Министерство образования и науки

Российской федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАНИЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Курганский государственный университет»

Кафедра «Программное обеспечение автоматизированных систем»

**КУРСОВАЯ РАБОТА**

**по дисциплине**

**«Теория вероятностей и математическая статистика»**

**РФ КГУ 09.03.04. КР24.260100**

Выполнил студент гр. ИТ- 0940322 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /Кузнецов В.А./

подпись, дата

Направление 09.03.04 - «Программная инженерия»

Проверил канд. физ.- мат. наук, доцент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / Черепанов О.С. /

подпись, дата

Работа защищена с оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2024

**Содержание**

[Введение 3](#_Toc167969088)

[1 Постановка задачи 4](#_Toc167969089)

[2 Распределение Лапласа. 5](#_Toc167969090)

[3 Описание оценок 6](#_Toc167969091)

[4 Расчет СКО 8](#_Toc167969092)

[5 Генерация выбросов 9](#_Toc167969093)

[6 Исследование оценок 11](#_Toc167969094)

[Заключение 16](#_Toc167969095)

[Список использованных источников 17](#_Toc167969096)

## Введение

В данной курсовой работе будет проведено сравнение двух оценок для нахождения наиболее эффективной оценки: выборочное среднее и оценка Ходжеса-Лемана.

Основная цель данной курсовой работы заключается в закрепление знаний по теории вероятности, и проведении исследования и сравнения оценок параметра распределения на основе бутстреп-метода.

## 1 Постановка задачи

На основе бутстреп-метода провести исследование и сравнение оценок выборочного среднего и Ходжеса-Лемана и параметра сдвига распределения Лапласа:

1. реализовать датчики псевдослучайных чисел из распределения и сформировать выборки заданного объема;
2. построить оценки плотности вероятностей случайной величины X по исходным выборкам;
3. построить оценки выборочного среднего и Ходжеса-Лемана.

Исследовать оценки:

1. На объемах выборки 50 и 500.

2. Без выбросов, при наличии симметричных и асимметричных выбросов модели Тьюки.

3. Используя бутстреп-метод, оценить среднее, дисперсию, среднеквадратическое отклонение (СКО).

Провести сравнение оценок выборочного среднего и Ходжеса-Лемана.

Сделать выводы.

## 

# 2 МЕТОДОЛОГИЯ ОЦЕНКИ ПАРАМЕТРОВ И РАСЧЁТ СКО

В данной главе представлена методология оценки параметров распределения Лапласа, включая использование оценок выборочного среднего и Ходжеса-Лемана, и описаны методы расчета среднеквадратического отклонения (СКО), бутстреп-метод, необходимый для повышения точности оценок. Так же рассмотрен метод формирования симметричных и асимметричных выбросов по модели Тьюки.

## 2.1 Распределение Лапласа

Распределение Лапласа, также известное как двустороннее экспоненциальное распределение, используется в статистике для моделирования данных, которые имеют центральное пиковое значение и хвосты, убывающие экспоненциально. Это распределение характеризуется двумя параметрами: μ (среднее или медиана) и b (параметр масштаба).

Функция распределения вероятностей для распределения Лапласа определяется как:

.

Плотность вероятности:

.

Свойства распределения Лапласа:

1. симметрия: Распределение Лапласа симметрично относительно параметра μ;
2. среднее:;
3. дисперсия: ;
4. куртозис: Распределение Лапласа имеет более "острые" пики и более "тяжелые" хвосты по сравнению с нормальным распределение.

## 2.2 Описание оценок

Оценка выборочного среднего — это статистический показатель, который используется для оценки среднего значения (математического ожидания) генеральной совокупности на основе данных, полученных из выборки [1]. Она представляет собой среднее арифметическое значений элементов выборки. Оценка выборочного среднего рассчитывается по следующей формуле:

.

где:

- — выборочное среднее,

- — объем выборки (число элементов в выборке),

- — значения элементов выборки.

Эта оценка используется в различных статистических анализах и служит базовым инструментом для проверки гипотез, анализа данных и построения доверительных интервалов.

Оценка Ходжеса-Лемана — это робастная оценка центральной тенденции, которая используется для оценки среднего значения генеральной совокупности на основе выборки. В отличие от выборочного среднего, оценка Ходжеса-Лемана менее чувствительна к выбросам и отклонениям от нормальности распределения данных.

Оценка Ходжеса-Лемана определяется как медиана всех попарных средних значений выборки. Формально она рассчитывается следующим образом:

1. Для выборки из элементов вычисляются все возможные средние попарных комбинаций:

.

2. Находится медиана среди всех полученных значений .

Таким образом, оценка Ходжеса-Лемана:

.

Эта оценка используется в задачах, где важна устойчивость к выбросам и более точное отражение центральной тенденции в данных с аномальными значениями.

## 2.3 Расчёт среднеквадратической ошибки (СКО)

Среднеквадратичная ошибка (СКО) является одним из наиболее важных и широко используемых показателей в статистике для оценки разброса данных относительно их среднего значения. Оно используется для определения степени разброса или вариации данных и помогает в оценке их надежности и точности.

Среднеквадратичная ошибка (СКО) определяет, насколько значения в наборе данных отклоняются от среднего значения этого набора. Меньшее СКО указывает на меньший разброс данных, тогда как большее СКО свидетельствует о большем разбросе.

Формулы расчёта СКО для выборки:

1. Дисперсия выборки:

.

где - выборочное среднее:

.

где n – количество значений выборки, Xi – отдельное значение выборки.

1. Среднеквадратичная ошибка выборки:

.

## 2.4 Генерация выбросов

Выброс — в статистике результат измерения, выделяющийся из общей выборки. Статистический метод, способный действовать в условиях выбросов, называется робастным. Оценка Ходжеса-Лемана является робастной характеристикой, а выборочное среднее — нет.

Модель выбросов Тьюки может быть представлена следующим образом: , где – основное распределение, – распределение выбросов, – доля выбросов.

Генерация выбросов происходит путем добавления в исходную выборку значения непрерывного равномерного распределения.

Непрерывное равномерное распределение в теории вероятностей — распределение случайной вещественной величины, принимающей значения, принадлежащие некоторому промежутку конечной длины, характеризующееся тем, что плотность вероятности на этом промежутке почти всюду постоянна.

Функция распределения вероятностей для непрерывного равномерного распределения определяется как:

.

Плотность вероятности:

*.*

Генерация симметричных выбросов происходит путем добавления 10% от исходной выборки значений непрерывного равномерного распределения с параметрами a = -10 и b = 10.

Генерация асимметричных выбросов происходит путем добавления 10% от исходной выборки значений непрерывного равномерного распределения с параметрами a = 15 и b = 16.

## 2.5 Бутстреп-метод

Бутстреп-метод (bootstrap) — это мощный статистический инструмент, разработанный для оценки свойств выборок и параметров распределений без необходимости использования предположений о форме распределения генеральной совокупности. Метод бутстрепа был введен Брэдли Эфроном в 1979 году и с тех пор стал широко использоваться в статистике и машинном обучении.

Бутстреп используется для:

* **Оценки точности статистических оценок:** Определение дисперсии, среднеквадратической ошибки (СКО), доверительных интервалов и других параметров.
* **Тестирования гипотез**: Проведение статистических тестов без необходимости жестких предположений о распределении данных.
* **Улучшения стабильности моделей:** Повышение надежности оценок в условиях ограниченных или малых выборок.

Существует несколько разновидностей бутстреп-методов:

* **Наивный бутстреп:** Генерация ревыборок путем случайного извлечения с возвращением из исходной выборки.
* **Параметрический бутстреп:** Генерация ревыборок из параметрической модели, основанной на исходных данных.
* **Полу-наивный бутстреп:** Комбинация наивного и параметрического подходов.

**Блоковый бутстреп:** Используется для временных рядов и данных с зависимостями, где данные делятся на блоки перед выборкой.

## 3 ИССЛЕДОВАНИЕ ОЦЕНОК

В этой главе проводится исследование эффективности оценок выборочного среднего и оценки Ходжеса-Лемана для параметра сдвига распределения Лапласа с использованием бутстреп-метода. Основное внимание уделяется анализу точности и надёжности этих оценок при различных условиях, включая разные объемы выборок и наличие выбросов. Мы рассмотрим, как выборочное среднее и оценка Ходжеса-Лемана ведут себя при симметричных и асимметричных выбросах.

Для исследования оценок было реализовано следующие:

* генератор псевдослучайных величин;
* «Наивный» бутстреп-метод;
* генератор псевдослучайных величин модели выбросов Тьюки.

Генерация выборки значений распределения Лапласа происходила с параметрами сдвига – 0 и масштаба – 1.

## 3.1 Исследование оценок в отсутствии выбросов

Таблица 1 - Исследование оценок на основе генератора псевдослучайных величин

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Объем исходной выборки | Дисперсия для выборочной средней оценки | Дисперсия для оценки Ходжеса-Лемана | Квадрат смещения для выборочной средней оценки | Квадрат смещения для оценки Ходжеса-Лемана |
| 50 | 0.03233 | 0.01950 | 0.004456 | 0.005969 |
| 500 | 0.00346 | 0.00242 | 0.003054 | 0.002350 |

|  |  |  |  |
| --- | --- | --- | --- |
| СКО для выборочной средней оценки | СКО для оценки Ходжеса-Лемана | Соотношение СКО | График рисунок |
| 0.03679 | 0.02547 | 1.444 | Рисунок 1 |
| 0.006523 | 0.004776 | 1.366 | Рисунок 2 |

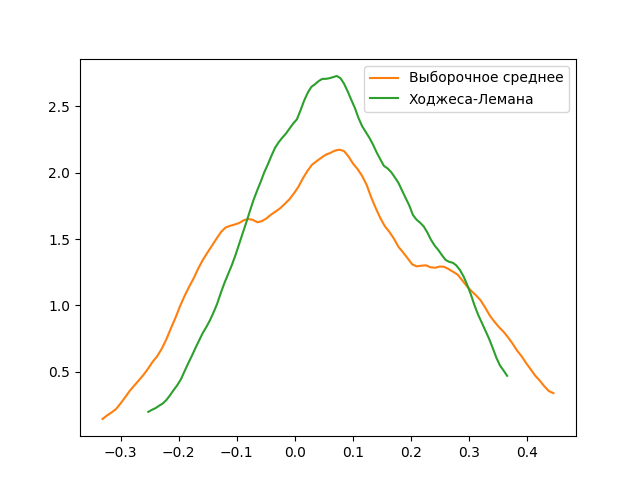


Рисунок 1 – Оценка плотности при объеме выборки = 50 и кол-во ре-выборок = 100

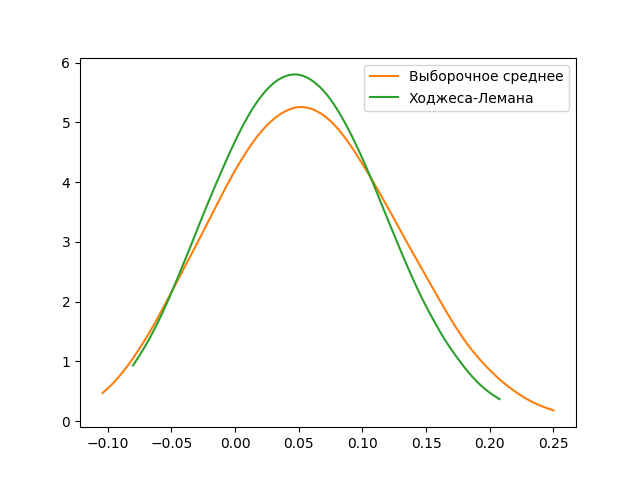


Рисунок 2 – Оценка плотности при объеме выборки = 500 и кол-во ре-выборок = 1000

При объеме исходной выборки равной 50 мы видим, что оценка Ходжеса-Лемана имеет меньшую дисперсию и СКО, но имеет больший квадрат смещения чем у оценки выборочного среднего.

При увеличении объема исходной выборки до 500 оценка Ходжеса-Лемана имеет меньшие значения дисперсии, квадрата смещения и СКО, чем у оценки выборочного среднего.

Так же у обоих оценок уменьшаются значения каждой из характеристик с увеличением объема исходной выборки значений. Это вызвано несколькими факторами:

1. Закон больших чисел.

Согласно закону больших чисел, при увеличении объема выборки, среднее значение выборки (выборочное среднее) приближается к математическому ожиданию генеральной совокупности. Это уменьшает смещение оценок.

1. Дисперсия выборочного среднего.

Дисперсия выборочного среднего обратно пропорциональна объему выборки. Соответственно с увеличением объема выборки дисперсия уменьшается.

1. Среднеквадратическая ошибка (СКО).

Среднеквадратическая ошибка состоит из двух компонентов: смещения и дисперсии. Так как при увеличении объема выборки дисперсия оценок и смещение уменьшаются, то и СКО тоже уменьшается.

## 3.2 Исследование оценок при наличии симметричных выбросов

Таблица 2 - Исследование оценок на основе генератора псевдослучайных величин модели выбросов Тьюки Симметричные выбросы

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Объем исходной выборки | Сдвиг | Масштаб | Дисперсия для выборочной средней оценки | Дисперсия для оценки Ходжеса-Лемана | Квадрат смещения для выборочной средней оценки | Квадрат смещения для оценки Ходжеса-Лемана |
| 50 | 0 | 1 | 0.04767 | 0.02609 | 0.06621 | 0.02606 |
| 500 | 0 | 1 | 0.00951 | 0.00294 | 0.00309 | 0.00074 |

*Окончание таблицы 2*

|  |  |  |  |
| --- | --- | --- | --- |
| СКО для выборочной средней оценки | СКО для оценки Ходжеса-Лемана | Соотношение СКО | График рисунок |
| 0.11389 | 0.05216 | 2.183 | Рисунок 3 |
| 0.01260 | 0.00369 | 3.755 | Рисунок 4 |

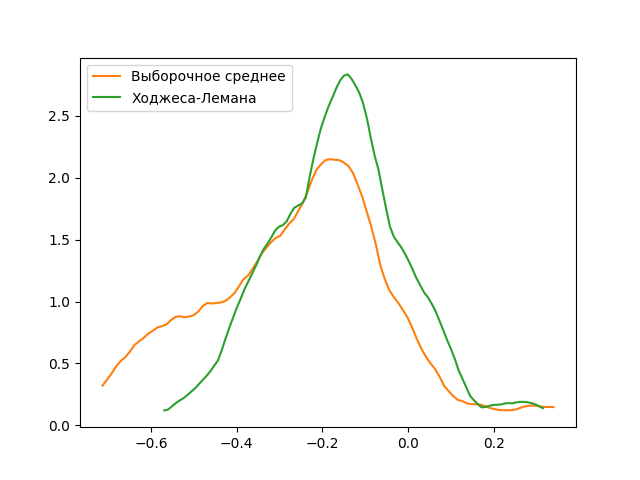


Рисунок 3 – Оценка плотности при объеме выборки = 50 и кол-во ре-выборок = 100 (с симметричными выбросами)

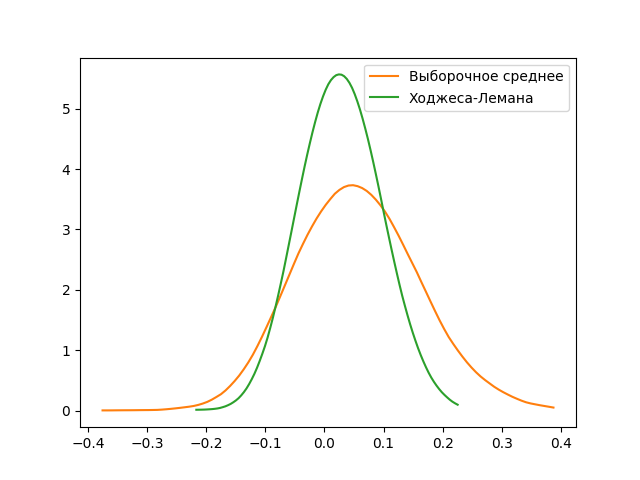


Рисунок 4 – Оценка плотности при объеме выборки = 500 и кол-во ре-выборок = 1000 (с симметричными выбросами)

При объеме исходной выборки равной 50 мы видим, что оценка Ходжеса-Лемана имеет меньшую дисперсию, СКО и квадрат смещения чем у оценки выборочного среднего.

При увеличении объема исходной выборки до 500 оценка Ходжеса-Лемана имеет меньшие значения дисперсии, квадрата смещения и СКО, чем у оценки выборочного среднего.

Можем заметить увеличение значений дисперсии для каждой из оценок в сравнение с исследованием без выбросов. Это вызвано

Таблица 3 - Исследование оценок на основе генератора псевдослучайных величин модели выбросов Тьюки Асимметричные выбросы

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Объем исходной выборки | Сдвиг | Масштаб | Дисперсия для выборочной средней оценки | Дисперсия для оценки Ходжеса-Лемана | Квадрат смещения для выборочной средней оценки | Квадрат смещения для оценки Ходжеса-Лемана |
| 50 | 0 | 1 | 0.47563 | 0.02893 | 1.85287 | 0.05389 |
| 500 | 0 | 1 | 0.03738 | 0.00353 | 1.90844 | 0.03508 |

|  |  |  |  |
| --- | --- | --- | --- |
| СКО для выборочной средней оценки | СКО для оценки Ходжеса-Лемана | Соотношение СКО | График рисунок |
| 2.32850 | 0.08282 | 28.115 | Рисунок 5 |
| 1.94582 | 0.03862 | 50.384 | Рисунок 6 |

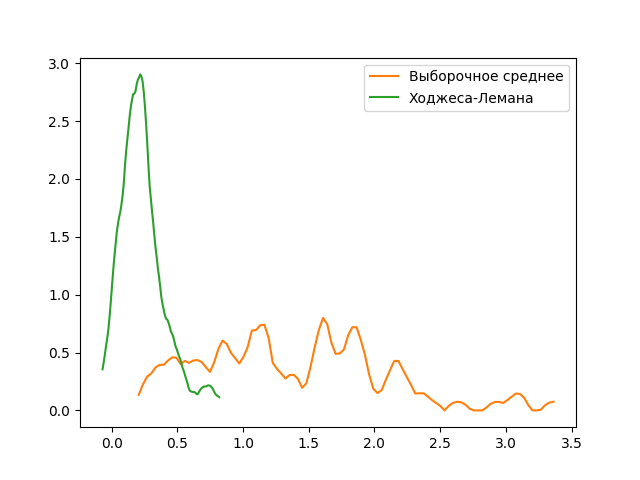


Рисунок 5 – Оценка плотности при объеме выборки = 50 и кол-во ре-выборок = 100 (с асимметричными выбросами)

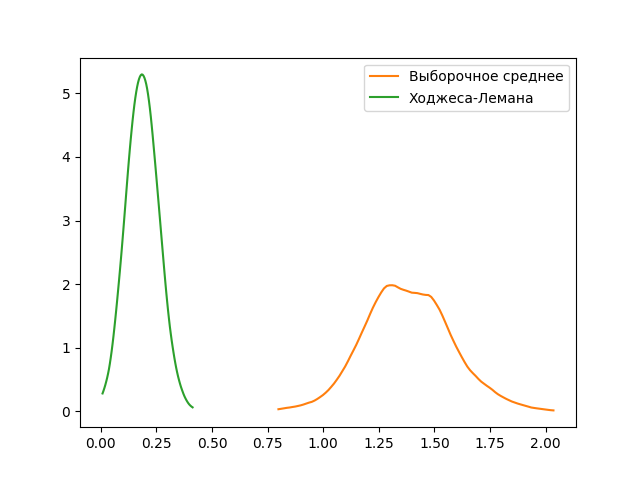


Рисунок 6 – Оценка плотности при объеме выборки = 500 и кол-во ре-выборок = 1000 (с асимметричными выбросами)

## Заключение

В результате работы мы провели исследование двух оценок для параметра сдвига в распределении Лапласа и выяснили, что оценка Ходжеса-Лемана лучше, чем выборочное среднее.

## Список использованных источников

1. Симахин, В. А. Теория вероятностей и математическая статистика: методические рекомендации к выполнению курсовых работ для студентов направления 09.03.04, 090303.65 / В. А. Симахин, О. С. Черепанов. – Курган: ПОАС, 2015. – 42с. – URL: http://dspace.kgsu.ru/xmlui/bitstream/handle/123456789/4141/%d0%a1%d0%b8%d0%bc%d0%b0%d1%85%d0%b8%d0%bd-%d0%92%d0%90\_2015\_%d0%9c%d0%a3.pdf?sequence=1&isAllowed=y (дата обращения 30.05.2024). – Текст: электронный