Санкт-Петербургский политехнический университет имени Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

Математическая статистика
Отчёт по лабораторным работам №1-4

Выполнил:

Студент: Золин Иван
Группа: 5030102/00201

Принял:

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

2023 г.

Содержание

Список иллюстраций

Список таблиц

1 Постановка задачи

Для четырех распределений:

- Нормальное распределение: $N(x, 0, 1)$
- Распределение Коши: $C(x, 0, 1)$
- Распределение Пуассона: $P(k, 10)$
- Равномерное распределение: $U(x, -\sqrt{3}, \sqrt{3})$

Выполнить следующие задачи:

1. Сгенерировать выборки размером 10, 50 и 1000 элементов. Построить на одном рисунке гистограмму и график плотности распределения. $\xi \sim \mathcal{N}(0, 1)$
2. Сгенерировать выборки размером 10, 100 и 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных: $\overline{x}, medx, z_R, z_Q, z_{tr}$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

$$E(z) = \overline{z} \tag{1}$$

   Вычислить оценку дисперсии по формуле:

$$D(z) = \overline{z^2} - \overline{z}^2 \tag{2}$$

   Представить полученные данные в виде таблиц.
3. Сгенерировать выборки размером 20 и 100 элементов.
   Построить для них боксплот Тьюки. Для каждого распределения определить долю выбросов экспериментально (сгенерировав выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.
4. Сгенерировать выборки размером 20, 60 и 100 элементов. Построить на них эмпирические функции распределения и ядерные оценки плотности распределения на отрезке $[-4; 4]$ для непрерывных распределений и на отрезке $[6; 14]$ для распределения Пуассона.

2 Теория

2.1 Распределения

Плотности классических распределений:

- Нормальное распределение

$$N(x,0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3}$$

- Распределение Коши

$$C(x,0,1) = \frac{1}{\pi}\frac{1}{x^2+1} \tag{4}$$

- Распределение Лапласа

$$L(x,0,\frac{1}{\sqrt{2}}) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \tag{5}$$

- Распределение Пуассона

$$P(k,10) = \frac{10^k}{k!} e^{-10} \tag{6}$$

- Равномерное распределение

$$U(x,-\sqrt{3},\sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при} \quad |x| \le \sqrt{3} \\ 0 & \text{при} \quad |x| > \sqrt{3} \end{cases} \tag{7}$$

2.2 Гистограмма

2.2.1 Определение

Гистограмма в математической статистике — это функция, приближающая плотность вероятности некоторого распределения, построенная на основе выборки из него [1].

2.2.2 Графическое описание

Графически гистограмма строится следующим образом. Сначала множество значений, которое может принимать элемент выборки, разбивается на несколько интервалов. Чаще всего эти интервалы берут одинаковыми, но это не является строгим требованием. Эти интервалы откладываются на горизонтальной оси, затем над каждым рисуется прямоугольник. Если все интервалы были одинаковыми, то высота каждого прямоугольника пропорциональна числу элементов выборки, попадающих в соответствующий интервал. Если интервалы разные, то высота прямоугольника выбирается таким образом, чтобы его площадь была пропорциональна числу элементов выборки, которые попали в этот интервал [1].

2.2.3 Использование

Гистограммы применяются в основном для визуализации данных на начальном этапе статистической обработки.
Построение гистограмм используется для получения эмпирической оценки плотности распределения случайной величины. Для построения гистограммы наблюдаемый диапазон изменения случайной величины разбивается на несколько интервалов и подсчитывается доля от всех измерений, попавшая в каждый из интервалов. Величина каждой доли, отнесенная к величине интервала, принимается в качестве оценки значения плотности распределения на соответствующем интервале [1].
Они являются мощным инструментом для исследования неизвестных распределений.
В частности, на этом построены различные способы обработки сигналов, изображений и других статистических объектов. Применение гистограмм к обработке экспериментальных данных позволяет устранять артефакты - шумы и выбросы, мешающие работе с данными и не являющиеся содержательными.
Шумы определяются как горизонтальные участки сигнала. Присутствуют они только до и после полезного сигнала, на нем они быть не могут. Определяются участки шума так: находятся границы 2-ух самых больших

столбцов гистограммы полученного сигнала, затем путем прохода по сигналу в обе стороны окном определнного размера и сравнения процентного соотношения значений внутри границ большого столбца с выбранным порогом, принимается решение о пометке участка как шумового при превышении этого порога.
Выбросы — экстремальные значения во входных данных, находящиеся далеко за пределами других наблюдений. На гистограмме выбросы будут формировать одиночные пики. Выбросы следует заменить чем-то разумным (средним, медианой в окрестности). Нужно идти по гистограмме скользящим окном (параметр), и, что вылетает далеко по гистограмме (или за 3 сигмы), заменять медианным значением. Тогда ожидается, что колонны гистограммы сравняются с пейзажем.

2.3 Вариационный ряд

Вариационным рядом называется последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются [2, с. 409].
Запись вариационного ряда: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.
Элементы вариационного ряда $x_{(i)}$ $(i = 1, 2, \dots, n)$ называются порядковыми статистиками.

2.4 Выборочные числовые характеристики

С помощью выборки образуются её числовые характеристики. Это числовые характеристики дискретной случайной величины X^*, принимающей выборочные значения $x_1, x_2, \dots, x_n$ [2, с. 411].

2.4.1 Характеристики положения

- Выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

- Выборочная медиана

$$med\, x = \begin{cases} x_{(l+1)} & \text{при} \quad n = 2l+1 \\ \dfrac{x_{(l)} + x_{(l+1)}}{2} & \text{при} \quad n = 2l \end{cases} \tag{9}$$

- Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{10}$$

- Полусумма квартилей

  Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном,} \\ x_{(np)} & \text{при } np \text{ целом.} \end{cases} \tag{11}$$

  Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{12}$$

- Усечённое среднее

$$z_{tr} = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} x_{(i)}, \quad r \approx \frac{n}{4} \tag{13}$$

2.4.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{14}$$

2.5 Боксплот Тьюки

2.5.1 Определение

Боксплот (англ. box plot) — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

2.5.2 Описание

Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы [3].

2.5.3 Построение

Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину «усов» определяют разность первого квартиля и полутора межквартильных расстояний и сумма третьего квартиля и полутора межквартильных расстояний. Формула имеет вид

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), \;\; X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1), \tag{15}$$

где X_1 — нижняя граница уса, X_2 — верхняя граница уса, Q_1 — первый квартиль, Q_3 — третий квартиль.

Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков [3].

2.6 Теоретическая вероятность выбросов

Встроенными средствами языка программирования R в среде разработки RStudio можно вычислить теоретические первый и третий квартили распределений ($Q_1^{\text{т}}$ и $Q_3^{\text{т}}$ соответственно). По формуле (15) можно вычислить теоретические нижнюю и верхнюю границы уса ($X_1^{\text{т}}$ и $X_2^{\text{т}}$ соответственно). Выбросами считаются величины x, такие что:

$$\begin{bmatrix} x < X_1^{\text{т}} \\ x > X_2^{\text{т}} \end{bmatrix} \tag{16}$$

Теоретическая вероятность выбросов для непрерывных распределений

$$P_{\text{в}}^{\text{т}} = P(x < X_1^{\text{т}}) + P(x > X_2^{\text{т}}) = F(X_1^{\text{т}}) + \Big(1 - F(X_2^{\text{т}})\Big), \tag{17}$$

где $F(X) = P(x \le X)$ - функция распределения.

Теоретическая вероятность выбросов для дискретных распределений

$$P_{\text{в}}^{\text{т}} = P(x < X_1^{\text{т}}) + P(x > X_2^{\text{т}}) = \Big(F(X_1^{\text{т}}) - P(x = X_1^{\text{т}})\Big) + \Big(1 - F(X_2^{\text{т}})\Big), \tag{18}$$

где $F(X) = P(x \le X)$ - функция распределения.

2.7 Эмпирическая функция распределения

2.7.1 Статистический ряд

Статистическим рядом называется последовательность различных элементов выборки $z_1, z_2, \ldots, z_k$, расположенных в возрастающем порядке с указанием частот $n_1, n_2, \ldots, n_k$, с которыми эти элементы содержатся в выборке.

Статистический ряд обычно записывается в виде таблицы

z	z_1	z_2	...	z_k
n	n_1	n_2	...	n_k

Таблица 1: Статистический ряд

2.7.2 Определение

Эмпирической (выборочной) функцией распределения (э. ф. р.) называется относительная частота события $X < x$, полученная по данной выборке:

$$F_n^*(x) = P^*(X < x). \tag{19}$$

2.7.3 Описание

Для получения относительной частоты $P^*(X < x)$ просуммируем в статистическом ряде, построенном по данной выборке, все частоты n_i, для которых элементы z_i статистического ряда меньше x. Тогда $P^*(X < x) = \frac{1}{n} \sum\limits_{z_i<x} n_i$. Получаем

$$F^*(x) = \tfrac{1}{n} \sum_{z_i<x} n_i. \tag{20}$$

$F^*(x)$ — функция распределения дискретной случайной величины X^*, заданной таблицей распределения

X^*	z_1	z_2	...	z_k
P	$\frac{n_1}{n}$	$\frac{n_2}{n}$	...	$\frac{n_k}{n}$

Таблица 2: Таблица распределения

Эмпирическая функция распределения является оценкой, т. е. приближённым значением, генеральной функции распределения

$$F_n^*(x) \approx F_X(x). \tag{21}$$

2.8 Оценки плотности вероятности

2.8.1 Определение

Оценкой плотности вероятности $f(x)$ называется функция $\widehat{f}(x)$, построенная на основе выборки, приближённо равная $f(x)$

$$\widehat{f}(x) \approx f(x). \tag{22}$$

2.8.2 Ядерные оценки

Представим оценку в виде суммы с числом слагаемых, равным объёму выборки:

$$\widehat{f}_n(x) = \tfrac{1}{nh_n} \sum_{i=1}^{n} K\left(\tfrac{x-x_i}{h_n}\right). \tag{23}$$

Здесь функция $K(u)$, называемая ядерной (ядром), непрерывна и является плотностью вероятности, $x_1, \ldots, x_n$ — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, обладающая свойствами

$$h_n \underset{n\to\infty}{\longrightarrow} 0; \qquad \frac{h_n}{n^{-1}} \underset{n\to\infty}{\longrightarrow} \infty. \tag{24}$$

Такие оценки называются непрерывными ядерными [2, с. 421-423].

Замечание. Свойство, означающее сближение оценки с оцениваемой величиной при $n \to \infty$ в каком-либо смысле, называется состоятельностью оценки.

Если плотность $f(x)$ кусочно-непрерывная, то ядерная оценка плотности является состоятельной при соблюдении условий, накладываемых на параметр сглаживания h_n, а также на ядро $K(u)$.

Гауссово (нормальное) ядро [4, с. 38]

$$K(u) = \tfrac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{25}$$

Правило Сильвермана [4, с. 44]

$$h_n = 1.06\hat{\sigma}n^{-1/5}, \tag{26}$$

где $\hat{\sigma}$ - выборочное стандартное отклонение.

2.8.3 Оценка качества ядерных приближений

После получения результатов возникает необходимость оценить качество ядерных приближений. Приведем в пример один из приемов количественных описаний сходства кривых - метод Фреше.

Расстояние Фреше − это мера сходства кривых, принимающая во внимание число и порядок точек вдоль кривых. Расстояние названо по имени французского математика Мориса Фреше. Метрика Фреше принимает во внимание течение лвух кривых, посколько пары точек, расстояние между которыми определяет расстояние Фреше, «пробегают» вдоль кривых. Расстояние Фреше между двумя кривыми − это не длина самого короткого поводка, с котрым можно пройти все пути, а самый короткий, при котором можно пройти этот путь.

Определим кривую как непрерывное отображение $f : [a, b] \to V$, где $a, b \in R$ и $a \leq b$ и $(V, d)-$ метрическое пространство. Даны две кривые $f : [a, b] \to V$ и $g : [a', b'] \to V$, их расстояние Фреше определено в виде:

$$\delta F(f, g) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \{d(f(\alpha(t)), g(\beta(t)))\} \tag{27}$$

где α (соответственно β) − произвольная непрерывная неубывающая функция из $[0, 1]$ на $[a, b]$ (соттветственно $[a', b']$).
При вычислении расстояния Фреше между произвольными кривыми обычно аппроксимируют кривые многоугольными кривыми. Многоульная кривая − это кривая $P : [0, n] \to V$, где $n-$ натуральное число, такое, что для каждого $i \in [0, n - 1]$ ограничение P к интервалу $[i, i + 1]$ является аффинным, то есть $P(i + \lambda) = (1 - \lambda)P(i) + \lambda P(i + 1)$.
Для заданных многоугольных кривых P и Q их дискретное расстояние Фреше определяется как: где $L - PQ$.
Необязательно решение − пара точек, между которыми найдено расстояние Фреше, − является единственным.

Рассмотрим расстояние Фреше для двух кривых: ядерной оценки плотности для равномерного распределения и самого равномерного распределения при $n = 100$:

3 Результаты

3.1 Гистограмма

3.2 Характеристики положения и рассеяния

3.3 Боксплот Тьюки

3.4 Доля выбросов

3.5 Теоретическая вероятность выбросов

3.6 Эмпирическая функция распределения

3.7 Ядерные оценки плотности распределения

4 Обсуждение

5 Реализация

5.1 Описание

Данная лабораторная работа была выполнена с использованием языка программирования Python 3.10 в среде разработки PyCharm с использованием следующих библиотек:

- scipy версии 1.8.0
- numpy версии 1.22.0

- matplotlib версии 3.5.1
- seaborn версии 0.11.2

5.2 Ссылка на репозиторий

https://github.com/IMZolin/Math-statistics-labs - GitHub репозиторий

Список литературы

[1] Histogram. URL: https://en.wikipedia.org/wiki/Histogram

[2] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

[3] Box plot. URL: https://en.wikipedia.org/wiki/Box_plot

[4] Анатольев, Станислав (2009) «Непараметрическая регрессия», Квантиль, №7, стр. 37-52.