Санкт-Петербургский политехнический университет
Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Отчет
по лабораторной работе №9
по дисциплине
«Математическая статистика»

Выполнил студент:
Иванова А.С.
группа: 5030102/00101

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

Санкт-Петербург
2023 г.

Содержание

Список иллюстраций

Список таблиц

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.
Нужно иллюстрировать данные выборки, построить диаграмму рассеяния, найти базовые оценки исходной выборки, 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:

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

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. $\boldsymbol{I} \longleftarrow \cap_{i=1}^{n} \boldsymbol{x}_i$
2. **if** $I \neq \varnothing$ **then**
   mode $\boldsymbol{X} \longleftarrow \boldsymbol{I}$; $\mu \longleftarrow n$
   else
   Помещаем все концы $\underline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_n, \overline{\boldsymbol{x}}_n$ инетрвалов рассматриваемой выборки $\boldsymbol{X}$ 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 $\boldsymbol{X}$, включающих интервал z_i;
   Вычисляем $\mu \longleftarrow \max\limits_{1 \le i \le 2n-1} \mu_i$;
   Вычисляем номера k интервалов z_k, для которых μ_k равно мксимальному, т.е. $\mu_k = \mu$ и формируем из таких k множество $K = \{k\} \subseteq \{1, 2, \ldots, 2n-1\}$;
   mode $\boldsymbol{X} \longleftarrow \cup_{k \in K} z_k$
   end if

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

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 $\boldsymbol{y}_i = [\overset{\circ}{y}_i - \epsilon_i, \overset{\circ}{y}_i + \epsilon_i]$ 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 такеи 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 (напрмиер, 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. за-

претить вырьировать их неопределенность.

Задачи поиска коэффициентов масштабирования величны неопределенности сформулирована для распространенного случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределенности измерений $\boldsymbol{y}_i$ при сохранении базовых значений интервалов $\overset{\circ}{y}_i$ неподвижными. При необходимости, постановка задачи легко обощается. 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:

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

$i = 1, \ldots, n$

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 инфинума по включению к супренуму по включению – индекс Жаккара:

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

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

$$-1 \leq J_i(\boldsymbol{X}) \leq 1$$

2.6 Относительная ширина моды

Относительная ширина моды равна:

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

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

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

3 Реализация

Лабораторная работа была выполнена с помощью встроенных средств языка программирования Python (библиотеки: NumPy, SciPy, Matplotlib, Pandas) в среде разработки Visual Studio Code. Исходный код работы приведен в приложении.

Ссылка на репозиторий с исходным кодом: `https://github.com/anivse/MathematicalStatistics`

4 Результаты

4.1 Оценки исходной выборки

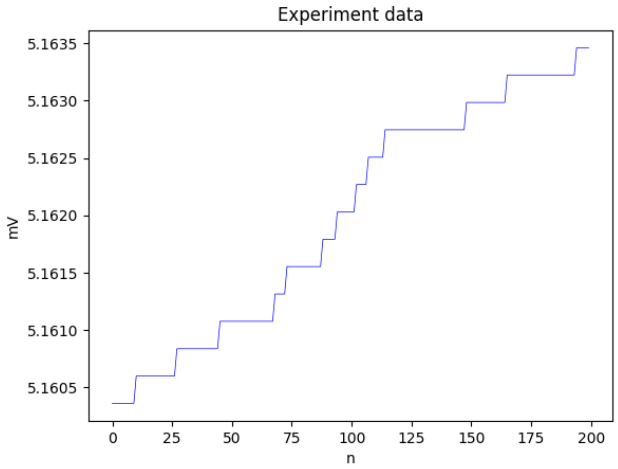

Рис. 1: Данные выборки $\boldsymbol{X}_1$

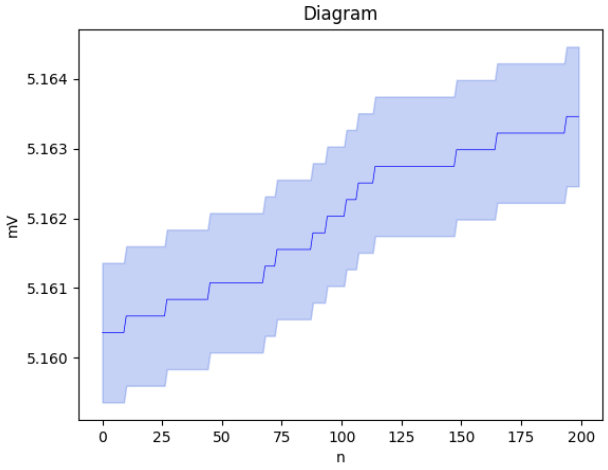

Рис. 2: Диаграмма рассеяния выборки $\boldsymbol{X}_1$ с уравновешанным интервалом неопределенности

Вычисленные внешние оценки выборки:

$$\underline{\boldsymbol{J}}_1 = 5.15936, \quad \overline{\boldsymbol{J}}_1 = 5.16446$$

4.2 Мода и максимальная клика выборки

Мода выборки:

$$\text{mode}(\boldsymbol{X}_1) = [5.16246, 5.16255]$$

Максимальная клика:

$$\max\mu_j(\boldsymbol{X}_1) = 127$$

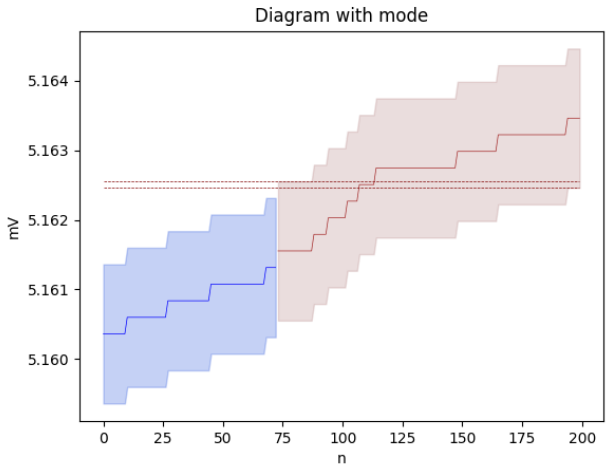

Рис. 3: Элементы выборки $\boldsymbol{X}_1$, в которые входит мода

4.3 Варьирование неопределенности изменений

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

$$\beta = 5.16191$$

Коэффициент растяжения интервала неопределенности:

$$w = 1.54950$$

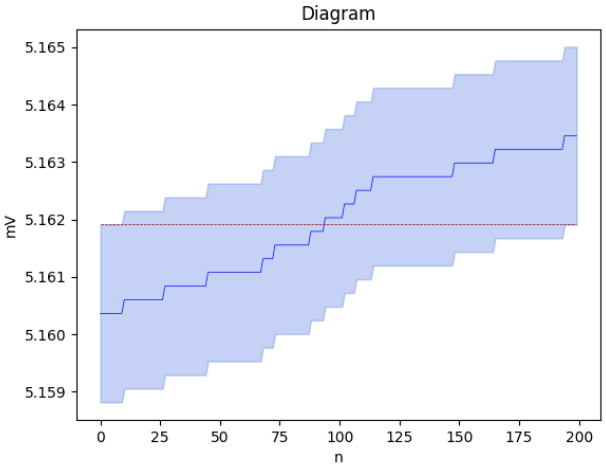

Рис. 4: Диаграмма рассеяния $\boldsymbol{X}_1$ с увеличенным в w раз интервалом неопределенности

Красной пунктирной линией обозначена оценка постоянной.

4.4 Коэффициент Жакара и относительная ширина моды

Получены следующие значения:
Индекс Жакара:

$$J_i(\boldsymbol{X}_1) = -0.21553$$

Относительная ширина моды:

$$\rho(\text{mode}(\boldsymbol{X}_1)) = 0.01824$$

5 Обсуждение

5.1 Оценки исходной выборки

На основе полученных результатов можно сделать вывод, что верхние и нижние вершины оценок $\boldsymbol{J}_1$ совпадают с границами отображения на рис.3.

5.2 Мода и максимальная клика выборки

Полученная мода входит в большую часть элементов выборки, что свидетельствуют о невысокой степени несовместности выборки.

5.3 Варьирование неопределенности изменений

Полученная оценка постоянной β получилась очень близка к вычисленной ранее моде.
Величина однородного расширения интервалов невелика, что свидетельствует о невысокой степени несовместности выборки.

5.4 Коэффициент Жакара и относительная ширина моды

Отрицательности коэффициента Жакара свидетельствует о несовместности выборки, а модуль - о степени несовместности. В данном случае, можно сделать вывод, что выборка несовместна, но степень несовместности невелика.

Величина относительной ширины моды составляет менее 2% внешней оценки выборки $\boldsymbol{X}_1$.