Санкт-Петербургский политехнический университет имени Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

Отчет
по лабораторной работе #1
по дисциплине
"Анализ данных с интервальной неопределённостью"

Выполнил:

Студент: Шварц Александр
Группа: 5040102/40201

Принял:

к. ф.-м. н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2026 г.

Содержание

1 Цель работы

Получить практические навыки вычисления интервальных описательных статистик (моды, медиан), работы с коэффициентом Жаккара и применения методов оптимизации для интервальных данных. Сравнить эффективность различных функционалов на основе интервальных статистик для оценивания параметров моделей.

2 Теоретические сведения

2.1 Интервальная арифметика

Интервалом называется замкнутое подмножество вещественной прямой:

$$\mathbf{x} = [\underline{x},\ \overline{x}] = \{x \in \mathbb{R} \mid \underline{x} \le x \le \overline{x}\}.$$

Середина и радиус интервала:

$$\operatorname{mid}\mathbf{x} = \frac{\underline{x} + \overline{x}}{2}, \qquad \operatorname{rad}\mathbf{x} = \frac{\overline{x} - \underline{x}}{2}.$$

Основные арифметические операции над интервалами $\mathbf{x} = [\underline{x}, \overline{x}]$ и $\mathbf{y} = [\underline{y}, \overline{y}]$:

$$\mathbf{x} + \mathbf{y} = [\underline{x} + \underline{y},\ \overline{x} + \overline{y}], \tag{1}$$

$$a + \mathbf{x} = [a + \underline{x},\ a + \overline{x}], \quad a \in \mathbb{R}, \tag{2}$$

$$t \cdot \mathbf{x} = \begin{cases} [t\,\underline{x},\ t\,\overline{x}], & t \ge 0, \\ [t\,\overline{x},\ t\,\underline{x}], & t < 0. \end{cases} \tag{3}$$

Пересечение двух интервалов:

$$\mathbf{x} \cap \mathbf{y} = [\max(\underline{x}, \underline{y}),\ \min(\overline{x}, \overline{y})],$$

определено при $\max(\underline{x}, \underline{y}) \le \min(\overline{x}, \overline{y})$.

2.2 Коэффициент Жаккара

Коэффициент Жаккара (Jaccard index) для двух интервалов определяется как отношение длины пересечения к длине объединения:

$$Ji(\mathbf{x}, \mathbf{y}) = \frac{\operatorname{wid}(\mathbf{x} \cap \mathbf{y})}{\operatorname{wid}(\mathbf{x} \cup \mathbf{y})},$$

где $\operatorname{wid}\mathbf{x} = \overline{x} - \underline{x}$ — ширина интервала, а объединение берётся как интервальная оболочка: $\mathbf{x} \cup \mathbf{y} = [\min(\underline{x}, \underline{y}),\ \max(\overline{x}, \overline{y})]$.

Для набора пар интервалов $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ коэффициент Жаккара обобщается как:

$$Ji = \frac{\sum_{i=1}^{n} \operatorname{wid}(\mathbf{x}_i \cap \mathbf{y}_i)}{\sum_{i=1}^{n} \operatorname{wid}(\mathbf{x}_i \cup \mathbf{y}_i)}.$$

2.3 Интервальная мода

Интервальная мода $\operatorname{mode}\mathbf{X}$ выборки интервалов $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ определяется как интервал максимальной глубины, т.е. точка на вещественной прямой, покрытая наибольшим числом интервалов из выборки. Формально:

$$\operatorname{mode}\mathbf{X} = \arg\max_{\mathbf{z}} \operatorname{depth}(\mathbf{z}, \mathbf{X}),$$

где $\operatorname{depth}(\mathbf{z}, \mathbf{X}) = \#\{i : \mathbf{z} \subseteq \mathbf{x}_i\}$. На практике мода вычисляется алгоритмом «заметающей прямой» (sweep line) за время $O(n \log n)$.

2.4 Интервальные медианы

Медиана Крейновича med_K определяется через медианы границ:

$$\operatorname{med}_K \mathbf{X} = [\operatorname{median}(\underline{x}_1, \ldots, \underline{x}_n),\ \operatorname{median}(\overline{x}_1, \ldots, \overline{x}_n)].$$

Медиана Пролубникова med_P определяется через медианы середин и радиусов:

$$\operatorname{med}_P \mathbf{X} = [m - r,\ m + r],$$

где $m = \operatorname{median}(\operatorname{mid}\mathbf{x}_1, \ldots, \operatorname{mid}\mathbf{x}_n)$, $r = \operatorname{median}(\operatorname{rad}\mathbf{x}_1, \ldots, \operatorname{rad}\mathbf{x}_n)$.

При одинаковых радиусах всех интервалов обе медианы совпадают: $\operatorname{med}_K = \operatorname{med}_P$.

2.5 Постановка задачи оптимизации

Рассматриваются две модели связи выборок $\mathbf{X}$ и $\mathbf{Y}$:

- **Аддитивная модель:** $a + \mathbf{X} = \mathbf{Y}$, т.е. $a + \mathbf{x}_i = \mathbf{y}_i$ для всех i.
- **Мультипликативная модель:** $t \cdot \mathbf{X} = \mathbf{Y}$, т.е. $t \cdot \mathbf{x}_i = \mathbf{y}_i$ для всех i.

Для аддитивной модели функционал вычисляется как:

$$F(a) = Ji\big(\{a + \mathbf{x}_i\},\ \{\mathbf{y}_i\}\big) = \frac{\sum_{i=1}^{n} \operatorname{wid}\big((a + \mathbf{x}_i) \cap \mathbf{y}_i\big)}{\sum_{i=1}^{n} \operatorname{wid}\big((a + \mathbf{x}_i) \cup \mathbf{y}_i\big)},$$

аналогично для мультипликативной модели $F(t) = Ji(\{t \cdot \mathbf{x}_i\},\ \{\mathbf{y}_i\})$.

Точечная оценка параметра $s \in \{a, t\}$ находится максимизацией функционала:

$$\hat{s} = \arg\max_{s} F(s, \mathbf{X}, \mathbf{Y}),$$

где рассматриваются четыре варианта функционала F:

B.1 $F(s) = Ji(s, \mathbf{X}, \mathbf{Y})$ — по полным выборкам,

B.2 $F(s) = Ji(s, \operatorname{mode}\mathbf{X}, \operatorname{mode}\mathbf{Y})$ — по модам,

B.3 $F(s) = Ji(s, \operatorname{med}_K \mathbf{X}, \operatorname{med}_K \mathbf{Y})$ — по медианам Крейновича,

B.4 $F(s) = Ji(s, \operatorname{med}_P \mathbf{X}, \operatorname{med}_P \mathbf{Y})$ — по медианам Пролубникова.

Интервальная оценка параметра строится как множество значений s, для которых $F(s) \geq \alpha \cdot F(\hat{s})$, где $\alpha \in (0, 1)$ — уровень доверия (в данной работе $\alpha = 0.95$).

3 Исходные данные

Загружены два файла данных диагностики томсоновского рассеяния:

- `-0.205_lvl_side_a_fast_data.bin` — выборка $\mathbf{X}$,
- `0.225_lvl_side_a_fast_data.bin` — выборка $\mathbf{Y}$.

Разрядность АЦП $N = 14$ бит, откуда $2^N = 16384$. Формула перевода кодов АЦП в вольты: $V = \mathrm{Code}/2^N - 0.5$.

Каждое измерение представляется интервалом $\mathbf{x}_i = [V_i - r,\ V_i + r]$ с радиусом $r = \operatorname{rad}\mathbf{x} = \operatorname{rad}\mathbf{y} = \frac{1}{2^{14}} = 0.0000610352$.

- $|\mathbf{X}| = 819200$ интервалов
- $|\mathbf{Y}| = 819200$ интервалов

4 Интервальные статистики

Таблица 1: Интервальные статистики выборок

Статистика	$\mathbf{X}$	$\mathbf{Y}$
mode	$[-0.171448,\ -0.171387]$	$[0.172058,\ 0.172119]$
med_K	$[-0.170593,\ -0.170471]$	$[0.172913,\ 0.173035]$
med_P	$[-0.170593,\ -0.170471]$	$[0.172913,\ 0.173035]$

5 Оптимизация параметров

Оптимизация проводилась методом перебора по равномерной сетке в два этапа: (1) грубый поиск на 500–1000 точках в широком диапазоне, (2) уточнение на 500 точках в окрестности найденного оптимума. Интервальная оценка параметра строилась как множество значений s, для которых $F(s) \geq 0.95 \cdot Ji_{\max}$.

5.1 Аддитивная модель: $a + \mathbf{X} = \mathbf{Y}$

Таблица 2: Результаты оптимизации для аддитивной модели

Функционал	$\hat{a}$	$Ji_{\max}$	Интервальная оценка a (95%)
B.1	0.342365	0.000061	$[0.341210,\ 0.344737]$
B.2	0.343506	0.993292	$[0.343504,\ 0.343507]$
B.3	0.343506	0.996640	$[0.343503,\ 0.343508]$
B.4	0.343506	0.996640	$[0.343503,\ 0.343508]$

5.2 Мультипликативная модель: $t \cdot \mathbf{X} = \mathbf{Y}$

Таблица 3: Результаты оптимизации для мультипликативной модели

Функционал	$\hat{t}$	$Ji_{\max}$	Интервальная оценка t (95%)
B.1	-1.014830	0.000052	$[-1.031182,\ -1.006814]$
B.2	-1.003915	0.988506	$[-1.003923,\ -1.003907]$
B.3	-1.014314	0.985888	$[-1.014339,\ -1.014298]$
B.4	-1.014314	0.985888	$[-1.014339,\ -1.014298]$

6 Графики

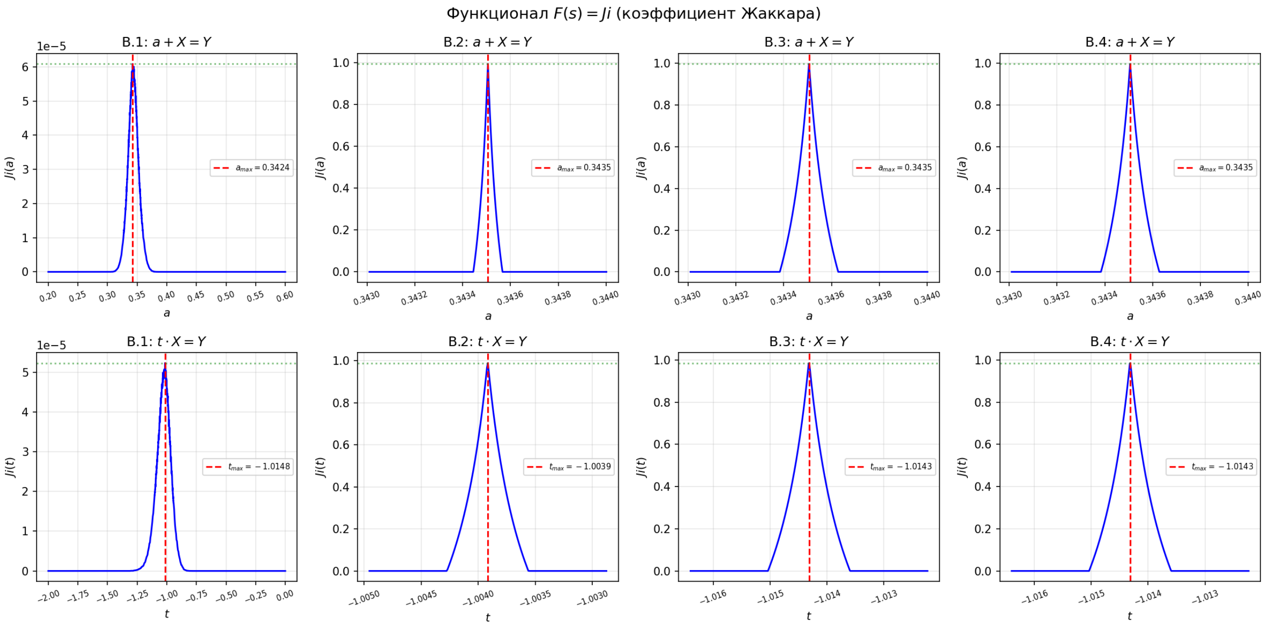

Рис. 1: Зависимость функционала $F(s) = Ji$ от параметра для аддитивной (верхний ряд) и мультипликативной (нижний ряд) моделей. Красная штриховая — $s_{\max}$, зелёная пунктирная — значение $Ji_{\max}$.

7 Сравнение результатов и выводы

1. Все четыре функционала дают согласованные оценки параметра a: среднее $\hat{a} \approx 0.3432$, разброс $\Delta a = 0.0011$.

2. Для мультипликативной модели $\hat{t} \approx -1.0118$. Коэффициент Жаккара для аддитивной модели выше во всех четырёх функционалах, что указывает на лучшее соответствие аддитивной модели данным.

3. При работе с полными выборками (B.1) $Ji_{\max} \approx 0.000$, тогда как для статистик (B.2–B.4) $Ji_{\max} \approx 0.99$. Малое значение для B.1 объясняется тем, что при 819200 парах интервалов с малым радиусом ($r \approx 6{,}1 \cdot 10^{-5}$) и значительным разбросом середин большинство пар $(a + \mathbf{x}_i) \cap \mathbf{y}_i$ имеют малое перекрытие. Для статистик же оптимизация проводится по одной паре представительных интервалов, что даёт почти полное совпадение при оптимальном s.

4. Использование интервальных статистик (мода, медианы Крейновича и Пролубникова) вместо полных выборок значительно ускоряет вычисления при близких точечных оценках параметров.

5. Медиана Крейновича и медиана Пролубникова дают идентичные результаты, поскольку все радиусы интервалов одинаковы ($\mathrm{rad}\,\mathbf{x}_i = \mathrm{const}$).

6. Для аддитивной модели медианы (B.3, B.4) дают $Ji_{\max} = 0.9966$, что выше моды $Ji_{\max} = 0.9933$. Для мультипликативной модели, напротив, мода ($Ji_{\max} = 0.9885$) превосходит медианы ($Ji_{\max} = 0.9859$). Это показывает, что выбор представительной статистики зависит от модели.

7. Интервальные оценки по полным выборкам (B.1) на 3 порядка шире, чем по статистикам (B.2–B.4): $\Delta a_{\mathrm{B.1}} \approx 0.004$ против $\Delta a_{\mathrm{B.2}} \approx 2 \cdot 10^{-6}$. Это следствие агрегации данных в одну пару представительных интервалов.

Репозиторий

Исходный код доступен по ссылке: `https://github.com/AleksandrShvartz/InterStat1`