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:
Чевыкалов Григорий Андреевич
Группа: 5030102/90201
Проверил:
к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2022

Содержание

Список иллюстраций

1 Постановка задачи

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

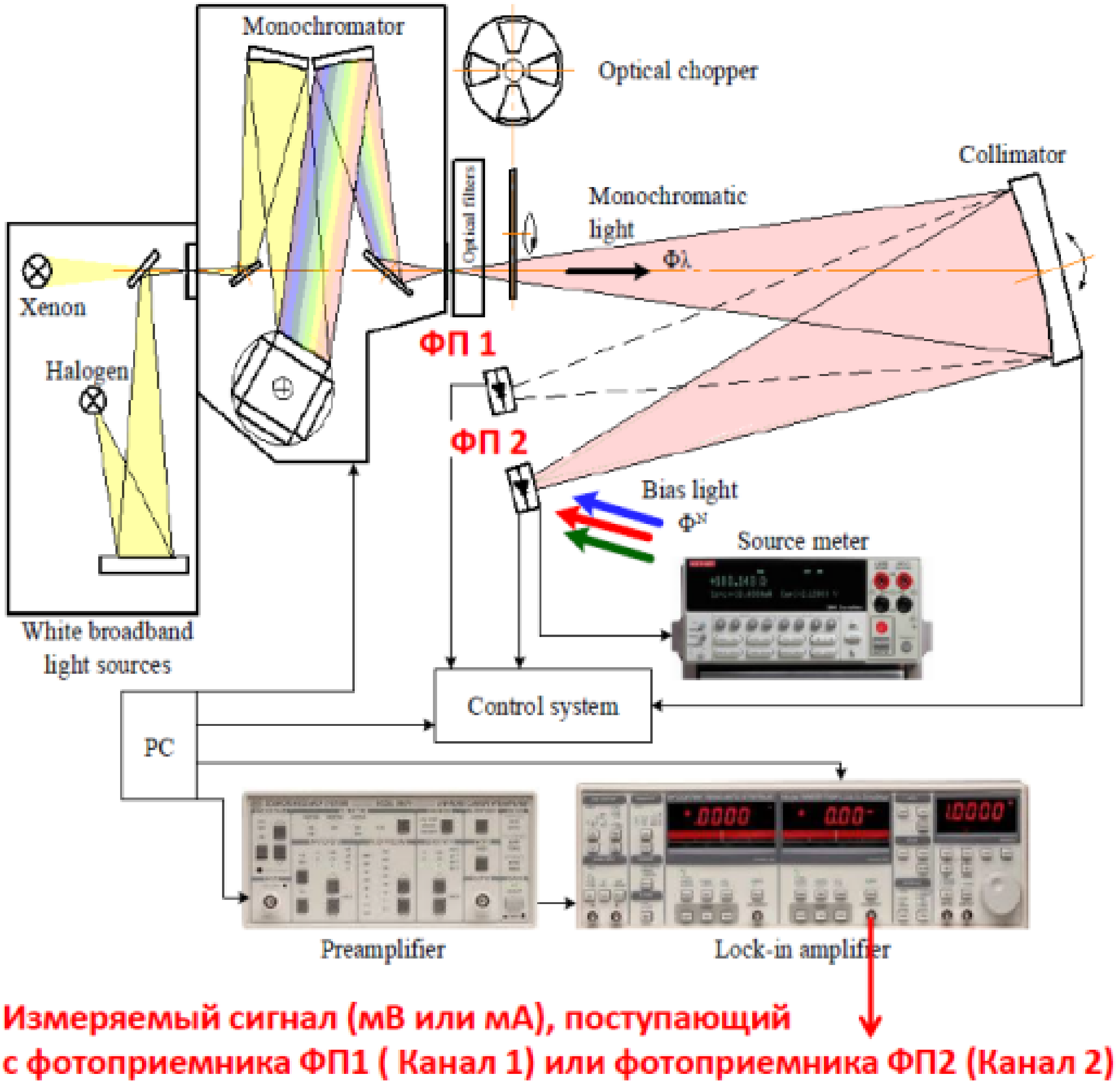

Рис. 1: Схема установки

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

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

$QE_{1,2}$ – 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 ФП2. Данные представлены в виде двух csv файлов с числом отсчетов 200.

Требуется определить: коэффициент калибровки

$$R_{21} = \frac{I_2}{I_1}. \tag{2}$$

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 интревальной неопределенностью. Один из распространненых 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 ϵ:

$$x = \dot{x} + \epsilon \tag{3}$$

Интервал погрешности зададим как

$$\epsilon = [-\xi, \xi]. \tag{4}$$

В конкретных измерениях примем $\xi = 10^{-4}$ мВ.
Согласно терминологии интервального анализа, рассматриваемая выборка — это вектор интервалов, или интервальный вектор $x = (x_1, x_2, x_3, ...)$.
Информационным множеством в случае оценивания единичной физической величины по выборке итервальных 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:

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i, \, i \in \bar{1,n} \tag{5}$$

где β_0, β_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 $f(x_n)$:

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

В данном случае мы ставим задачу линейного программирования таким образом, чтобы найти не только β_0 и β_1, но и вектор w множителей коррекции. Тогда задача ставится в следующем виде:

$$\sum |w_i| \to min, i \in \bar{1,n} \tag{7}$$

При ограничениях:

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

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

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:

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

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:

$$I^f_{1,2}(n) = \dot{x}(n) + \epsilon * w_{1,2}(n), n \in \bar{1,N}. \tag{11}$$

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

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

2.5 Коэффициент Жаккара

Нами рассматривается модификация индекса Жаккара для интервальных данных:

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

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:

$$JK(x_{1\oplus 2}) \to max \tag{14}$$

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

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

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 программирования Python (v3.10.2) и редактора Visual Studio Code (расширение Python v2022.2.1924087327), а также Octave (v7.1.0). Для реализации использовались библиотеки matplotlib (v3.5.1), scipy (v1.8.0) и пакет Octave interval.
Отчет подготовлен с использованием TeXstudio и TeX Live.

4 Результаты

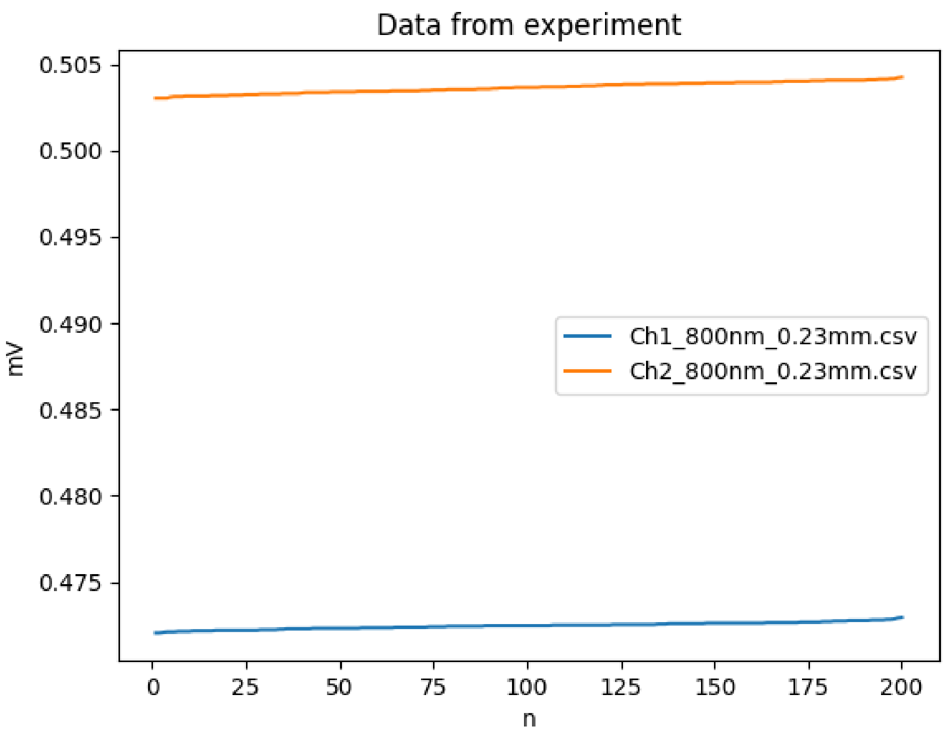

Рис. 2: Исходные выборки

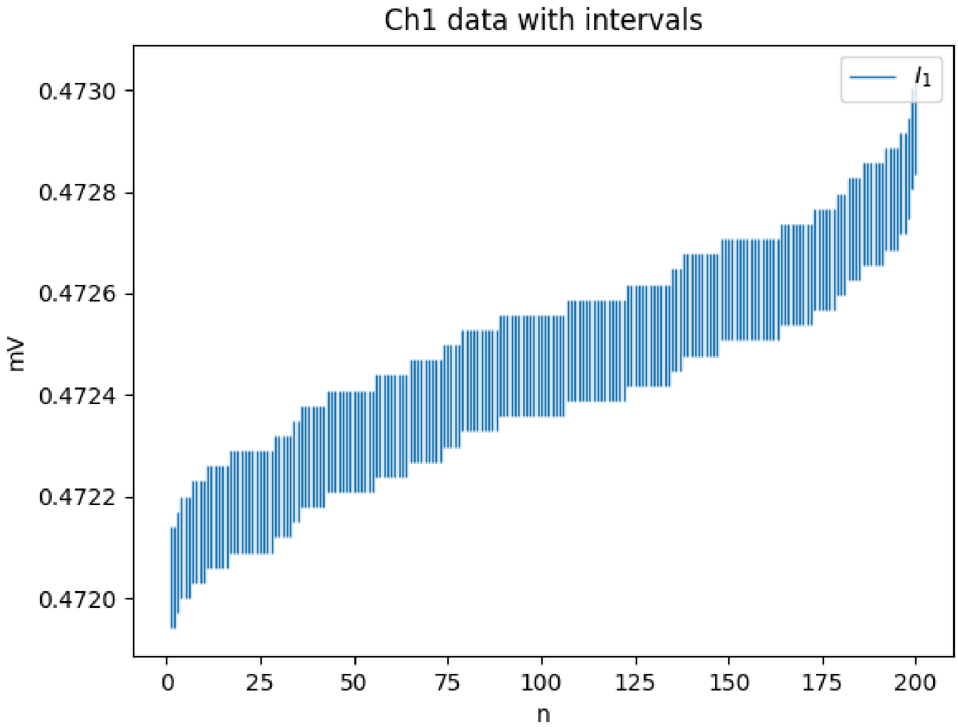

Рис. 3: "Обинтерваленные"значения первой выборки

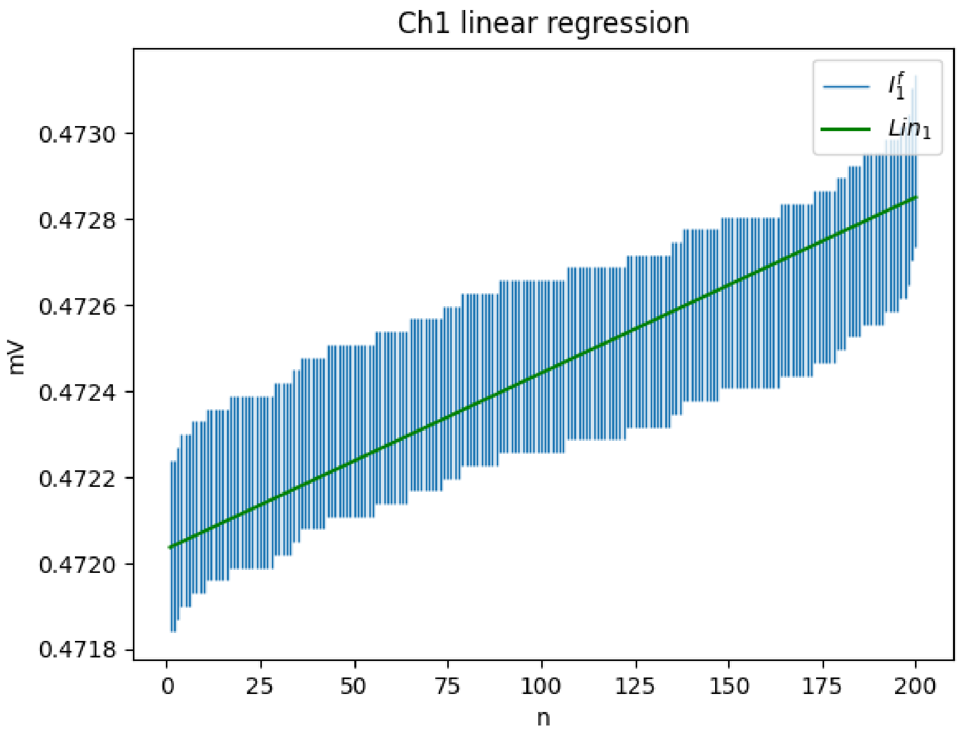

Рис. 4: Линейная регрессия для первой выборки

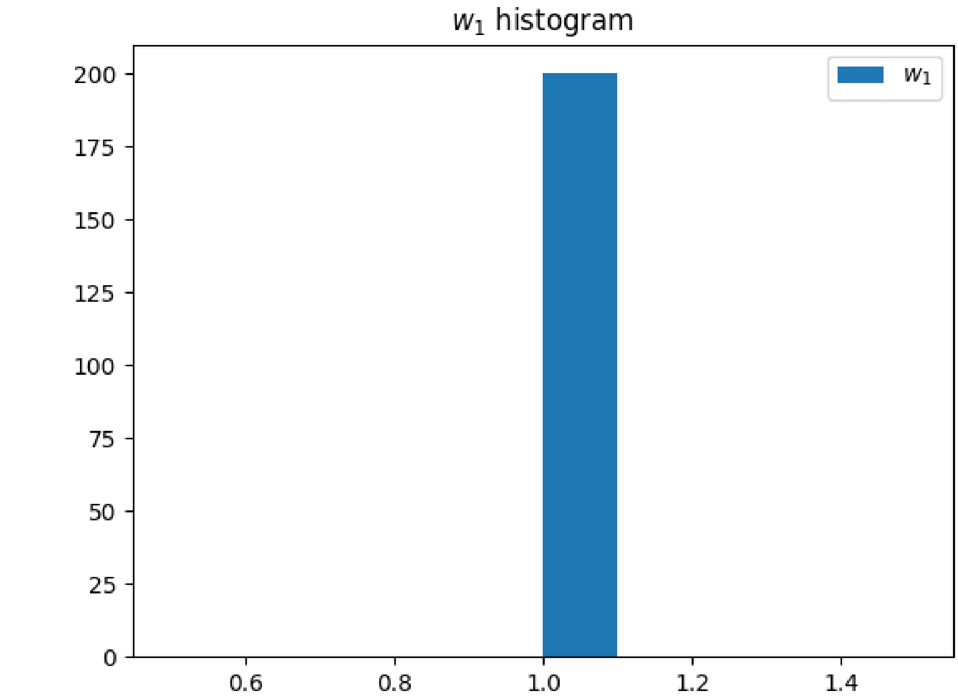

Рис. 5: Гистограмма коэффициентов коррекции для первой выборки

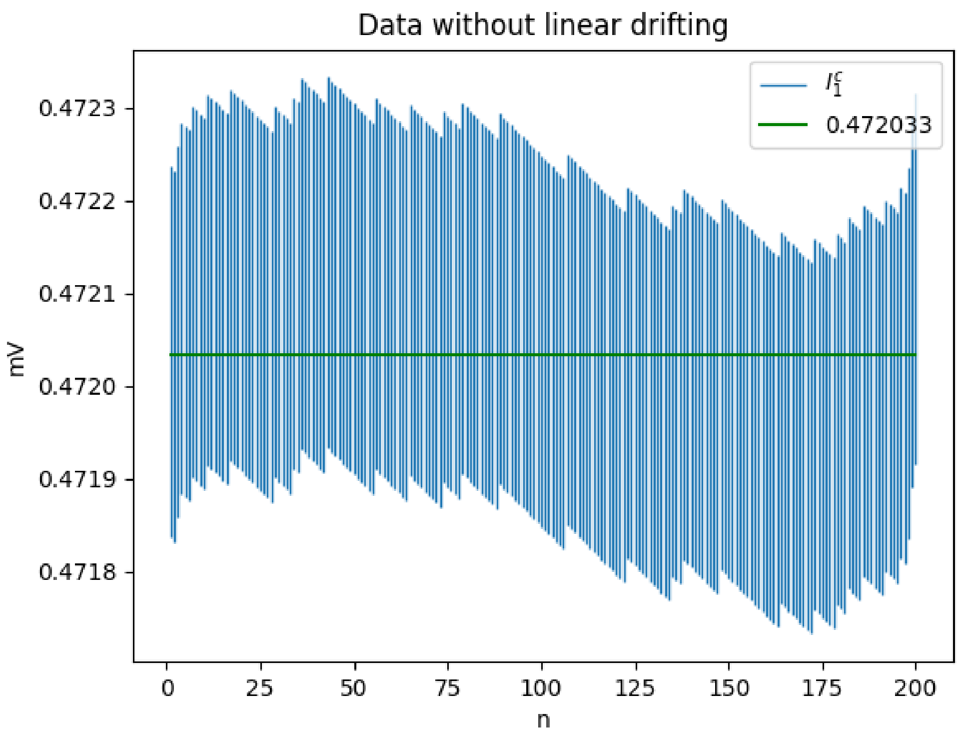

Рис. 6: "Спрямленные"значения первой выборки

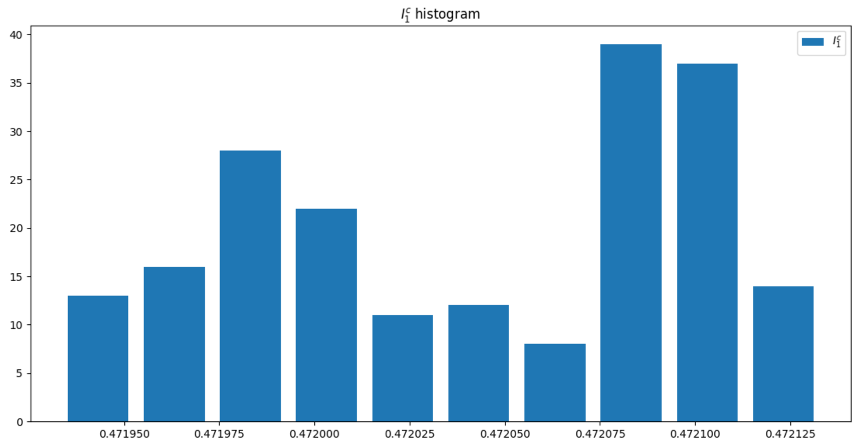

Рис. 7: Гистограмма для скорректированной модели данных первой выборки

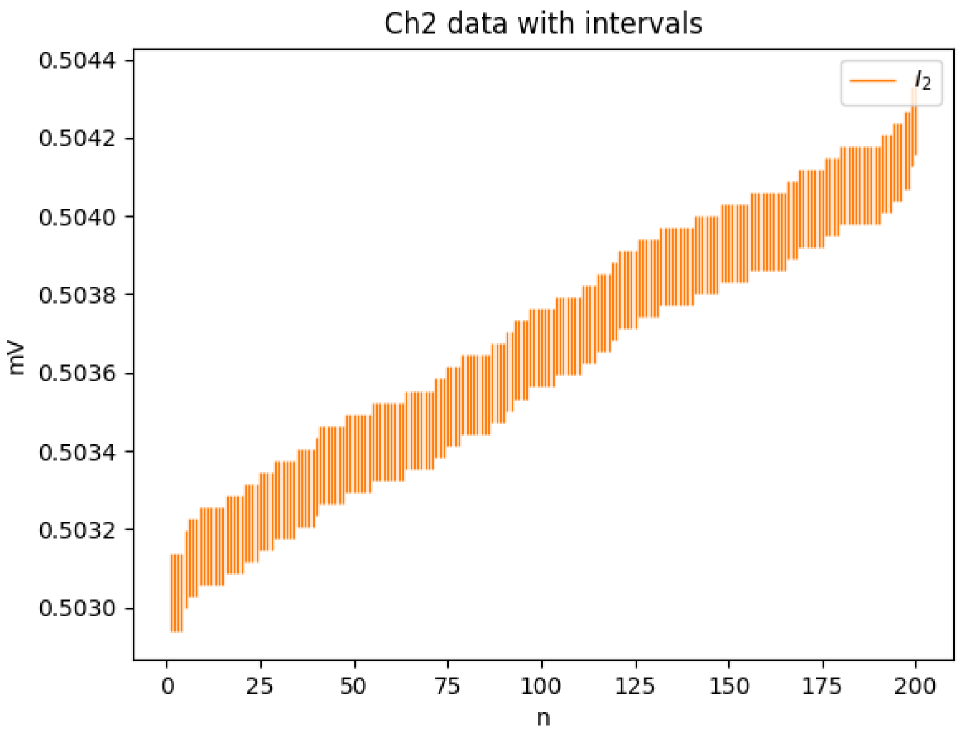

Рис. 8: "Обинтерваленные"значения второй выборки

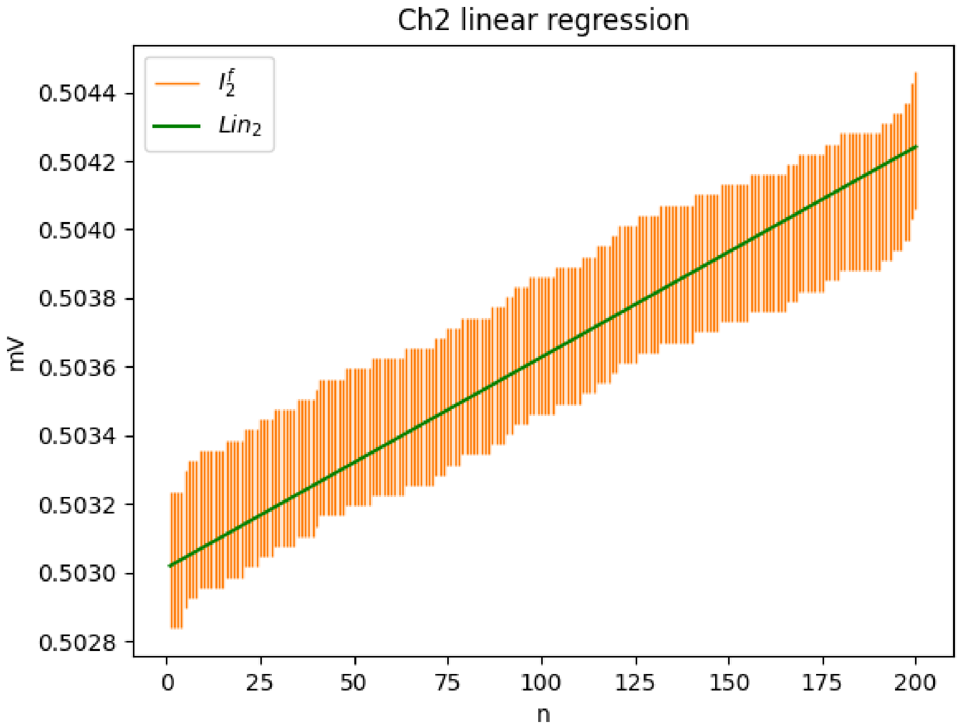

Рис. 9: Линейная регрессия для второй выборки

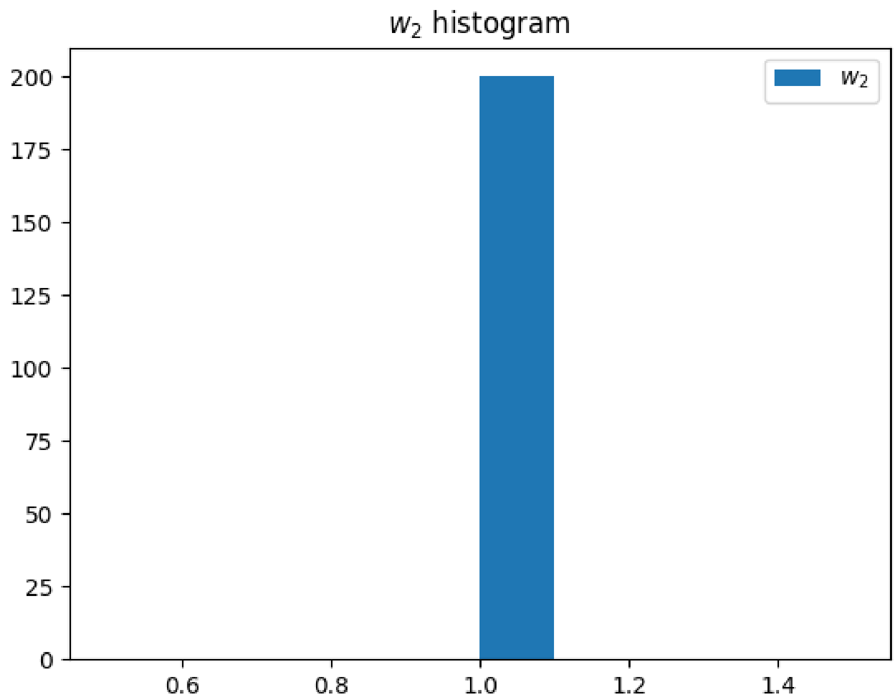

Рис. 10: Гистограмма коэффициентов коррекции для второй выборки

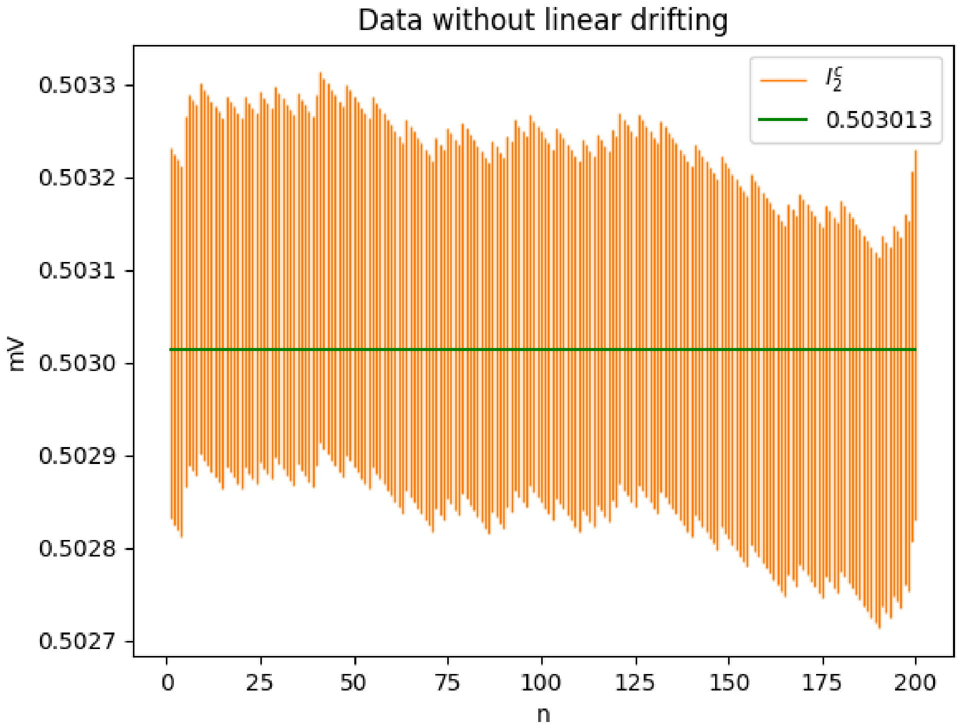

Рис. 11: "Спрямленные"значения второй выборки

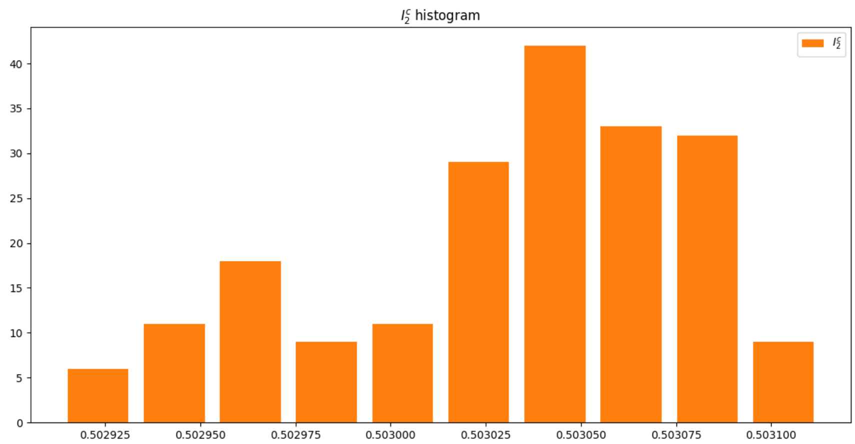

Рис. 12: Гистограмма для скорректированной модели данных второй выборки

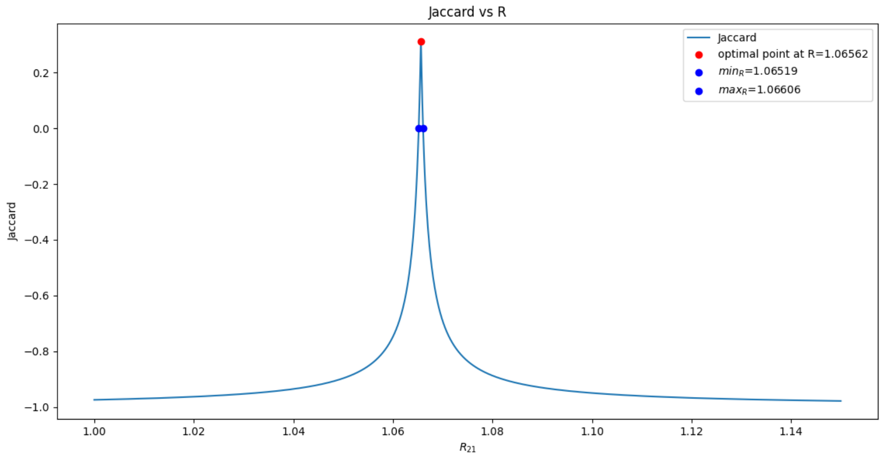

Рис. 13: Коэффициент Жаккара от калибровочного коэффициента

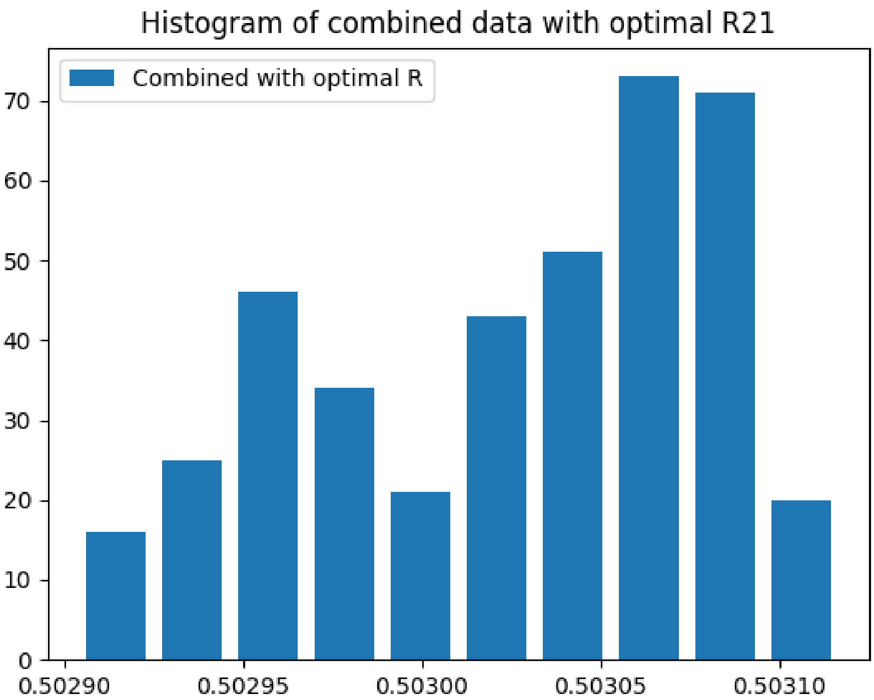

Рис. 14: Гистограмма объединенной выборки

5 Обсуждение

5.1 Модель дрейфа

Из гистограмм для w_1 и w_2 видно, что данным не потребовалась коррекция. Это означает, что выбор линейной модели дрейфа данных является разумным приближением.

5.2 Гистограммы скорректированных моделей и объединенной выборки

Характерной особенностью обеих выборок является наличие двух "пиков"на гистограммах для их скорректированных моделей, причем правый пик в обоих случаях выше. Эта особенность перенеслась и на совмещенную выборку, при этом 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 $R_{21} = 1.06562$, при этом в качестве интревала для R_{21}, исходя из замечания сделанного в пункте 2.6, можно рассматривать [1.06519,1.06606].

6 Ссылка на репозиторий

Репозиторий с исходным кодом: `https://github.com/gchevykalov/MathStat`.

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.

[3] С.И. Жилин. Примеры анализа интервальных данных в Octave. URL: https://github.com/szhilin/octave-interval-examples.