Санкт-Петербургский политехнический университет
Высшая школа прикладной математики и вычислительной физики, ФизМех

Направление подготовки
«01.03.02 Прикладная математика и информатика»

Отчет по лабораторной работе № 3
**дисциплина "Математическая статистика"**

Выполнил студент гр. 5030102/00201 Соболев Д.В.
Преподаватель: Баженов А.Н.

Санкт-Петербург

**2023**

# Содержание

# Список иллюстраций

# 1 Постановка задачи

Предметная область относится к физике полупроводников — исследованиям фотоэлектрических характеристик испытываемого датчика, проводимым специалистами лаборатории фотоэлектрических преобразователей Физико-технического института им. А. Ф. Иоффе [23]. Развёрнутое описание задачи дано в статье [5].

**Данные выборки.** Имеется выборка данных $X_1$ с интервальной неопределённостью. Число отсчётов в выборке равно 200.

На Рис. 2.1 представлены сырые данные с прибора [23].

Для дальнейшей работы используется модель данных с уравновешенным интервалом погрешности.

$$x = \mathring{x} + \varepsilon, \varepsilon = [-\varepsilon, \varepsilon], \quad \varepsilon > 0, \tag{1}$$

Здесь $\mathring{x}$ — данные прибора, $\varepsilon = 10^{-4}$ — погрешность прибора [23].

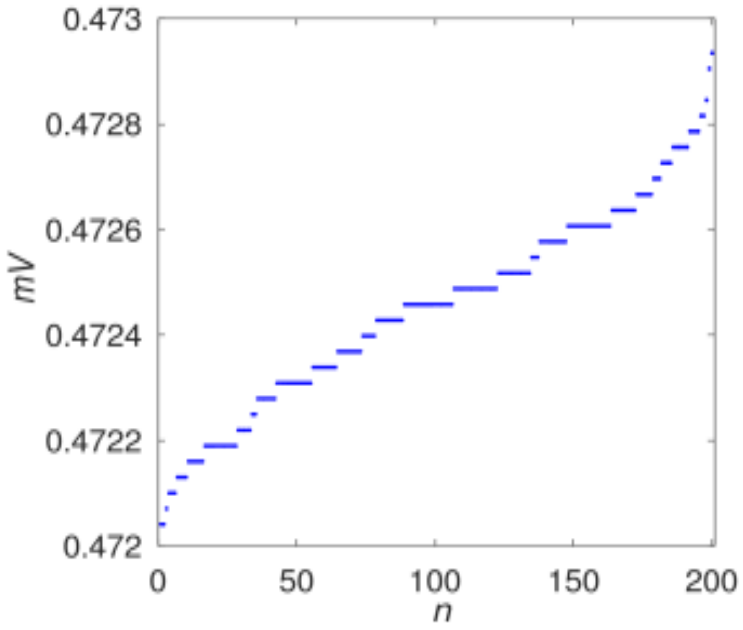


Рис. 1: Данные выборки $X_1$ [23].

**Диаграмма рассеяния.** Привести диаграмму рассеяния выборки с учётом погрешности прибора.

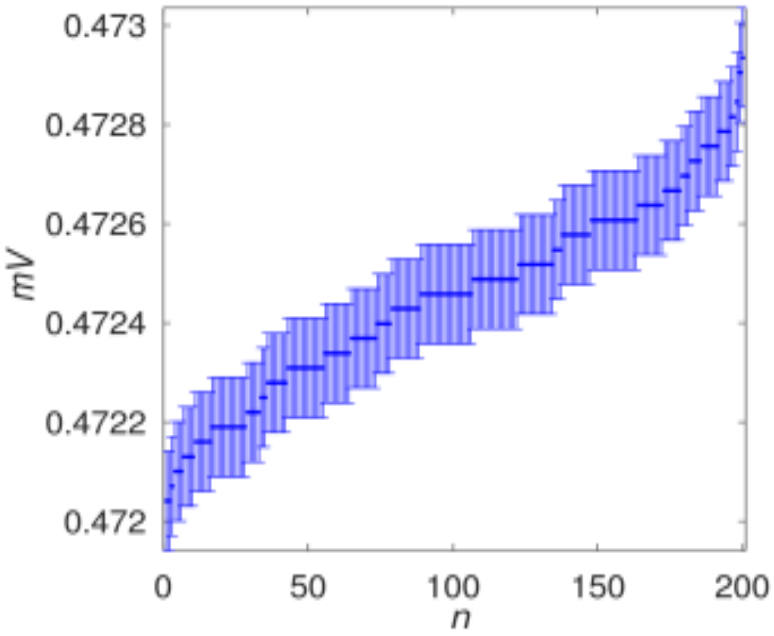


Рис. 2: Диаграмма рассеяния выборки $X_1$ с уравновешенным интервалом погрешности (1).

**Оценки исходной выборки.** Вычислим базовые оценки исходной выборки. Эти выборки несовместны и их информационные множества пусты. Внешние оценки найдем как

$$\underline{J} = \min_{x \le k \le n} \underline{x_k}, \ \overline{J} = \max_{x \le k \le n} \overline{x_k}. \tag{2}$$

Вычисления дают следующие результаты:

$$J_1 = 0.47194, wid\, J_1 = 0.47304. \tag{3}$$

Верхние и нижние вершины оценок $J_1$ совпадают с границами отображения на рис. 2.

## 2 Теория

Оценим выборку с помощью набора мер совместности [4].

Сначала значения мер совместности качества возьмём в исходном, ненормированном виде:

1. Размер максимальной клики $\max \mu_j$
2. Величина коэффициента вариабельности по Оскорбину $k_O$
3. Мера совместности Жаккара $J_i$

Вычисление моды выборки и максимальной клики. Имеет смысл распространить понятие моды на обработку интервальных данных, где она будет обозначать интервал тех значений, которые наиболее часты, т. е. встречаются в интервалах обрабатываемых данных наиболее часто. Фактически, это означает, что точки из моды интервальной выборки накрываются наибольшим числом интервалов этой выборки. Ясно, что по самому своему определению понятие моды имеет наибольшее значение (и наибольший смысл) лишь для накрывающих выборок. Иначе, если выборка ненакрывающая, то смысл «частоты» тех или иных значений в пределах рассматриваемых интервалов этой выборки в значительной мере теряется, хотя и не обесценивается.

Мода является пересечением интервалов максимальной совместной подвыборки, и если максимальных подвыборок имеется более одной, то мода будет объединением их пересечений, т. е. мультиинтервалом. Простой алгоритм вычисления моды интервальной выборки можно найти в [6]. Псевдокод специализированного алгоритма для нахождения моды выборки интервальных измерений и её частоты приведён в Табл. 1.

Ключевым в алгоритме Табл. 1 является формирование множества элементарных подинтервалов измерений из упорядоченных вершин (концов интервалов) $\underline{x_1}, \overline{x_1}, \underline{x_2}, \overline{x_2}, ..., \underline{x_n}, \overline{x_n}$ исходной выборки $X$.

Отметим также, что мода интервальной выборки — это интервал или мультиинтервал, который не обязан совпадать с каким-либо из интервалов обрабатываемой выборки.

**Пример 2.1.** Рассмотрим пример вычисления моды интервальной выборки из 4 элементов

$$X = \{[1, 4], [5, 9], [1.5, 4.5], [6, 9]\}. \tag{4}$$

В соответствии с алгоритмом Табл. 1, проверим совместность X. Пересечение элементов выборки пусто

$$I = \bigcap_{i=0}^{n} x_i = \varnothing.$$

Таким образом, необходимо выполнить шаги алгоритма Табл. 1 после ключевого слова ELSE.

Сформируем массив интервалов $z$ из концов интервалов X

$$z = \{[1, 1.5], [1.5, 4], [4, 4.5], [4.5, 5], [5, 6], [6, 9], [9, 9]\}. \tag{5}$$

Для каждого интервала $z_i$ подсчитываем число $\mu_i$ интервалов из выборки $X$, включающих $z_i$, получаем массив значений $\mu_i$ в виде

$$\{1, 2, 1, 0, 1, 2, 2\}. \tag{6}$$

Максимальные $\mu_i$, равные $\max \mu = 2$, достигаются для индексного множества

$$K = \{2, 6, 7\},$$

Как итог, мода является мультиинтервалом

$$modeX = \bigcup_{k \in K} z_k = [1.5, 4] \cup [6, 9] \cup [9, 9] = [1.5, 4] \cup [6, 9]. \tag{7}$$

Перейдём к практическому примеру. Проведём вычисление моды выборки $X_1$ по алгоритму Табл. 1.

Таблица 1: Алгоритм для нахождения моды интервальной выборки

**Вход**

Интервальная выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ длины $n$.

**Выход**

Мода $\operatorname{mode}\boldsymbol{X}$ выборки $\boldsymbol{X}$ и её частота $\mu$.

**Алгоритм**

$\boldsymbol{I} \leftarrow \bigcap_{i=1}^{n} \boldsymbol{x}_i$ ;

IF $\boldsymbol{I} \neq \varnothing$ THEN

  $\operatorname{mode}\boldsymbol{X} \leftarrow \boldsymbol{I}$ ;

  $\mu \leftarrow n$

ELSE

  помещаем все концы $\underline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_1$, $\underline{\boldsymbol{x}}_2$, $\overline{\boldsymbol{x}}_2$, …, $\underline{\boldsymbol{x}}_n$, $\overline{\boldsymbol{x}}_n$ интервалов рассматриваемой выборки $\boldsymbol{X}$ в один массив $Y = (y_1, y_2, \ldots, y_{2n})$ ;

  упорядочиваем элементы в $Y$ по возрастанию значений;

  порождаем интервалы $\boldsymbol{z}_i = [y_i, y_{i+1}]$, $i = 1, 2, \ldots, 2n-1$ (назовём их *элементарными подинтервалами измерений*);

  для каждого $\boldsymbol{z}_i$ подсчитываем число $\mu_i$ интервалов из выборки $\boldsymbol{X}$, включающих интервал $\boldsymbol{z}_i$ ;

  вычисляем $\mu \leftarrow \max\limits_{1 \le i \le 2n-1} \mu_i$ ;

  выбираем номера $k$ интервалов $\boldsymbol{z}_k$, для которых $\mu_k$ равно максимальному, т. е. $\mu_k = \mu$, и формируем из таких $k$ множество $K = \{k\} \subseteq \{1, 2, \ldots, 2n-1\}$ ;

  $\operatorname{mode}\boldsymbol{X} \leftarrow \bigcup_{k \in K} \boldsymbol{z}_k$

END IF

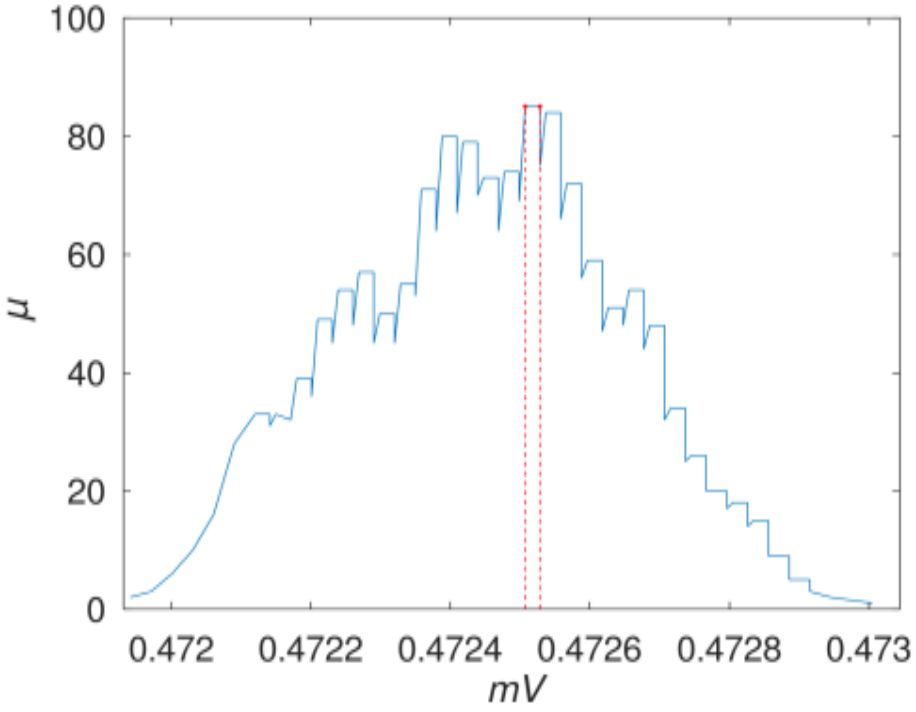


Рис. 3: График частот при вычислении моды выборки $X_1$

По результатам вычисления моды выборки находим размер максимальной клики $\max \mu_j$:

$$\max \mu_j(X_1) = 85. \tag{8}$$

Индексы таких элементов образуют множество $K$, а из них образуется мода

$$K = \{79, 80, ..., 163\}, \tag{9}$$

$$mode(X_1) = \bigcup_{k \in K} z_k = [0.47251, 0.47253]. \tag{10}$$

На рис. 4 показаны элементы выборки $X_1$, в которые входит мода.

**Варьирование неопределённости измерений.** Один из приёмов выявления достижения совместности выборки интервальных наблюдений основан на представлении о причине несовместности как недооценённой величины неопределённости [27, 28]. Закономерным шагом в этом случае становится поиск некоторой минимальной коррекции величин неопределённости интервальных наблюдений, необходимой для обеспечения совместности задачи построе-

ния зависимости. Если величину коррекции каждого интервального наблюдения $y_i = [\overset{\circ}{y}_i - \epsilon_i, \overset{\circ}{y}_i + \epsilon_i] = (\overset{\circ}{y}_i, \epsilon_i)$ выборки $S_n$ выражать коэффициентом его уширения $\omega_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде

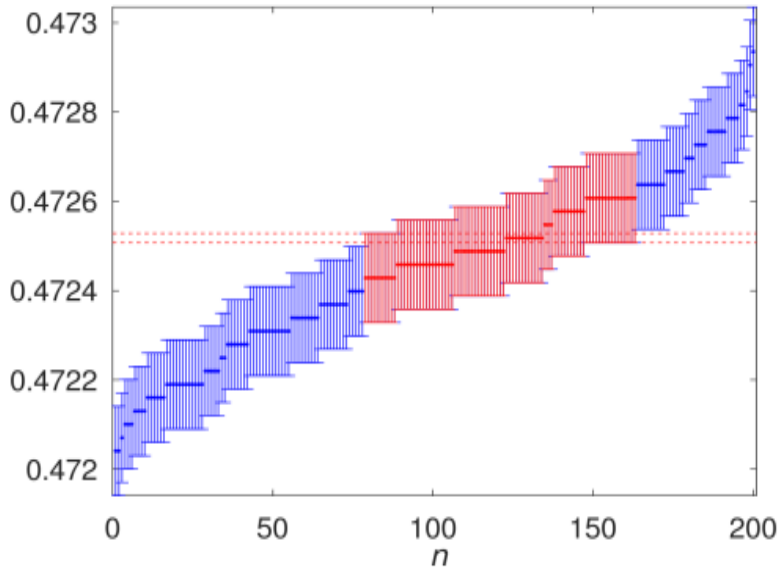


Рис. 4: Элементы выборки $X_1$, в которые входит мода (3).

вектора коэффициентов $\omega^* = (\omega_1^*, ..., \omega_n^*)$, необходимая для совместности задачи построения зависимости$y = f(x, \beta)$ может быть найдена решением задачи условной оптимизации

$$\min_{\omega,\beta} \sum_{i=1}^{n} \omega_i \tag{11}$$

при ограничениях

$$\begin{cases} \overset{\circ}{y}_i - \omega_i \epsilon_i \leq f(x_i, \beta) \leq \overset{\circ}{y}_i + \omega_i \epsilon_i, \\ \omega_i \geq 1, \end{cases} \quad i = 1, ..., n. \tag{12}$$

Результирующие значения коэффициентов $\omega_i^*$, строго превосходящие единицу, указывают на наблюдения, которые требуют уширения интервалов неопределённости для обеспечения совместности данных и модели. Именно такие наблюдения заслуживают внимания при анализе данных на выбросы. Значительное количество подобных наблюдений может говорить либо о неверно выбранной

структуре зависимости, либо о том, что величины неопределённости измерений занижены во многих наблюдениях (например, в результате неверной оценки точности измерительного прибора).

Следует отметить значительную гибкость языка неравенств. Он даёт возможность переформулировать и расширять систему ограничений (12) для учёта специфики данных и задачи при поиске допустимой коррекции данных, приводящей к разрешению исходной несовместности. Например, если имеются основания считать, что величина неопределённости некоторой группы наблюдений одинакова и при коррекции должна увеличиваться синхронно, то система ограничений (12) может быть пополнена равенствами вида

$$\omega_{i_1} = \omega_{i_2} = ... = \omega_{i_K}$$

где $i_1, ..., i_K$ — номера наблюдений группы. В случае, когда в надёжности каких-либо наблюдений исследователь уверен полностью, при решении задачи (11)–(12) соответствующие им величины $\omega_i$ можно положить равными единице, т.е. запретить варьировать их неопределённость.

Задача поиска коэффициентов масштабирования величины неопределённости (11)–(12) сформулирована для распространённого случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределённости измерений $y_i$ при сохранении базовых значений интервалов $\overset{\circ}{y}_i$ неподвижными. При необходимости постановка задачи легко обобщается. Например, если интервалы наблюдений не уравновешены относительно базовых значений (то есть $y_i = [\overset{\circ}{y}_i - \epsilon_i^-, \overset{\circ}{y}_i + \epsilon_i^+]$ $\epsilon_i^- \neq \epsilon_i^+$, то границы интервальных измерений можно варьировать независимо, масштабируя величины неопределённости $\epsilon_i^-$ и $\epsilon_i^+$ с помощью отдельных коэффициентов $\omega_i^-$ и $\omega_i^+$:

$$\min_{\omega^-, \omega^+, \beta} \sum_{i=1}^{n} (\omega^- + \omega^+) \tag{13}$$

при ограничениях

$$\begin{cases} \overset{\circ}{y}_i - \omega_i^- \epsilon_i^- \leq f(x_i, \beta) \leq \overset{\circ}{y}_i + \omega_i^+ \epsilon_i^+, \\ \omega_i^- \geq 1, \\ \omega_i^+ \geq 1, \end{cases} \qquad i = 1, ..., n. \tag{14}$$

Для линейной по параметрам $\beta$ зависимости $y = f(x, \beta)$ задача (11)–(12) представляет собою задачу линейного программирования, для решения которой широко доступны хорошие и апробированные программы в составе библиотек на различных языках программирования, в виде стандартных процедур систем

компьютерной математики, а также в виде интерактивных подсистем электронных таблиц.

**Оптимизация по Оскорбину.** Перейдём к практическому примеру выборки $X_1$. Поставим задачу линейного программирования (13) — (14) в простейшем виде

$$\min_{\omega,\beta} \omega \tag{15}$$

при ограничениях

$$\begin{cases} mid\, x_i - \omega\epsilon_i \le \beta \le mid\, x_i + \omega\epsilon_i, \\ \omega_i \ge 1, \end{cases} \quad i = 1, ..., n. \tag{16}$$

Проведём вычисление моды выборки $X_1$ с использованием программ С.И.Жилина [8]. Синтаксис вызова программы

$$[oskorbin\_center\_k, w] = estimate\_uncertainty\_center(X1) \tag{17}$$

Вычисления дают следующие результаты

$$oskorbin\_center\_k = 0.4725, \tag{18}$$

$$w = 4.4705. \tag{19}$$

Оценка постоянной (18) очень близка с вычисленной ранее модой. Величина однородного расширения интервалов (19) достаточно велика, что соответствует весьма большой степени несовместности выборки $X_1$.

На рис. 5 приведена диаграмма рассеяния выборки $X_1$ с увеличенным в w раз интервалом неопределённости.

Красной пунктироной линией показана оценка постоянной (18).

**Индекс Жаккара.** Для описания выборок, помимо оценок их размеров, желательно иметь дополнительную информацию о мере сходства элементов выборки. В различных областях анализа данных, биологии, информатике, в науках о Земле часто используют различные меры сходства множеств (см. [22]). Меру сходства между объектами $A$ и $B$ можно определить как двухместную вещественнозначную функцию $S(A, B)$, которая обладает следующими свойствами:

— ограниченность:$0 \le S(A, B) \le 1$;

— симметричность: $S(A, B) = S(B, F)$;

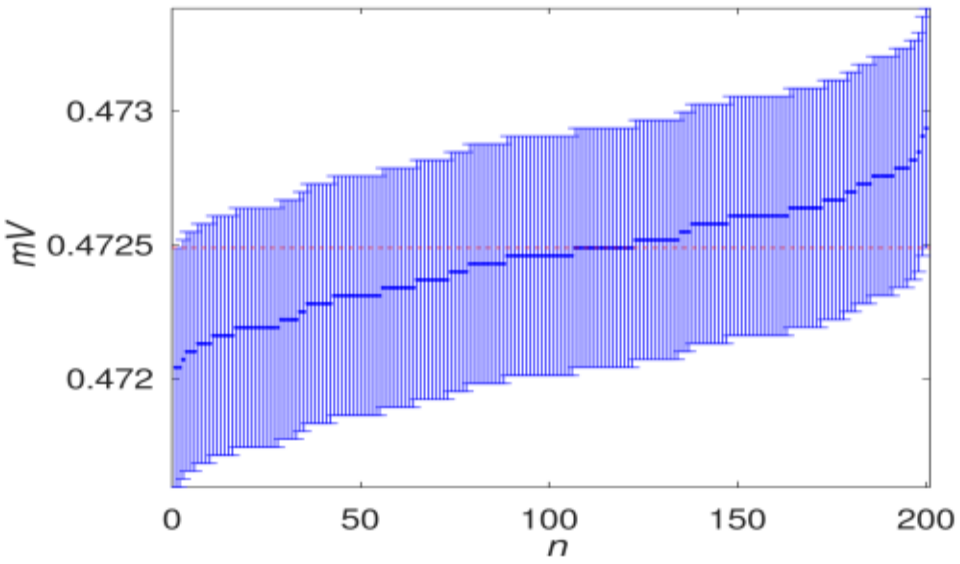


Рис. 5: Диаграмма рассеяния выборки $X_1$ с увеличенным в w раз интервалом неопределённости.

— неразличимость совпадающих элементов: $S(A,B) = 1 \Leftrightarrow A = B$;
— монотонность: $A \subseteq B \subseteq C \Rightarrow S(A,B) \geq S(A,C)$.

Таким образом, значение 1 этой меры соответствует совпадению множеств $A$ и $B$, а значение 0 означает их полное несходство. Отметим, что существуют и иные системы аксиом сходства. В компьютерных приложениях (обработка изображений, машинное обучение) меру сходства множеств обозначают как $I_oU$ (Intersection over Union). В математике и её приложениях для подобных конструкций часто используется термин индекс Жаккара, по имени исследователя, впервые предложившего эту меру.

В процессе развития интервального анализа были введены различные определения и конструкции оценки меры совместности интервальных объектов. Вместе с тем в практике обработки данных часто необходимо оперировать относительными величинами. В частности, это нужно в связи с необходимостью сопоставления допусков и размеров деталей, погрешности измерителей и значений измеряемых величин и т. п. [?].

Представим обобщение меры Жаккара на выборки интервалов [5]. В качестве числовой характеристики степени совпадения двух интервалов $x, y$ рассмотрим

величину

$$Ji := \frac{wid\,(\mathbf{x} \wedge \mathbf{y})}{wid\,(\mathbf{x} \vee \mathbf{y})} \tag{20}$$

В выражении (20) используется ширина интервала (см. стр. ??), а вместо операций пересечения и объединения множеств — операции взятия точной нижней грани ı∧ȷ (инфимума, см. (??)) и точной верхней грани ı∨ȷ (супремума, см. (??)) относительно включения для двух величин в полной интервальной арифметике Каухера. В обозначении $Ji(\mathbf{x}, \mathbf{y})$ буква J указывает на фамилию «Jaccard», а i — на интервальность его применения. В общем случае инфимум по включению в числителе выражения (20) может быть неправильным интервалом, и его ширина тогда отрицательна.

Рассмотренная мера обобщает обычное понятие меры совместности на различные типы взаимной совместности интервалов. Если пересечение интервалов $\mathbf{x}, \mathbf{y}$ пусто, т. е. $\mathbf{x} \cap \mathbf{y} = \varnothing$, то $\mathbf{x} \wedge \mathbf{y}$ — неправильный интервал и числитель формулы (20) имеет отрицательное значение. В предельном случае несовпадающих вещественных вырожденных интервалов $\mathbf{x} = x$ и $\mathbf{y} = y$, $x \neq y$, имеем

$$Ji(x, y) = -1.$$

В целом получаем

$$-1 \leq Ji(\mathbf{x}, \mathbf{y}) \leq 1. \tag{21}$$

Таким образом, величина $Ji$ непрерывно описывает ситуации от полной несовместности вещественных значений $x \neq y$ до полного перекрытия интервалов $\mathbf{x} = \mathbf{y}$. Следует заметить, что в отличие от случая вещественных величин, для которых индекс Жаккара может принимать только два значения, 0 и 1, формула (20) даёт характеризацию различных отношений сходства интервалов с помощью непрерывного ряда значений между $-1$ и 1.

Мера совместности, введённая для двух интервалов в форме (20), допускает естественное обобщение на случай интервальной выборки $\mathbf{X} = \{x_i\}, i = 1, 2, ..., n$. Определим меру $Ji(\mathbf{X})$ для этой выборки как

$$Ji(\mathbf{X}) = \frac{wid\,(\bigwedge\limits_i x_i)}{wid\,(\bigvee\limits_i x_i)} \tag{22}$$

Видно, что выражение (22) переходит в случае интервальной выборки из двух элементов в выражение (20).

В связи несовместностью выборки будем использовать следующую меру, которая имеет место и в случае несовместных выборок.

$$\rho(mode(\mathbf{X})) = \frac{wid(mode(\mathbf{X}))}{wid\,(\bigvee\limits_i x_i)} \tag{23}$$

Назовём конструкцию (23) относительная ширина моды. В отличие от минимума по включению, мода выборки всегда является правильным интервалом. В целом получаем

$$0 \leq \rho(mode(\mathbf{X})) \leq 1. \tag{24}$$

**Вычисление меры совместности.** Перейдём к практическому примеру выборки $\mathbf{X}_1$. Проведём вычисление с использованием программ на ресурсе [23].

$$Ji(\mathbf{X}) = \frac{wid\,(\bigwedge\limits_i x_i)}{wid\,(\bigvee\limits_i x_i)} = -0.6344. \tag{25}$$

Отрицательность меры (25) соответствует несовместности выборки $\mathbf{X}_1$, а её модуль — высокой степени этой несовместности.

Относительная ширина моды (23) равна

$$\rho(mode(\mathbf{X})) = \frac{wid(mode(\mathbf{X}))}{wid\,(\bigvee\limits_i x_i)} = 0.039. \tag{23}$$

Величина (26) составляет менее 4% внешней оценки выборки $\mathbf{X}_1$.

# 3 Результаты

## 3.1 Данные выборки

Данные для выборки взяты из файла $Channel\_1\_400nm\_2mm.csv, \varepsilon = 10^{-4}$.

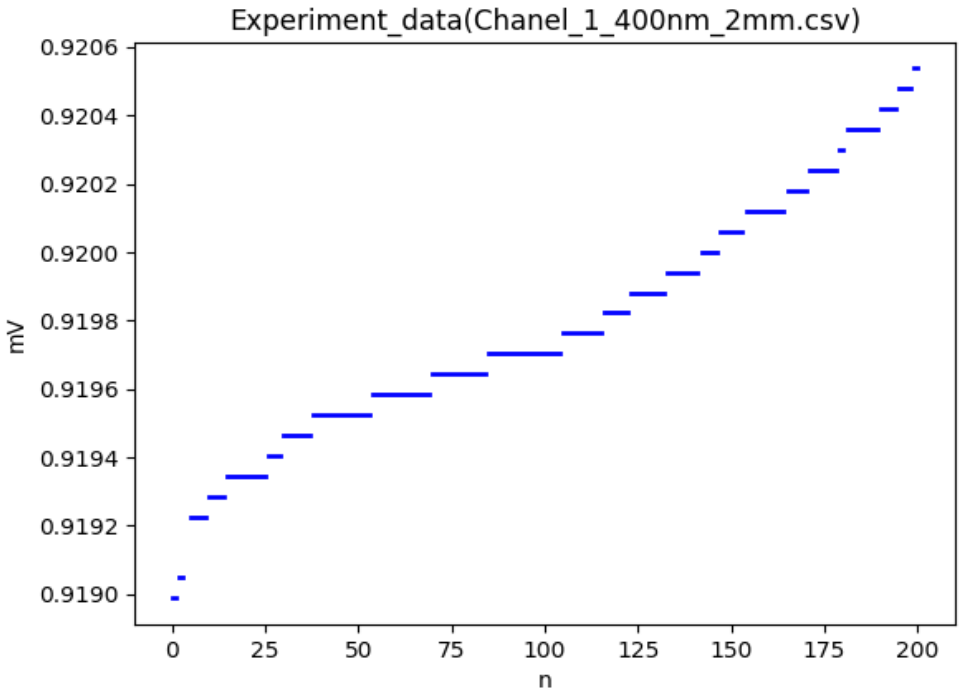


Рис. 6: Данные выборки $\mathbf{X}_1$

### 3.2 Диаграмма рассеяния

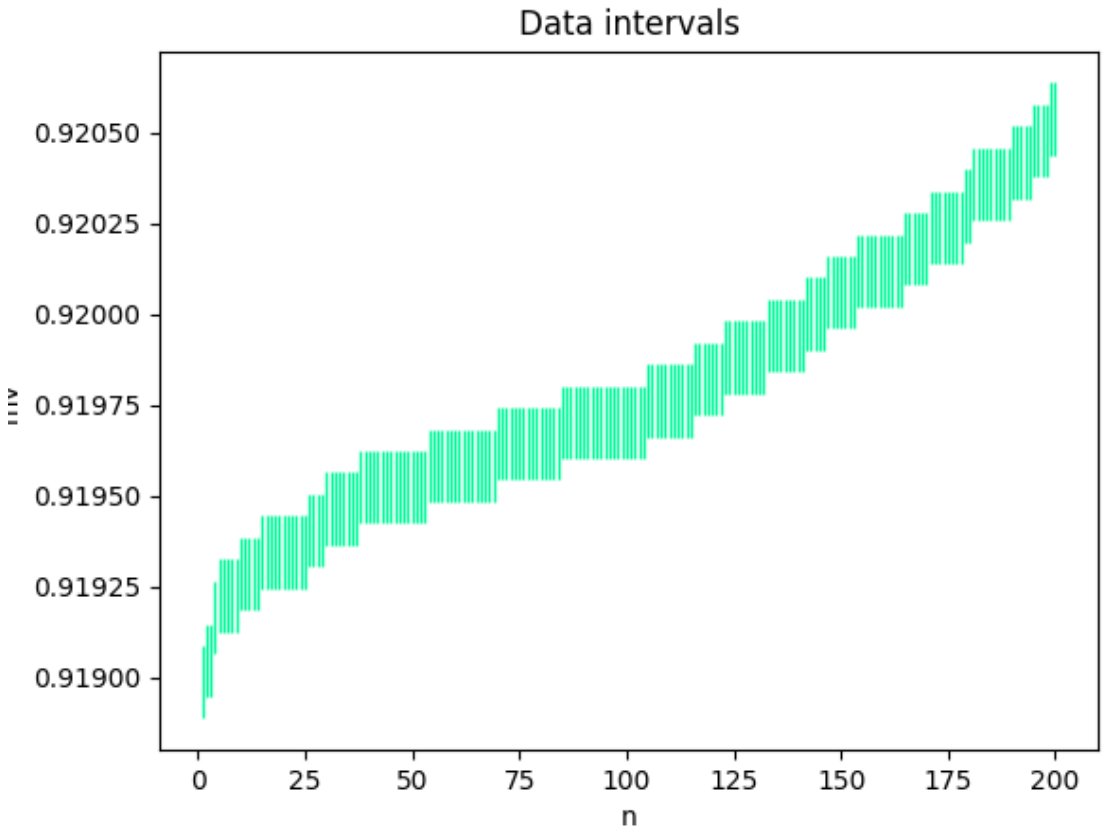


Рис. 7: Диаграмма рассеяния выборки $\mathbf{X}_1$ с уравновешенным интервалом погрешности (1).

### 3.3 Оценки исходной выборки

$$\underline{J} = \min_{x \le k \le n} \underline{x_k} = 0.9189881, \ \overline{J} = \max_{x \le k \le n} \overline{x_k} = 0.9205379.$$

$$mid\, J = 0.919763, \ rad\, J = 0.000775, \ wid\, J = 0.00155.$$

### 3.4 Вычисление моды выборки и максимальной клики

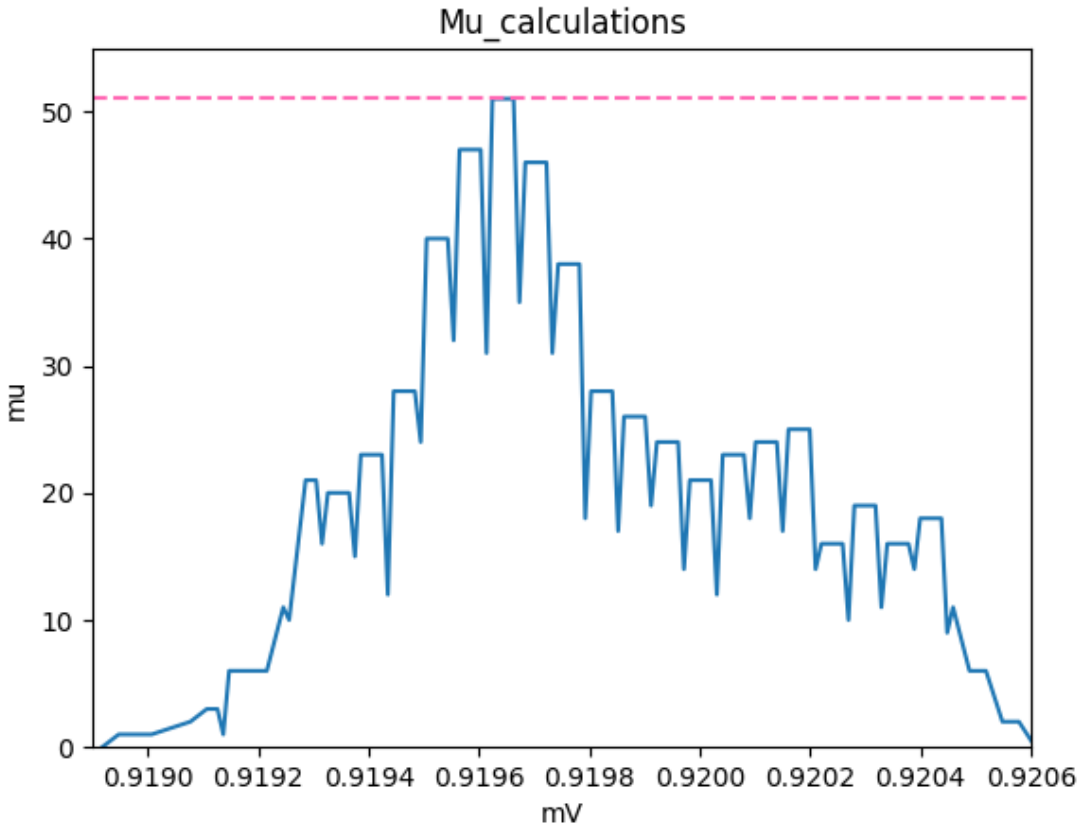


Рис. 8: График частот при вычислении моды выборки $\mathbf{X}_1$

По результам вычисления моды выборки находим размер максимальной клики $\max \mu_j$:

$$\max \mu_j(X_1) = 51.$$

Индексы таких элементов образуют множество $K$, а из них образуется мода

$$K = \{142, 143, ..., 167\},$$

$$mode(X_1) = \bigcup_{k \in K} z_k = [0.9196246, 0.919663].$$

На рис. 9 показаны элементы выборки $X_1$, в которые входит мода.

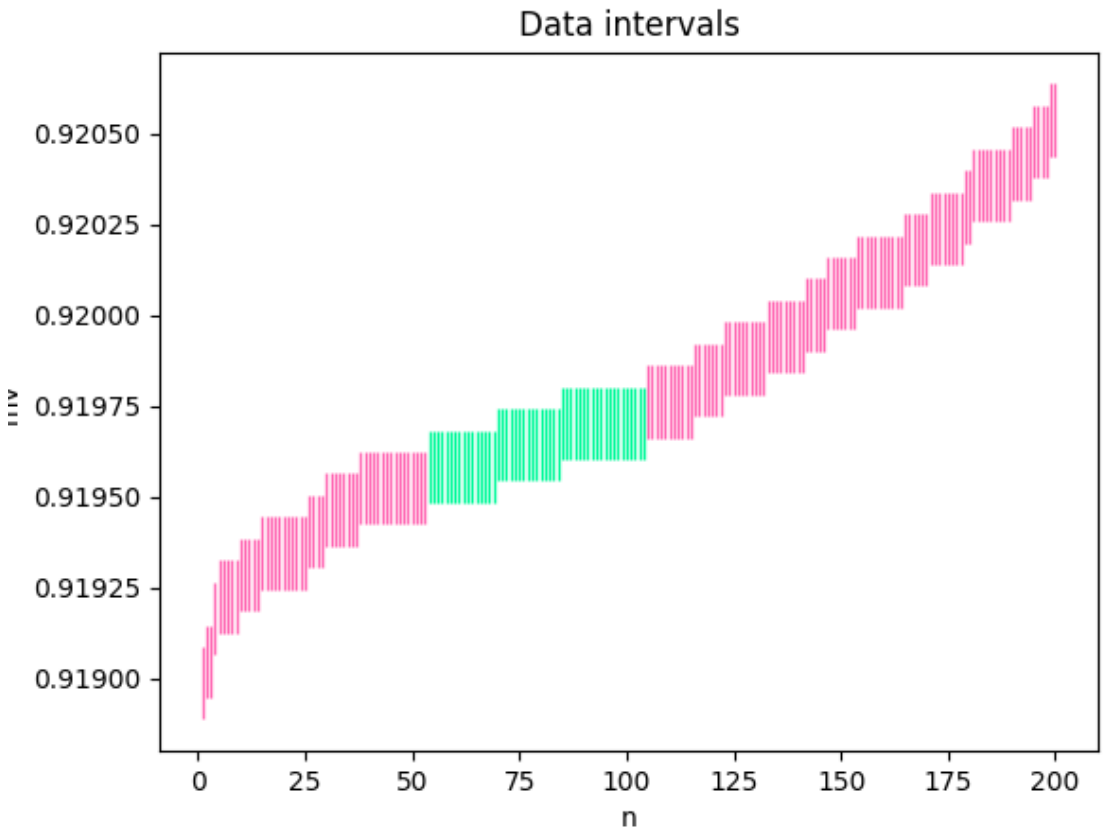


Рис. 9: Элементы выборки $\mathbf{X}_1$, в которые входит мода (3).

### 3.5 Оптимизация по Оскорбину

Вычисления дают следующие результаты

$$oskorbin\_center\_k = 0.919763,$$

$$w = 7.7490.$$

На рис. 10 приведена диаграмма рассеяния выборки $\mathbf{X}_1$ с увеличенным в w раз интервалом неопределённости.

Красной линией показана оценка постоянной $w$.

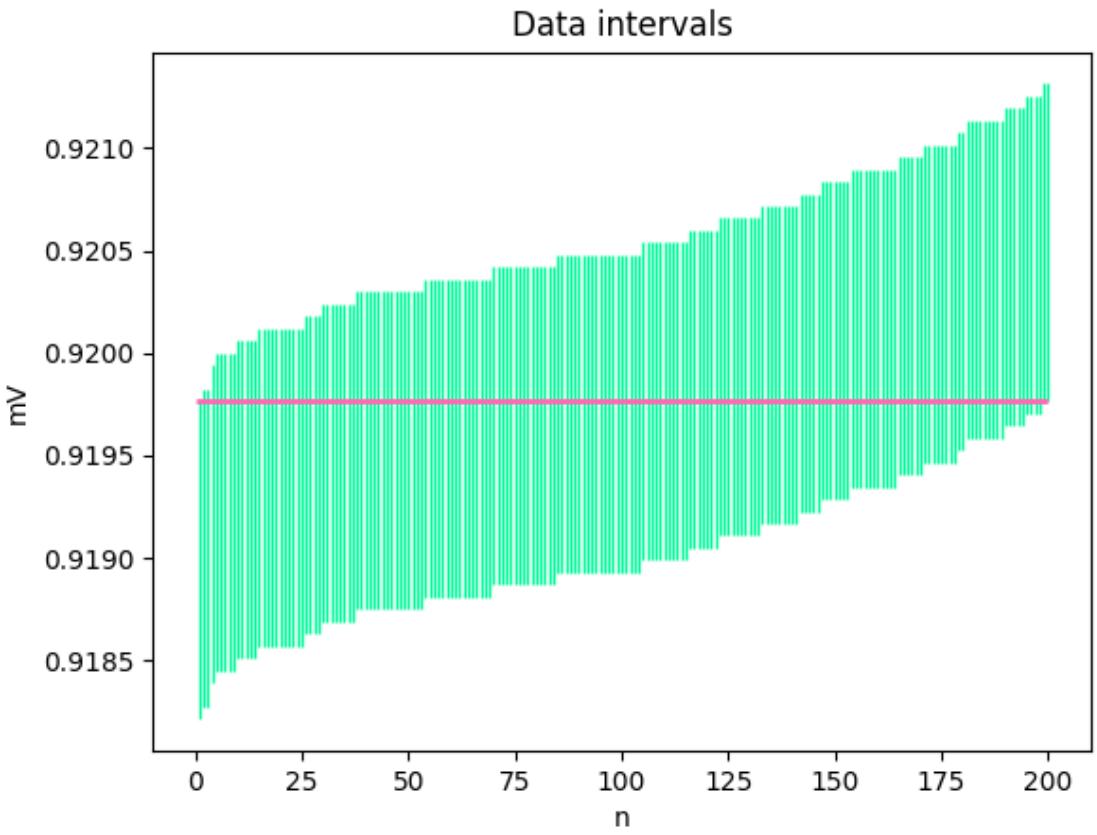


Рис. 10: Диаграмма рассеяния выборки $\mathbf{X}_1$ с увеличенным в $w$ раз интервалом неопределённости.

### 3.6 Вычисление меры совместности

$$Ji(\mathbf{X}) = \frac{wid\left(\bigwedge\limits_i x_i\right)}{wid\left(\bigvee\limits_i x_i\right)} = -0.77140.$$

Относительная ширина моды равна

$$\rho(mode(\mathbf{X})) = \frac{wid(mode(\mathbf{X}))}{wid\left(\bigvee\limits_i x_i\right)} = 0.11972.$$

# 4 Обсуждение

## 4.1 Оптимизация по Оскорбину

Оценка постоянной $w$ близка с вычисленной ранее модой. Величина однородного расширения интервалов достаточно велика, что соответствует весьма большой степени несовместности выборки $\mathbf{X}_1$.

## 4.2 Вычисление меры совместности

Отрицательность меры $Ji(\mathbf{X})$ соответствует несовместности выборки $\mathbf{X}_1$, а её модуль — высокой степени этой несовместности.

Абсолютное значение ширины моды зависит как от расстояния между интервалами мультивыборки, так и от ширины их максимума по включению. Величина ширины относительной моды составляет менее 12% внешней оценки выборки $\mathbf{X}_1$ .

# 5 Литература

Язык программирования *Python* 3.10
Подключенные библиотеки для *Python*: $numpy, math, seaborn, matplotlib, scipy, tabulate$
Ссылка на Github: `https://github.com/Parampaika/MatStat`