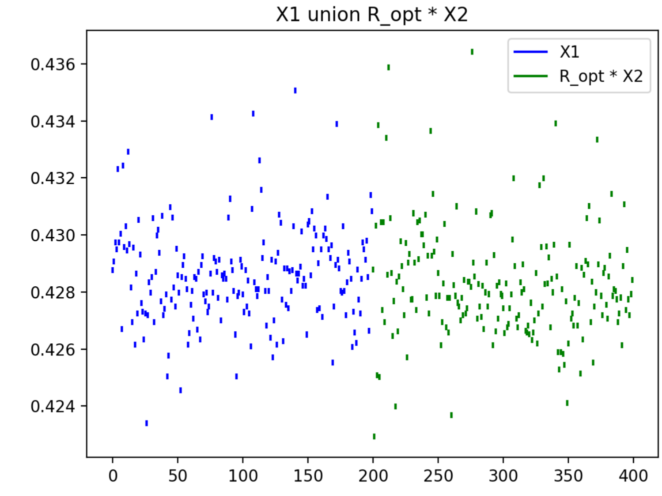

Рис. 10: Объединённая выборка $X_1 \cup R_{opt}X_2$

Индекс Жаккара полученной выборки равен $JK(X) = 0.00905$.

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

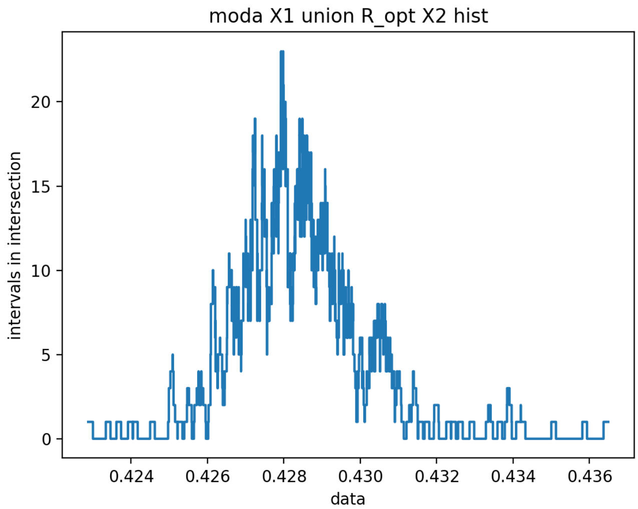

Рис. 11: Частота пересечений подинтервалов с интервалами выборки $X_1 \cup R_{opt} X_2$

Мода для объединённой выборки $X_1 \cup R_{opt} X_2$ равна интервалу $\mu_{X_1 \cup R_{opt} X_2} = [0.427926, 0.427928]$.

Посмотрим на зависимость частоты пересечений моды $\mu(R)$ с интервалами для объединённой выборки $X_1 \cup R X_2$ в зависимости от значений R.

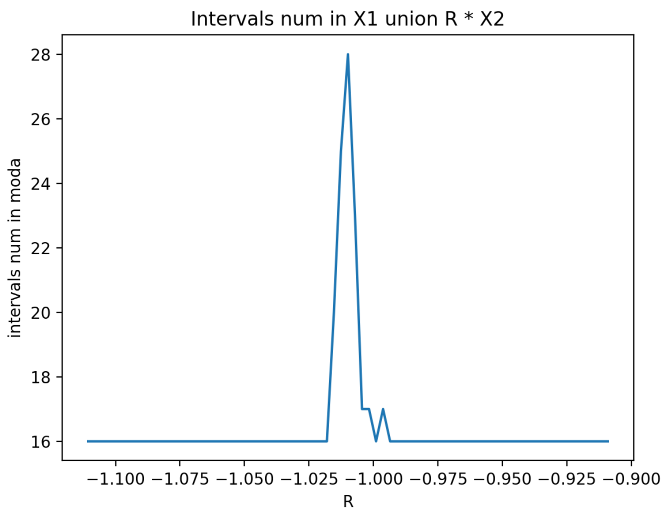

Рис. 12: Зависимость частоты пересечения моды с интервалами $X_1 \cup RX_2$

Найдём внутреннюю оценку $\mathbf{R}$ двумя способами: используя индекс Жаккара и моду. Для этого введём уровень доверия $\alpha = 0.95$ и найдем крайние значений R, удовлетворяющие $JK(R) > JK(R_{opt}) * \alpha$ в случае индекса Жаккара и $\mu(R) > \mu(R_{opt}) * \alpha$ в случае моды. Результаты представлены на рис. 13 (график $\mu(R)$ нормирован так, чтобы $\max_R \mu(R)$ и $\max_R JK(R)$ были равны).

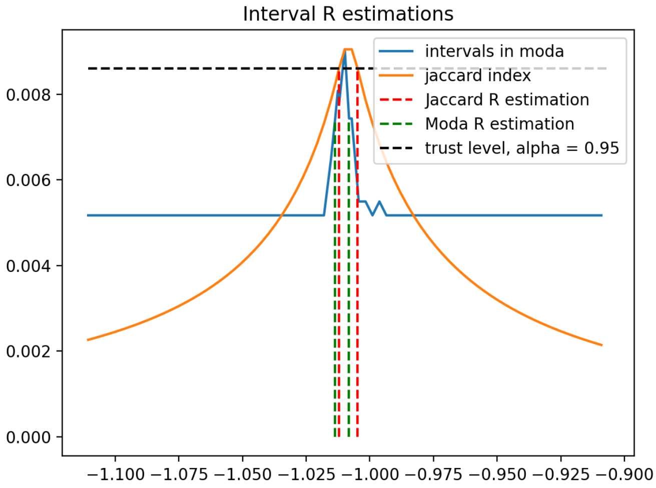

Рис. 13: Внутренняя и внешняя оценки R

В итоге получили следующие оценки: $R_{JK} = [-1.012119, -1.004806]$, $R_\mu = [-1.01361, -1.008163]$.

Внешнюю оценку получим по формулам 4, 5 $R_{out} = [-1.01062, -1.006362]$.

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.

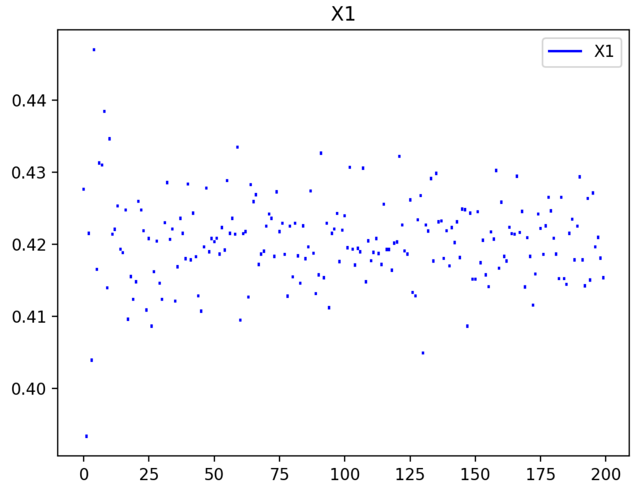

Рис. 14: Интервальная выборка X_1'

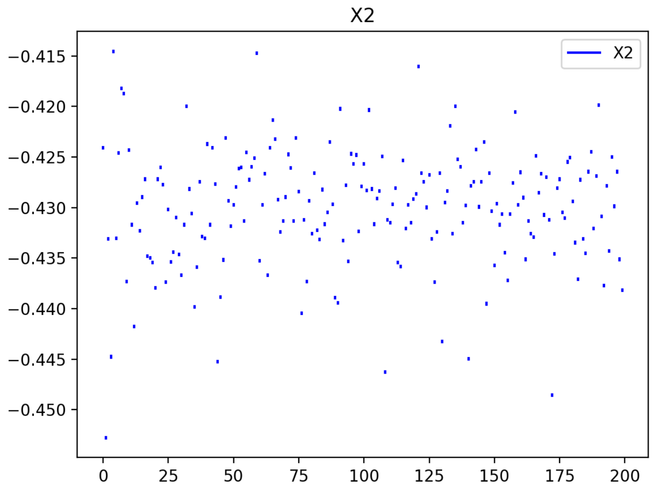

Рис. 15: Интервальная выборка X_2'

Вычислим индекс Жаккара $JK(X_1') = 0.00227, JK(X_2') = 0.00318$.
Зависимость индекса Жаккара от значения параметра R имеет вид.

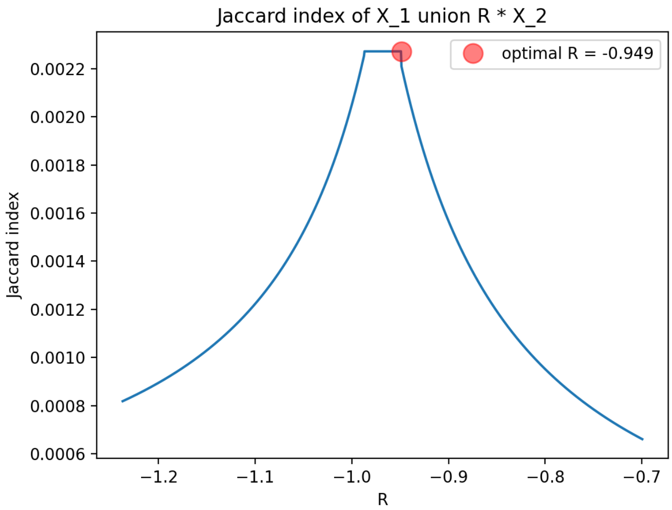

Рис. 16: Зависимость индекса Жаккара от значения R

Также построим зависимость числа интервалов в моде от параметра R.

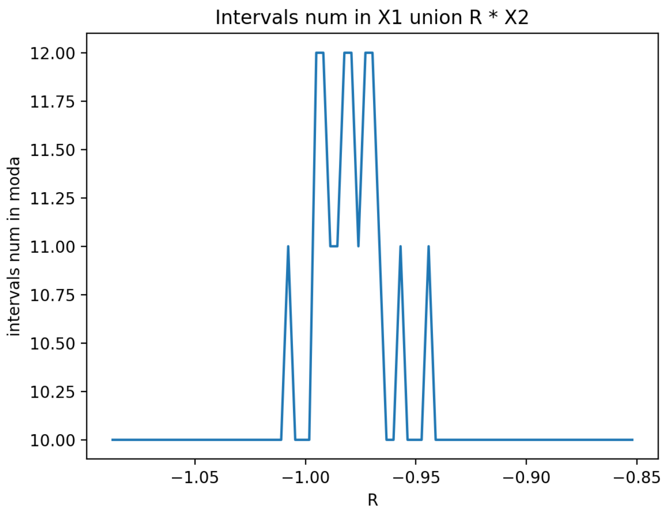

Рис. 17: Зависимость числа интервалов в моде от R

Видно, что оптимальное значение параметра R равно $R'_{opt} = -0.94892$, что значительно отличается от первого случая. Тогда объелинённая выборка $X_1' \cup R_{opt}' X_2'$ имеет вид.

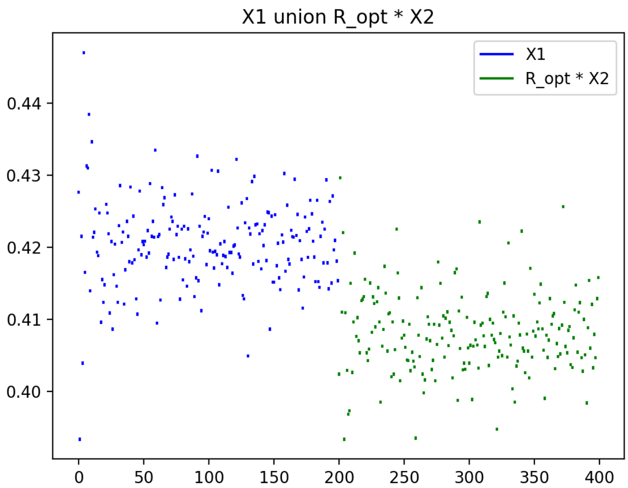

Рис. 18: Объединённая выборка $X_1' \cup R_{opt}' X_2'$

5 Обсуждение

Из полученных результатов можно заметить следующее. Как видно на рисунке 9 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 10.