Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

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

Выполнил студент:
Желудев К.И.
группа: 5030102/00101

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

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

1 Постановка задачи

Имеется выборка данных с интервальной неопределенностью. Число отсчетов в выборке равно 200. Используется модель данных с уравновешенным интервалом погрешности.

$\boldsymbol{x} = \overset{\circ}{x} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} = [-\epsilon, \epsilon]$ для некоторого $\epsilon > 0$,

Здесь $\overset{\circ}{x}$ – данные некоторого прибора, $\epsilon = 10^{-4}$ – погрешность прибора.
Необходимо:

- Иллюстрировать данные выборки
- Построить диаграмму рассеяния
- Найти базовые оценки исходной выборки
- Найти моду выборки и максимальную клику
- Произвести варьирование неопределенности измерений
- Вычислить меру совместности по индексу Жакара и относительную ширину моды

2 Теория

2.1 Оценки исходной выборки

Внешние оценки ищутся как:

$$\underline{J} = \min_{1 \le k \le n} \underline{x}_k, \quad \overline{J} = \min_{1 \le k \le n} \overline{x}_k \tag{1}$$

2.2 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.

Алгоритм для нахождения моды интервальной выборки:

1. $I \longleftarrow \cap_{i=1}^{n} x_i$
2. $I \neq \text{mode } X \longleftarrow I$; $\mu \longleftarrow n$
   Помещаем все концы $\underline{x}_1, \overline{x}_1, \ldots, \underline{x}_n, \overline{x}_n$ инетрвалов рассматриваемой выборки X в один массив $Y = (y_1, y_2, \ldots, y_{2n})$;
   Упорядочиваем элементы в Y по возрастанию значений;
   Порождаем интервалы $z_i = [y_i, y_{i+1}], i = 1, 2, \ldots, 2n-1$ (назовем их элементарными подинтервалами измерений);
   Для каждого z_i подсчитываем число μ_i интервалов из выборки X, включающих интервал z_i;
   Вычисляем $\mu \longleftarrow \max_{1 \le i \le 2n-1} \mu_i$;
   Вычисляем номера k интервалов z_k, для которых μ_k равно мксимальному, т.е. $\mu_k = \mu$ и формируем из таких k множество $K = \{k\} \subseteq \{1, 2, \ldots, 2n-1\}$;
   $\text{mode } X \longleftarrow \cup_{k \in K} z_k$

Значение максимальной клики равняется: max $\mu_j(X)$

2.3 Варьирование неопределенности изменений

Один из приемов выявления достижения совместности выборки интервальных наблюдений основан на представлении о причине несовместности как недооцененной величины неопределенности. Закономерным шагом в этом случае становится поиск некоторой минимальной коррекции величин неопределенности интервальных наблюдений, необходимой для обеспечения совместности задачи построения зависимости. Если величину коррекции каждого интервального наблюдения $y_i = [\overset{\circ}{y}_i - \epsilon_i, \overset{\circ}{y}_i + \epsilon_i]$ выборки S_n выражать коэффициентом его уширения $w_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде вектора коэффициентов $w^* = (w_1^*, \ldots, w_n^*)$, необходимая для совместности задачи построения $y = f(x, \beta)$ может быть решена решением задачи условной оптимизации:

Найти:

$$\min_{w,\beta} \sum_{i=1}^{n} w_i \tag{2}$$

При ограничениях:

$$\begin{cases} \overset{\circ}{y}_i - w_i\epsilon_i \leq f(x_i, \beta) \leq \overset{\circ}{y}_i + w_i\epsilon_i, \\ w_i \geq 1 \end{cases} \tag{3}$$

$i = 1, \ldots, n$

Результирующие значения коэффициентов w_i^*, строго превосхожящие единицу, указывают на наблюдения, которые требуют уширения интервалов неопределенности для обеспечения совместности данных и модели. Именно такие наблюдения заслуживают внимания пр ианализе данных на выбросы. Значительное количество подобных наблюдений может говорить либо о неверно выбранной структуре зависимости, либо о том, что величины неопределенности 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 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 может быть пополнена равенствами вида:

$$w_{i_1} = w_{i_2} = \ldots = w_{i_K},$$

где $i_1, \ldots, i_K$ – номера наблюдений группы. В случае, когда в надежности каких-либо наблюдений исследователь уверен полностью, при решении задачи (2) - (3) соответствующие им величины w_i можно положить равными единице, т.е. запретить вырьировать их неопределенность. Задачи поиска коэффициентов масштабирования величины неопределенности сформулирована для распространенного случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределенности измерений y_i при сохранении базовых значений интервалов $\overset{\circ}{y}_i$ неподвижными. При необходимости, постановка задачи легко обобщается. Например, если интервалы наблюдений не уравновешаны относительно базовых значений, то границы интервальных измерений можно варьировать независимо, масштабируя величины неопределенности ϵ_i^- и ϵ_i^+ с помощью отделимых коэффициентов w_i^- и w_i^+:

Найти:

$$\min_{w^-,w^+,\beta} \sum_{i=1}^{n} (w_i^- + w_i^+) \tag{4}$$

При ограничениях:

$$\begin{cases} \overset{\circ}{y}_i - w_i^- \epsilon_i^- \leq f(x_i, \beta) \leq \overset{\circ}{y}_i + w_i^+ \epsilon_i^+, \\ w_i^- \geq 1 \\ w_i^+ \geq 1 \end{cases} \tag{5}$$

$i = 1, \ldots, n$

Для линейной по параметрам β зависимости $y = f(x, \beta)$ задача представляет собой задачу линейного программирования, для решения которой 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.

2.4 Оптимизация по Оскорбину

Постановка задачи линейного программирования в простейшем виде:

Найти:

$$\min_{w,\beta} w \tag{6}$$

При ограничениях:

$$\begin{cases} \text{mid} x_i - w\epsilon_i \leq \beta \leq \text{mid} x_i + w\epsilon_i \\ w \geq 1 \end{cases} \tag{7}$$

$i = 1, \ldots, n$

2.5 Индекс Жаккара

Для описания выборок, помимо оценок их размеров, желательно иметь дополнительную информацию о мере сходимости элементов выборки. В различных областях анализа данных, биологии, информатике и науках о Земле часто используют различные меры сходства множеств.

Рассмотрим один из возможных коэффициентов совместности – это отношение инфимума по включению к супремуму по включению – индекс Жаккара:

$$J_i(X) = \frac{\mathrm{wid}(\wedge_i x_i)}{\mathrm{wid}(\vee_i x_i)} \tag{8}$$

Индекс Жакара непрерывно описывает ситуациии от полной несовместности выборок до полного перекрытия интервалов. Он может принимать значения:

$$-1 \leq J_i(X) \leq 1$$

2.6 Относительная ширина моды

Относительная ширина моды равна:

$$\rho(\mathrm{mode}(X)) = \frac{\mathrm{wid}(\mathrm{mode}(X))}{\mathrm{wid}(\vee_i x_i)} \tag{9}$$

В отличие от минимума по включению, мода выборки всегда является правильным интервалом. В целом получаем:

$$0 \leq \rho(\mathrm{mode}(X)) \leq 1$$

3 Результаты

3.1 Оценки исходной выборки

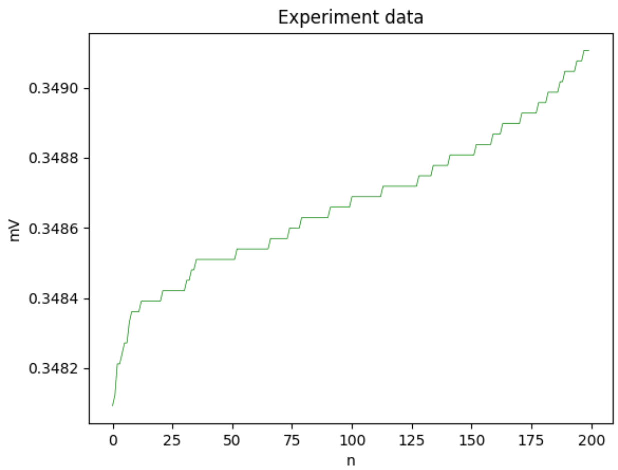

Рис. 1: Данные выборки X_1

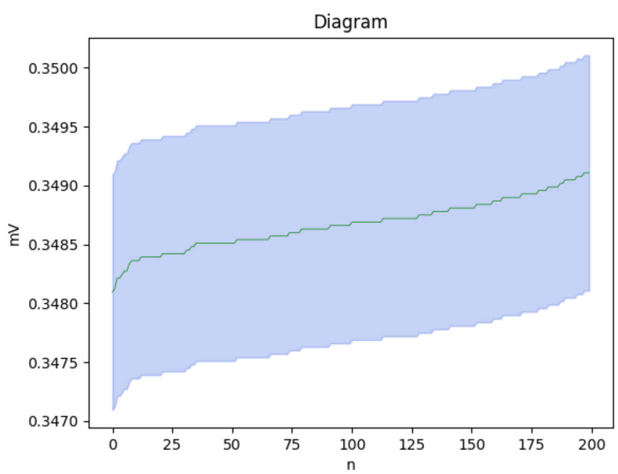

Рис. 2: Диаграмма рассеяния выборки X_1 с уравновешанным интервалом неопределенности

Вычисленные внешние оценки выборки:

$$\underline{J}_1 = 0.34709, \quad \overline{J}_1 = 0.35011$$

3.2 Мода и максимальная клика выборки

Мода выборки:

$$\text{mode}(X_1) = [0.34811, 0.34909]$$

Максимальная клика:

$$\max \mu_j(X_1) = 200$$

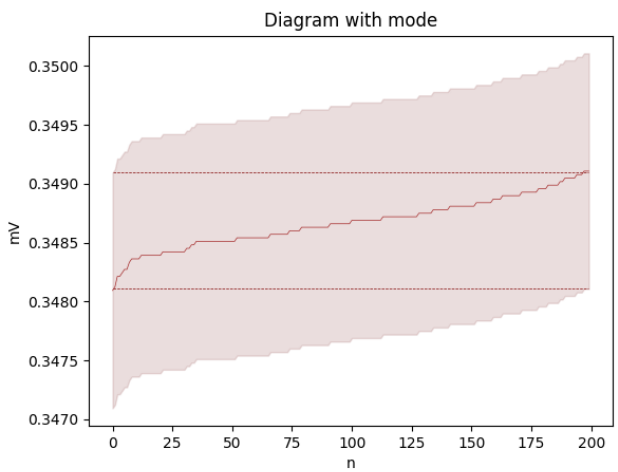

Рис. 3: Элементы выборки X_1, в которые входит мода

3.3 Варьирование неопределенности изменений

В результате решения задачи линейного программирования при оптимизации по Оскорбину были найдены следующие значения:
Оценка постоянной:

$$\beta = 0.34811$$

Коэффициент растяжения интервала неопределенности:

$$w = 1.0$$

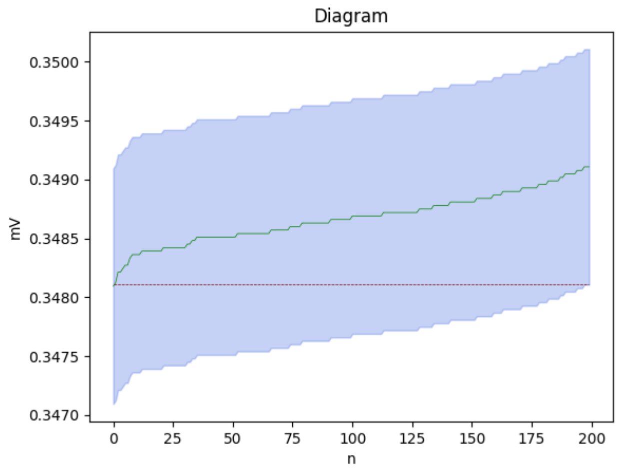

Рис. 4: Диаграмма рассеяния X_1 с увеличенным в w раз интервалом неопределенности

Красной пунктирной линией обозначена оценка постоянной.

3.4 Коэффициент Жакара и относительная ширина моды

Получены следующие значения:
Индекс Жакара:

$$J_i(X_1) = 0.32744$$

Относительная ширина моды:

$$\rho(\mathrm{mode}(X_1)) = 0.32744$$

4 Обсуждение

4.1 Оценки исходной выборки

На основе полученных результатов можно сделать вывод, что верхние и нижние вершины оценок J_1 совпадают с границами отображения на рис. 3.

4.2 Мода и максимальная клика выборки

Полученная мода входит во все элементы выборки, что свидетельствуют о полной совместности выборки.

4.3 Варьирование неопределенности изменений

Полученная оценка постоянной β почти совпала с нижней границей моды, вычисленной ранее. Величина однородного расширения равна единице, что свидетельствует о полной совместности выборки.

4.4 Коэффициент Жакара и относительная ширина моды

Положительность коэффициента Жакара свидетельствует о перекрытии интервалов выборки. В данном случае, можно сделать вывод, что выборка совместна. Коэффициент Жакара и относительная ширина моды совпали, поскольку ширина пересечения интервалов и ширина моды выборки тоже совпали.

5 Реализация

Лабораторная работа выполнена с использованием языка программирования Python 3.10 в среде разработки PyCharm Community с использованием библиотек

- pandas v.1.5.3
- matplotlib v.3.7.1
- numpy v.1.24.2
- scipy v.1.10.1

6 Приложение

Ссылка на репозиторий GitHub: `https://github.com/krzhld/mathstat`