

Санкт-Петербургский политехнический университет  
Петра Великого

Физико-механический институт  
Высшая школа прикладной математики и физики

**Отчёт**  
**по лабораторной работе №9**  
**по дисциплине**  
**«Математическая статистика»**

Выполнил студент:  
Павлов Илья Сергеевич  
Группа: 5030102/90201  
Проверил:  
к.ф.-м.н., доцент  
Баженов Александр Николаевич

Санкт-Петербург  
2022

# Содержание

<b>1</b>	<b>Постановка задачи</b>	<b>3</b>
<b>2</b>	<b>Теория</b>	<b>4</b>
2.1	Представление данных . . . . .	4
2.2	Простая линейная регрессия . . . . .	4
2.3	Метод наименьших модулей . . . . .	5
2.4	Предварительная обработка данных . . . . .	5
2.5	Коэффициент Жаккара . . . . .	5
2.6	Процедура оптимизации . . . . .	6
<b>3</b>	<b>Реализация</b>	<b>6</b>
<b>4</b>	<b>Результаты</b>	<b>6</b>
<b>5</b>	<b>Обсуждение</b>	<b>13</b>
5.1	Модель дрейфа . . . . .	13
5.2	Гистограммы скорректированных моделей и объединенной выборки . . . . .	13
5.3	Коэффициент Жаккара. Оптимальное значение коэффициента калибровки . .	13
<b>6</b>	<b>Ссылка на репозиторий</b>	<b>13</b>
	<b>Список литературы</b>	<b>14</b>

## Список иллюстраций

1	Схема установки . . . . .	3
2	Исходные выборки . . . . .	6
3	"Обинтерваленные" значения первой выборки . . . . .	7
4	Линейная регрессия для первой выборки . . . . .	7
5	Гистограмма коэффициентов коррекции для первой выборки . . . . .	8
6	"Спрявленные" значения первой выборки . . . . .	8
7	Гистограмма для скорректированной модели данных первой выборки . . . . .	9
8	"Обинтерваленные" значения второй выборки . . . . .	9
9	Линейная регрессия для второй выборки . . . . .	10
10	Гистограмма коэффициентов коррекции для второй выборки . . . . .	10
11	"Спрявленные" значения второй выборки . . . . .	11
12	Гистограмма для скорректированной модели данных второй выборки . . . . .	11
13	Коэффициент Жаккара от калибровочного коэффициента . . . . .	12
14	Гистограмма объединенной выборки . . . . .	12

# 1 Постановка задачи

Исследование из области солнечной энергетики. На Рис.1 показана схема установки для исследования фотоэлектрических характеристик.

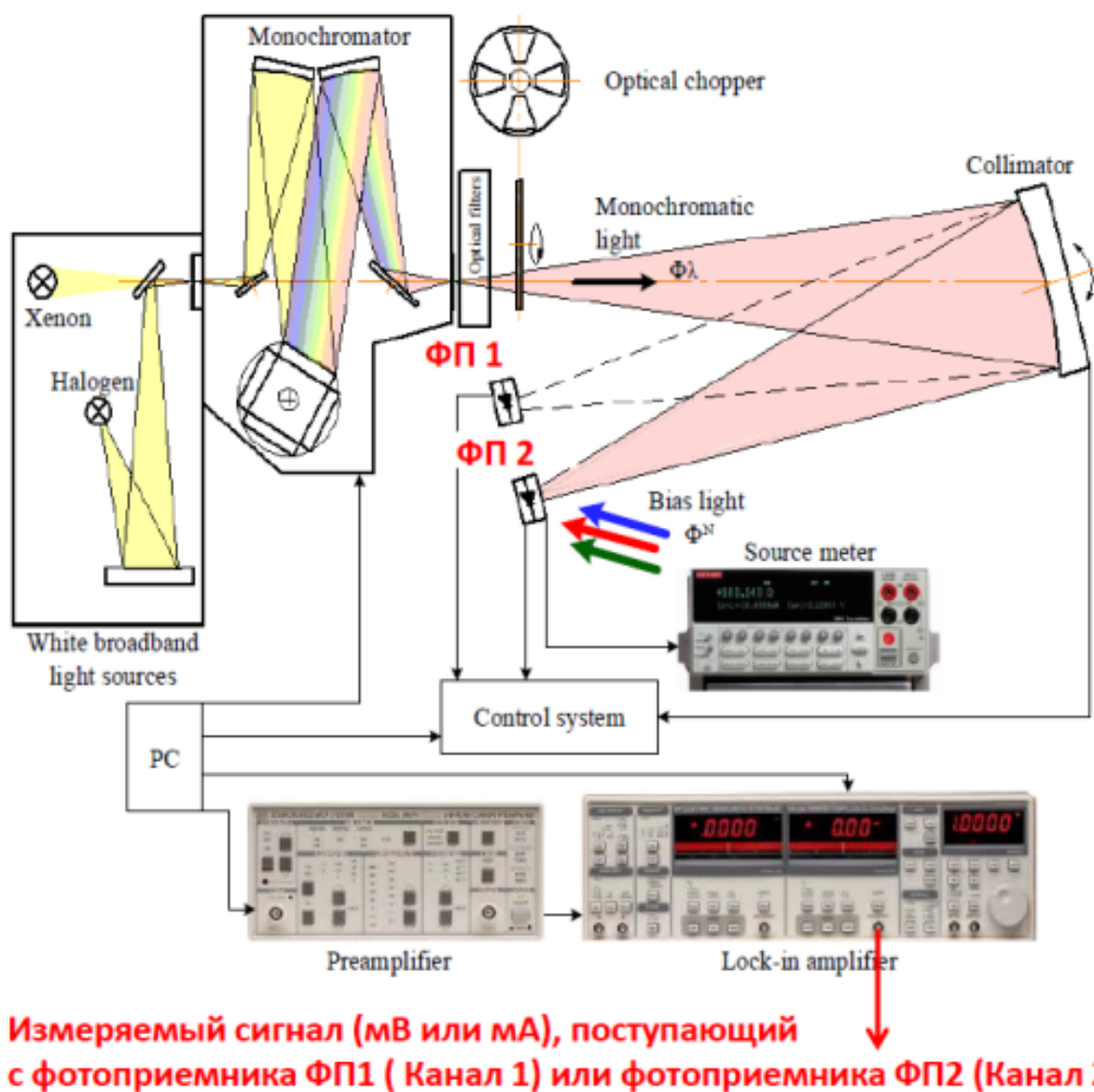


Рис. 1: Схема установки

Калибровка датчика ФП1 производится по эталону ФП2. Зависимость между квантовыми эффективностями датчиков предполагается постоянной для каждой пары наборов измерений

$$QE_2 = \frac{I_2}{I_1} * QE_1. \quad (1)$$

$QE_{1,2}$  – квантовые эффективности эталонного и исследуемого датчиков,  $I_{1,2}$  – измеренные токи.

**Исходные данные:** Имеется 2 выборки данных с интервальной неопределенностью. Одна из них относится к эталонному датчику ФП2. Другая выборка соответствует исследуемому датчику ФП1. Данные представлены в виде двух csv файлов с числом отсчетов 200.

**Требуется определить:** коэффициент калибровки

$$R_{21} = \frac{I_2}{I_1}. \quad (2)$$

при помощи линейной регрессии на множестве интервальных данных и коэффициента Жаккара.

## 2 Теория

### 2.1 Представление данных

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределенностью. Один из распространенных способов получения интервальных результатов в первичных измерениях — это "обинтерваливание" точечных значений, когда к точечному базовому значению  $\dot{x}$ , которое считывается по показаниям измерительного прибора прибавляется интервал погрешности  $\epsilon$ :

$$x = \dot{x} + \epsilon \quad (3)$$

Интервал погрешности зададим как

$$\epsilon = [-\xi, \xi]. \quad (4)$$

В конкретных измерениях примем  $\xi = 10^{-4}$  мВ.

Согласно терминологии интервального анализа, рассматриваемая выборка — это вектор интервалов, или интервальный вектор  $x = (x_1, x_2, x_3, \dots)$ .

Информационным множеством в случае оценивания единичной физической величины по выборке интервальных данных будет также интервал, который называют также информационным интервалом. Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые "совместны" с измерениями выборки.

### 2.2 Простая линейная регрессия

Регрессионную модель описания данных называют простой линейной, если заданный набор данных аппроксимируется прямой с внесенной добавкой в виде некоторой нормально распределенной ошибки:

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i, i \in \overline{1, n} \quad (5)$$

где  $\beta_0, \beta_1$  — параметры подлежащие оцениванию. В данном случае рассматриваем модель без внесения добавки.

## 2.3 Метод наименьших модулей

Данный метод основан на минимизации  $l^1$ -нормы разности последовательностей полученных экспериментальных данных  $y_n$  и значений аппроксимирующей функции  $f(x_n)$ :

$$\|f(x_n) - y_n\|_{l^1} \rightarrow \min \quad (6)$$

В данном случае мы ставим задачу линейного программирования таким образом, чтобы найти не только  $\beta_0$  и  $\beta_1$ , но и вектор  $w$  множителей коррекции. Тогда задача ставится в следующем виде:

$$\sum |w_i| \rightarrow \min, i \in 1, \bar{n} \quad (7)$$

При ограничениях:

$$\beta_0 + \beta_1 * x_i - w_i * \xi \leq y_i, i \in 1, \bar{n} \quad (8)$$

$$\beta_0 + \beta_1 * x_i + w_i * \xi \leq y_i, i \in 1, \bar{n} \quad (9)$$

## 2.4 Предварительная обработка данных

Из графического представления выборок ясно, что для оценки коэффициента калибровки необходима предварительная обработка данных. Для этого зададимся линейной моделью дрейфа:

$$Lin_{1,2} = A_{1,2} + B_{1,2} * n + \epsilon_i, n \in 1, \bar{N}. \quad (10)$$

Поставим задачу линейного программирования 7-9 и найдем коэффициенты  $A_{1,2}$ ,  $B_{1,2}$  и вектор  $w_{1,2}$  множителей коррекции данных. Множитель коррекции необходим для того чтобы получить данные, согласующиеся с полученной линейной моделью дрейфа:

$$I_{1,2}^f(n) = \dot{x}(n) + \epsilon * w_{1,2}(n), n \in 1, \bar{N}. \quad (11)$$

После построения линейной модели дрейфа необходимо построить "спрямленные" данные выборки:

$$I_{1,2}(n) = I_{1,2}^f(n) - B_{1,2} * n, n \in 1, \bar{N}. \quad (12)$$

## 2.5 Коэффициент Жаккара

Нами рассматривается модификация индекса Жаккара для интервальных данных:

$$JK(x) = \frac{wid(\cap x_i)}{wid(\cup x_i)} \quad (13)$$

В качестве меры рассматривается ширина интервала, а вместо пересечения и объединения — взятие минимума и максимума по включению двух величин в интервальной арифметике. Поскольку минимум по включению может быть неправильным интервалом, коэффициент нормирован на промежутке  $[-1; 1]$ .

## 2.6 Процедура оптимизации

Для поиска оптимального параметра калибровки поставим задачу максимизации:

$$JK(x_{1\oplus 2}) \rightarrow \max \quad (14)$$

где  $x_{1\oplus 2}$  выборка полученная как конкатенация двух выборок

$$x_{1\oplus 2} = I_1^f * R \oplus I_2^f. \quad (15)$$

При этом, поскольку знак коэффициента Жаккара может свидетельствовать о совместности двух выборок (исходя из правильности минимума по включению), в качестве интервала для  $R_{21}$  можно рассматривать область где  $JK(R) \geq 0$ .

## 3 Реализация

Лабораторная работа выполнена с использованием языка Python (Jupyter Notebook) и Octave. Для реализации использовались библиотеки `padndas`, `matplotlib`, `scipy`. Отчет подготовлен с использованием TeXstudio и MikTeX.

## 4 Результаты

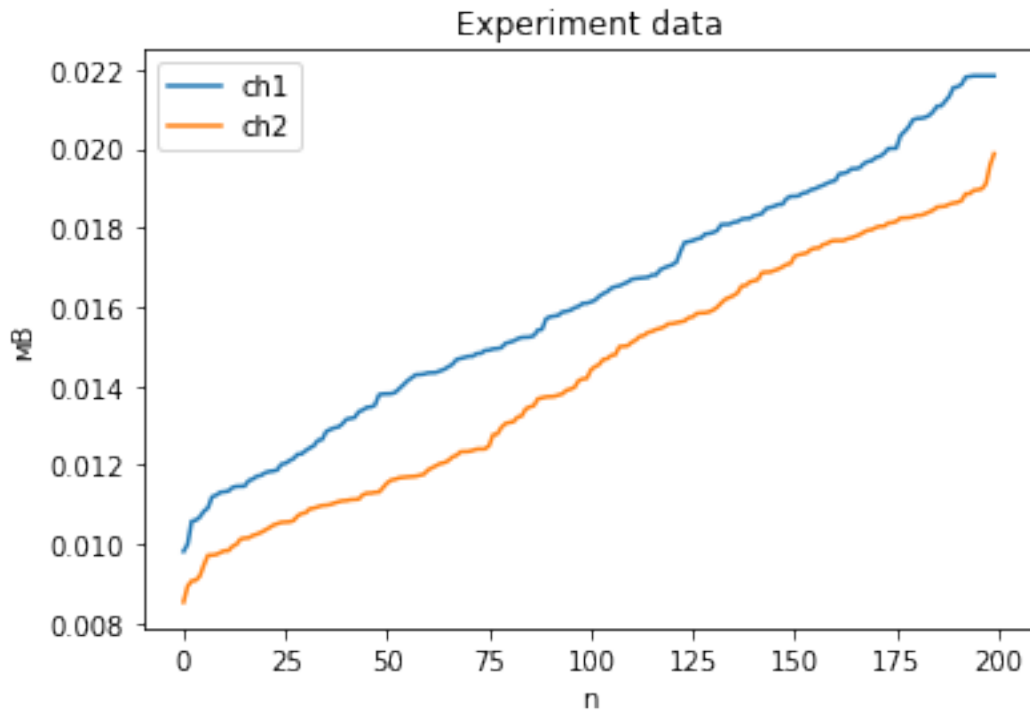


Рис. 2: Исходные выборки

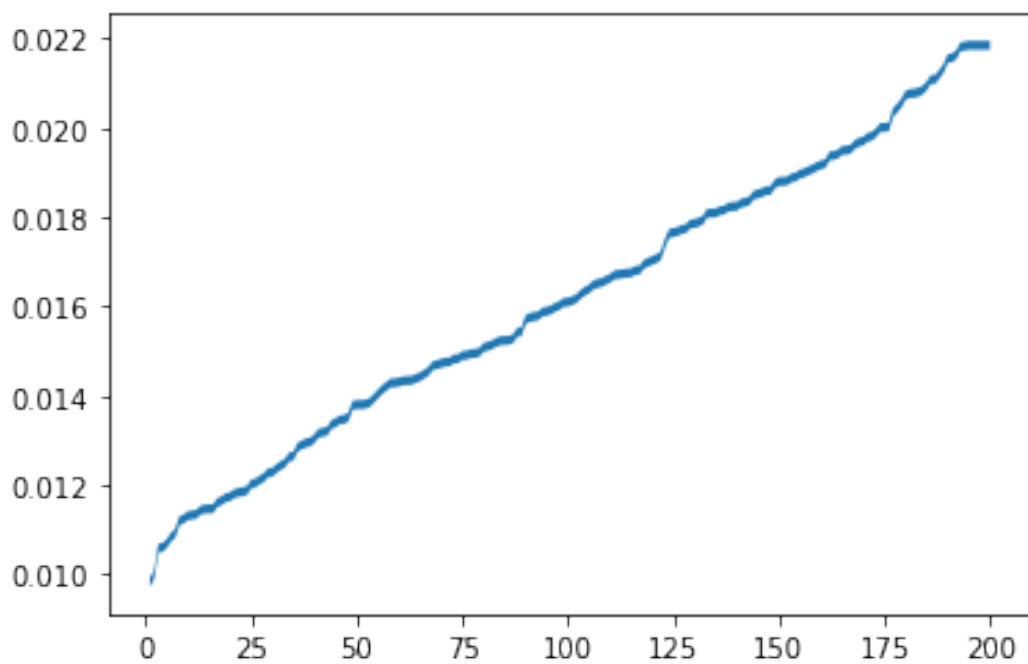


Рис. 3: "Обинтерваленные" значения первой выборки

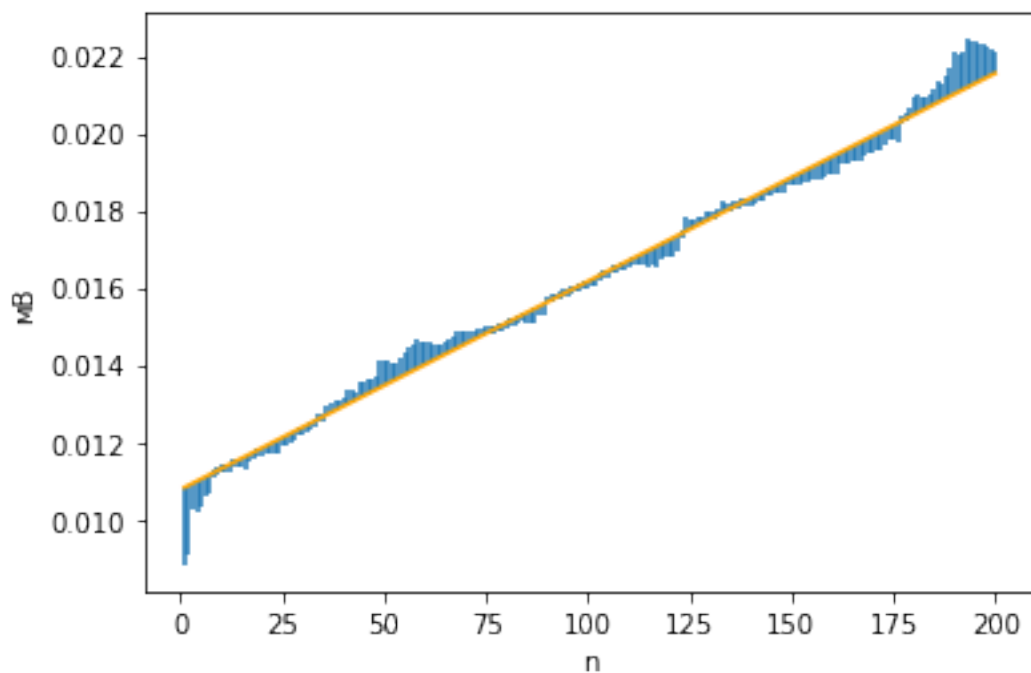


Рис. 4: Линейная регрессия для первой выборки



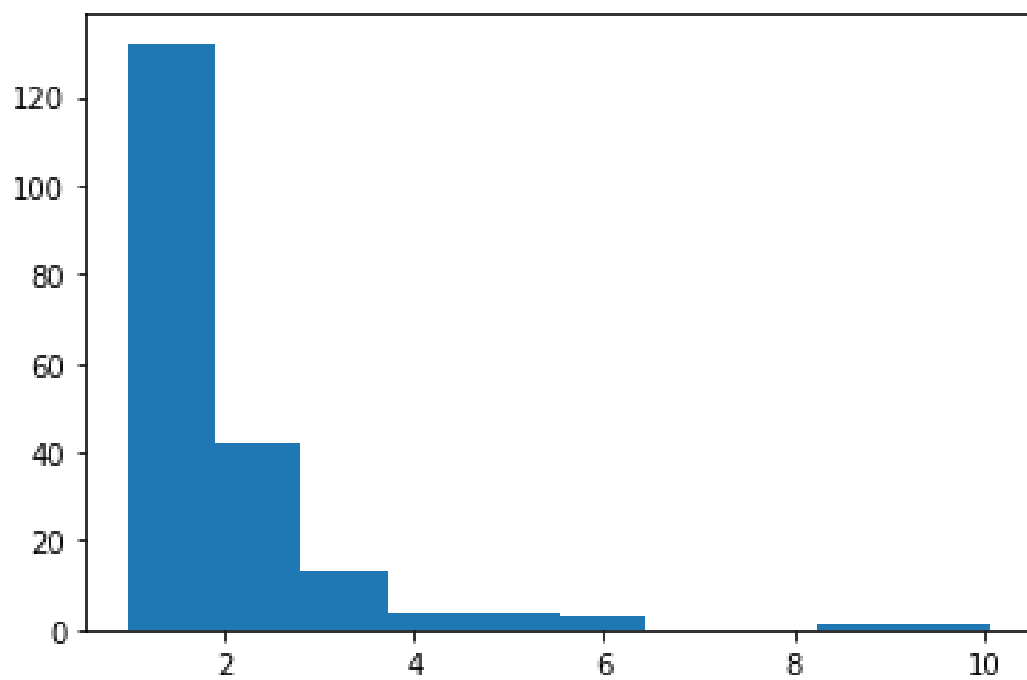


Рис. 5: Гистограмма коэффициентов коррекции для первой выборки

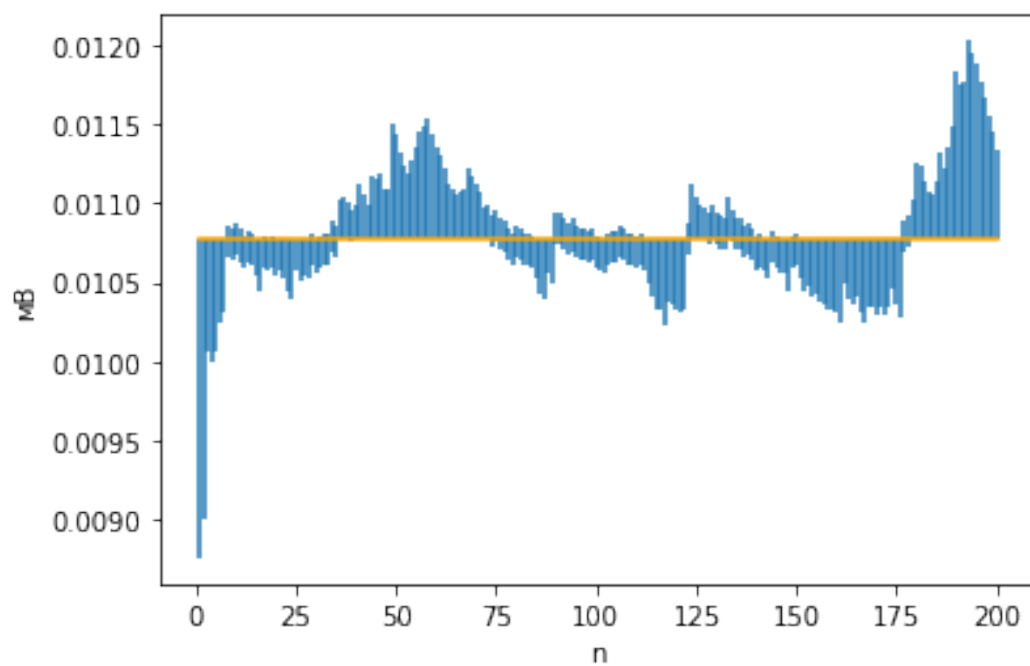


Рис. 6: "Спрявленные" значения первой выборки

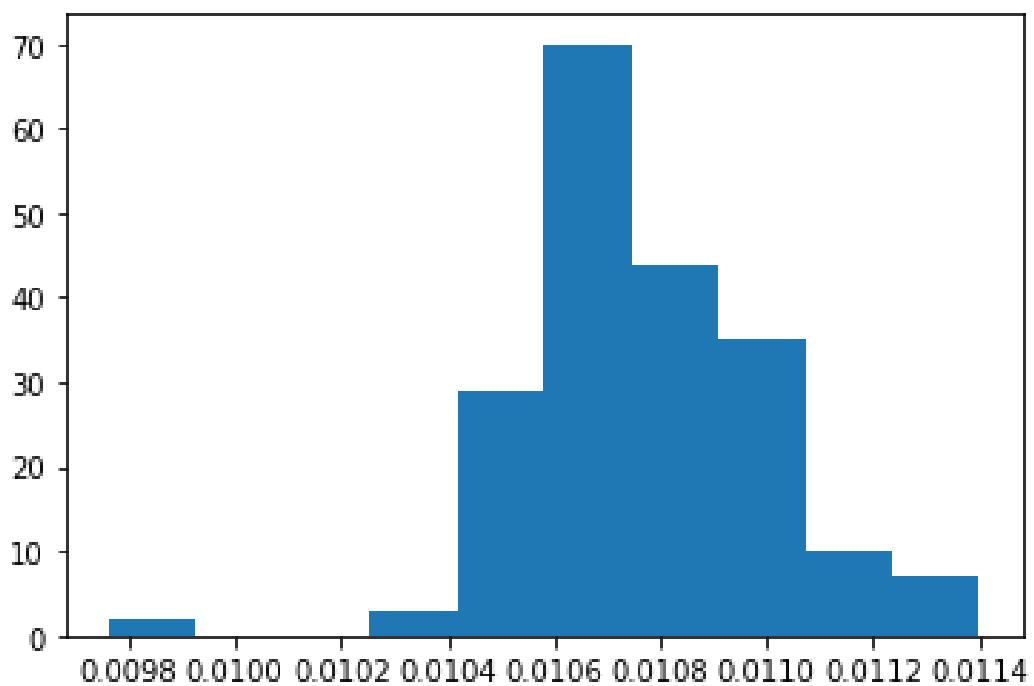


Рис. 7: Гистограмма для скорректированной модели данных первой выборки

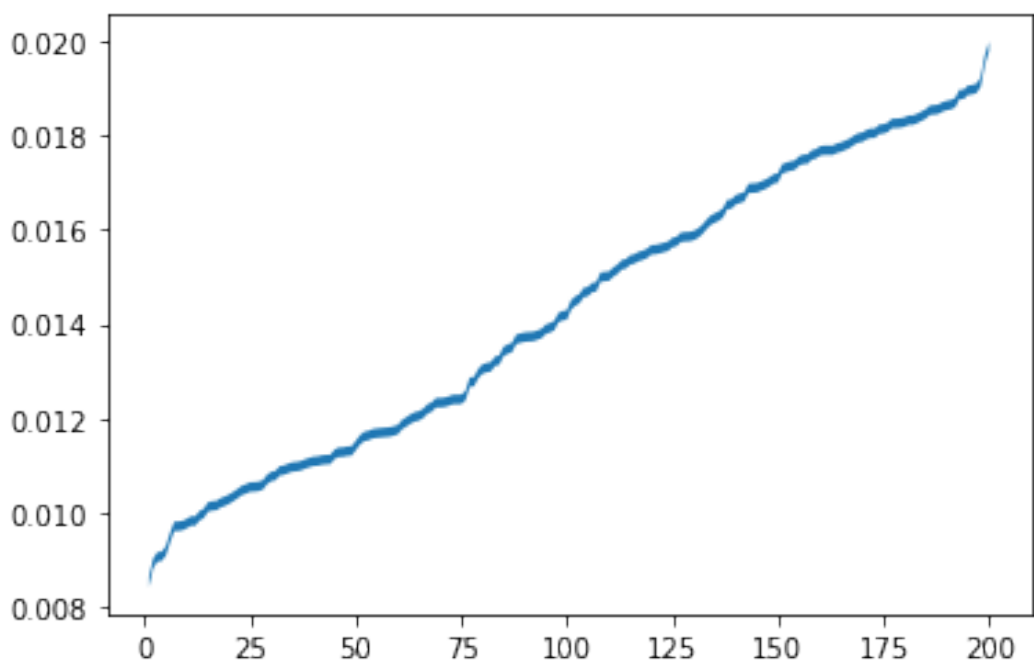


Рис. 8: "Обинтерваленные" значения второй выборки

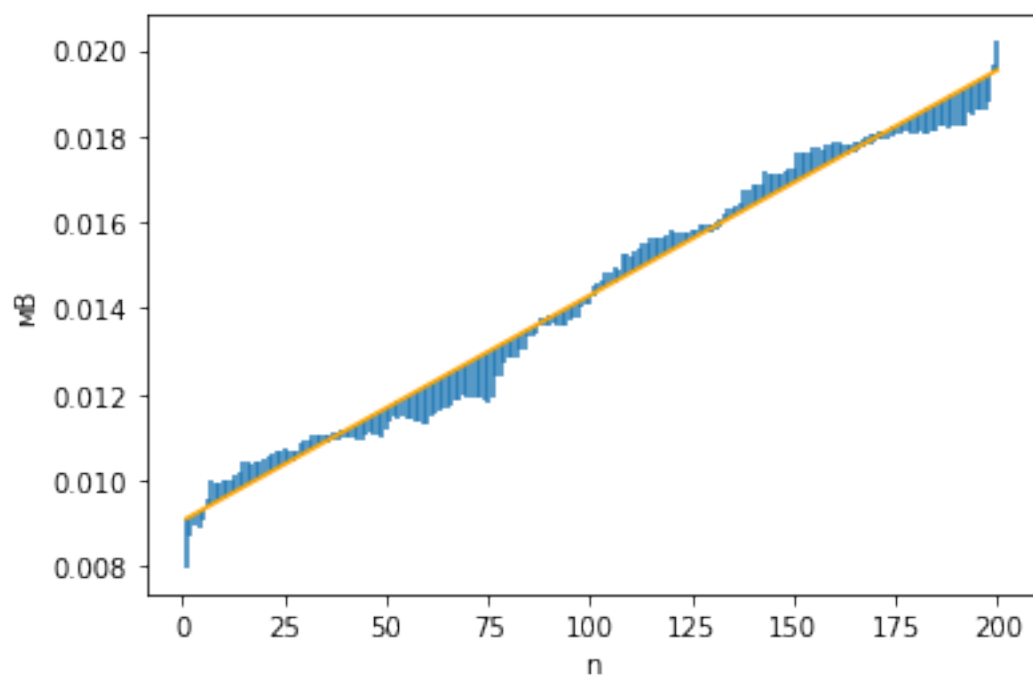


Рис. 9: Линейная регрессия для второй выборки

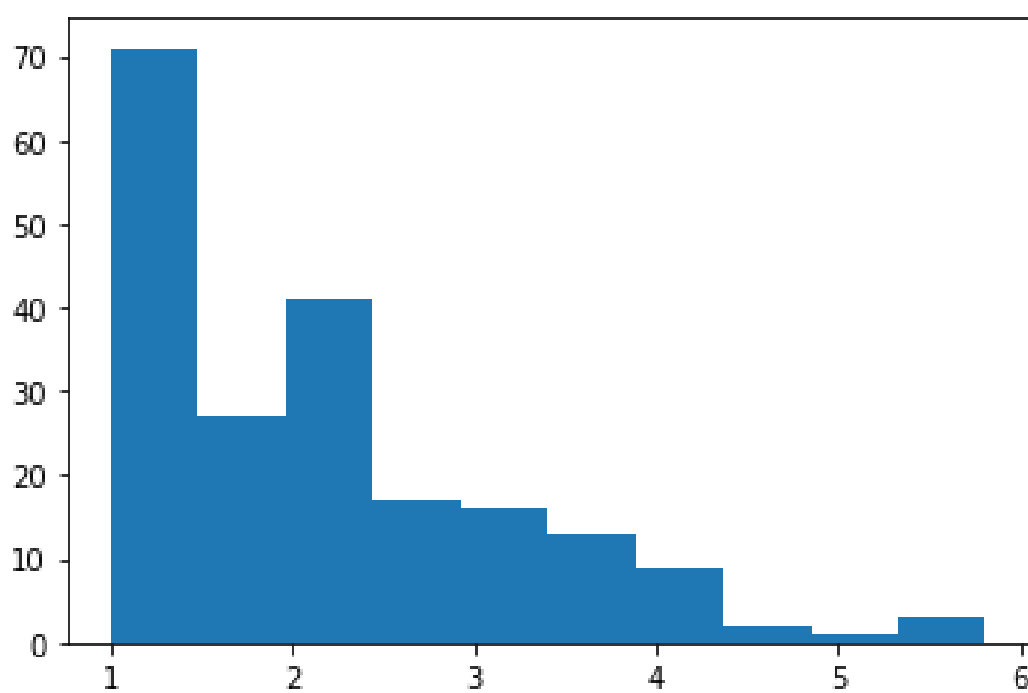


Рис. 10: Гистограмма коэффициентов коррекции для второй выборки

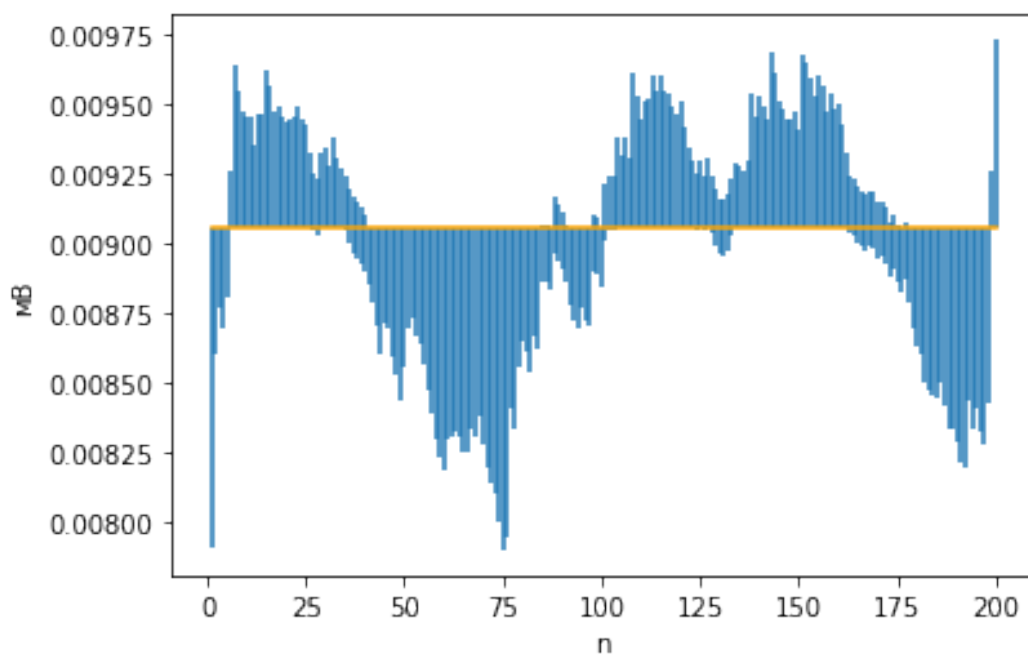


Рис. 11: "Спрявленные" значения второй выборки

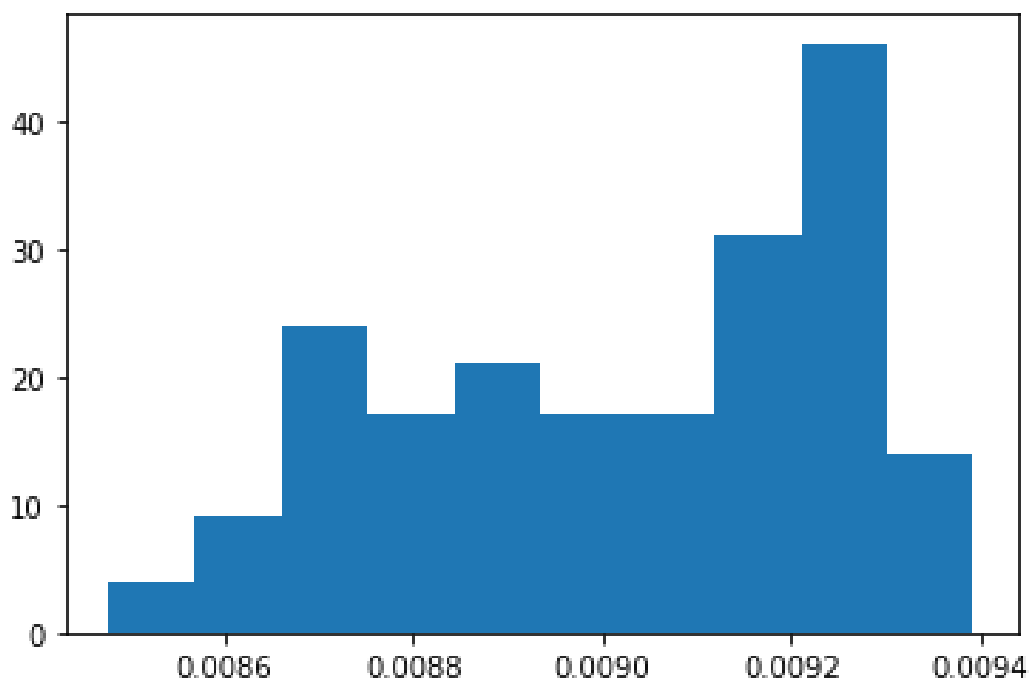


Рис. 12: Гистограмма для скорректированной модели данных второй выборки

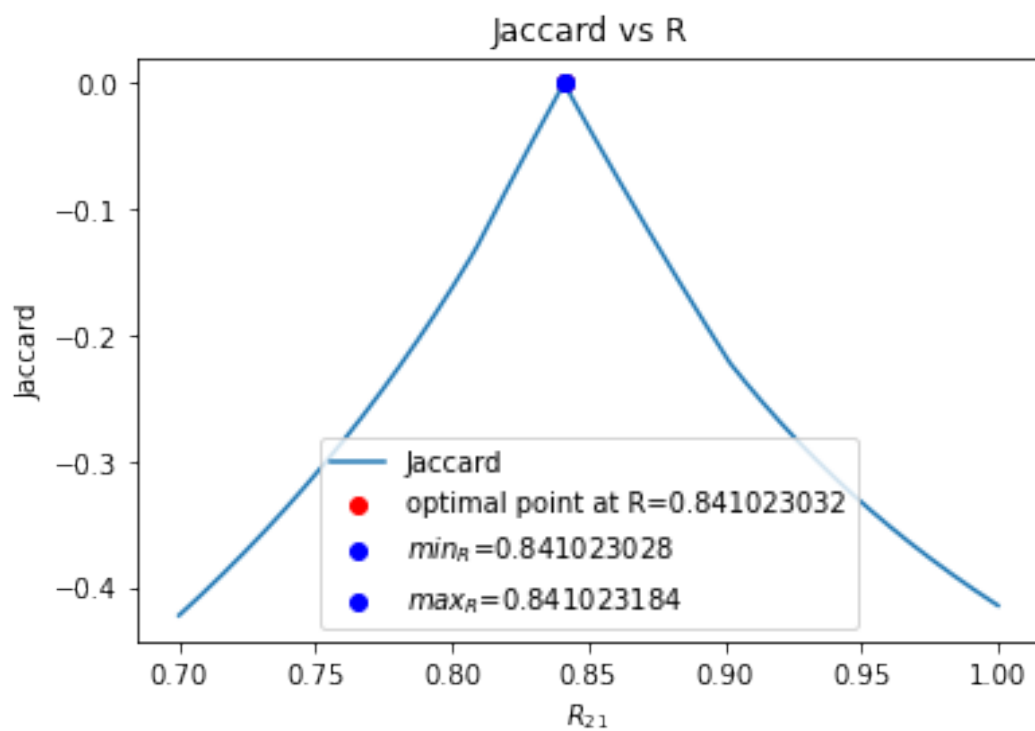


Рис. 13: Коэффициент Жаккара от калибровочного коэффициента

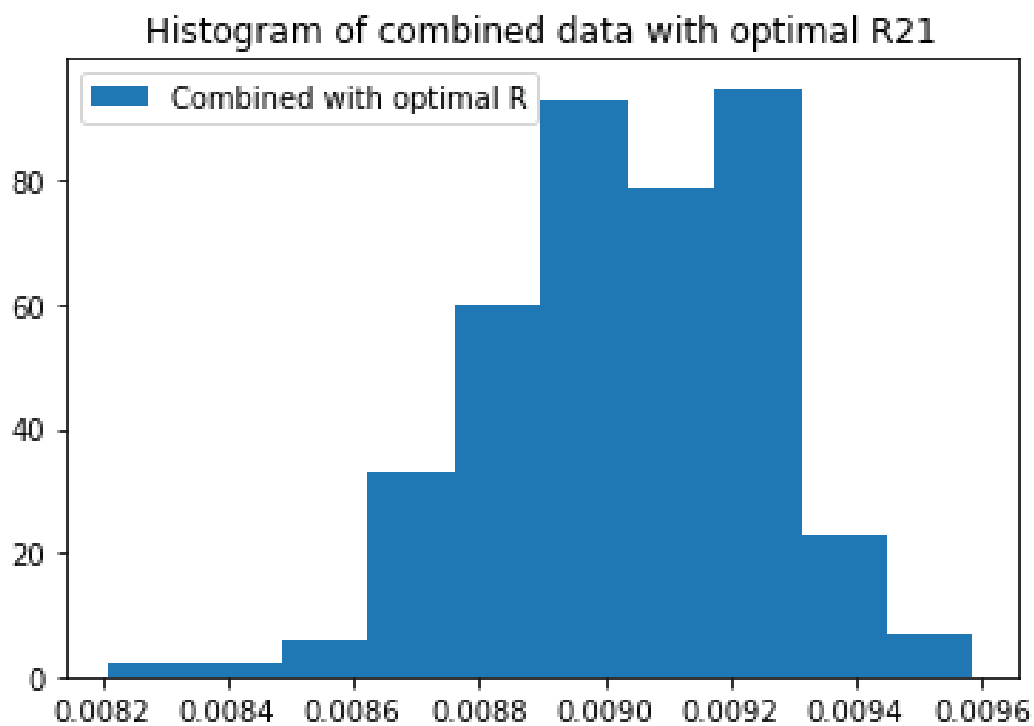


Рис. 14: Гистограмма объединенной выборки

## 5 Обсуждение

### 5.1 Модель дрейфа

Из гистограмм для  $w_1$  и  $w_2$  видно, что около половины данных потребовла коррекции, при этом величины коэффициентов коррекции не очень большие. Это говорит о том, что линейная модель дрейфа в целом неплохо описывает данные.

### 5.2 Гистограммы скорректированных моделей и объединенной выборки

Гистаграмма скорректированной второй выборки имеет характерный пик в районе 0.0092-0.0094, он же сохранился на гистограмме совмещённой выборки, при этом он стал шире, вероятно, за счёт смещения пика первой гистограммы. Заметим также, что границы гистограммы совмещенной выборки совпадают с границами для второй (эталонной) выборки.

### 5.3 Коэффициент Жаккара. Оптимальное значение коэффициента калибровки

Полученное с помощью коэффициента Жаккара оптимальное значение коэффициента калиброки  $R_{21} = 0.841023032$ , при этом видно, что интервал, на котором коэффициент Жаккара положительный очень мал (порядка  $1e-9$ ). Это указывает на то, что данные имеют ряд неточностей, которые сложно устранить.

## 6 Ссылка на репозиторий

Репозиторий с исходным кодом: <https://github.com/IlyaP01/Statistics>.

## Список литературы

- [1] А.Н. Баженов. Введение в анализ данных с интервальной неопределенностью. — Спб., 2022.
- [2] Коэффициент Жаккара. URL: [https://en.wikipedia.org/wiki/Jaccard\\_index](https://en.wikipedia.org/wiki/Jaccard_index).
- [3] С.И. Жилин. Примеры анализа интервальных данных в Octave. URL: <https://github.com/szhilin/octave-interval-examples>.