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 09.03.03. КР24.712143

Выполнил студент гр. ИТ-0930222 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / Чернов С.Ю./

подпись, дата

Направление 09.03.03 - «Прикладная информатика»

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

подпись, дата

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

2024

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc167999889)

[1 ПОСТАНОВКА ЗАДАЧИ 4](#_Toc167999890)

[2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 5](#_Toc167999891)

[2. 1. Равномерное распределение 5](#_Toc167999892)

[2. 2. Оценка Ходжеса-Лемана 5](#_Toc167999893)

[2. 3. Оценка на порядковых статистиках 5](#_Toc167999894)

[2. 4. Бутстреп 6](#_Toc167999895)

[2. 5. «Наивный» бутстреп-метод 6](#_Toc167999896)

[2. 6. Выбросы 7](#_Toc167999897)

[2. 7. Описание модели выбросов Тьюки 7](#_Toc167999898)

[2. 8. Генератор псевдослучайных величин модели выбросов Тьюки 8](#_Toc167999899)

[3 ИССЛЕДОВАНИЕ ОЦЕНОК 9](#_Toc167999900)

[ЗАКЛЮЧЕНИЕ 22](#_Toc167999901)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 23](#_Toc167999902)

# ВВЕДЕНИЕ

Теория вероятностей и математическая статистика являются важнейшими областями математики, находящими широкое применение в различных сферах деятельности человека. Изучение этих дисциплин помогает нам понять, какие результаты можно ожидать при проведении определенных экспериментов или исследований.

В данной курсовой работе мы будем рассматривать тему проведения исследования и сравнения оценки Ходжеса-Лемана и оценки на порядковых статистиках для параметра сдвига равномерного распределения.

Цель данной работы заключается в том, чтобы проанализировать свойства оценки Ходжеса-Лемана и оценки на порядковых статистиках при использовании их как оценок параметра сдвига равномерного распределения. Для достижения этой цели мы будем использовать теоретические выкладки и численные эксперименты.

Таким образом, данная курсовая работа имеет большое значение для понимания особенностей равномерного распределения и эффективности оценок параметра сдвига данного распределения.

**1 ПОСТАНОВКА ЗАДАЧИ**

**Вариант 17**

На основе бутстреп-метода провести исследование и сравнение оценок Ходжеса-Лемана – параметра положения равномерного распределения:

1. Реализовать генераторы псевдослучайных чисел из распределения и сформировать выборки заданного объема;
2. Построить оценки плотности вероятностей случайной величины *X* по исходным выборкам;
3. Построить оценки Ходжеса-Лемана и на порядковых статистиках.

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) и доверительные интервалы оценок.
4. Провести сравнение.
5. Сделать выводы.

# 2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

**2. 1. Равномерное распределение**

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.

Плотность вероятности равномерного распределения может быть записана как:

Интегрируя определённую выше плотность, получаем функцию распределения:

**2. 2. Оценка Ходжеса-Лемана**

Определим по выборке  набор из средних вида, называемых средними Уолша. Оценка Ходжеса-Лемана определяется как медиана средних Уолша, то есть медиана ряда. Следует отметить высокую устойчивость оценки к наличию аномальных наблюдений в выборке.

**2. 3. Оценка на порядковых статистиках**

Оценка параметра положения на порядковых статистиках определяется следующим выражением:

,

где – r-я порядковая статистика.

**2. 4. Бутстреп**

Бутстреп в статистике - это метод для оценки статистической неопределённости и построения доверительных интервалов для параметров модели или статистик на основе анализа повторных выборок из исходной выборки. Бутстреп широко применяется для оценки доверительных интервалов, оценки стандартных ошибок, проверки гипотез и других статистических процедур. Он особенно полезен в ситуациях, когда аналитические методы неприменимы или сложны. Бутстреп также позволяет учесть особенности исходных данных и моделировать более реалистичные распределения.

**2. 5. «Наивный» бутстреп-метод**

Бутстреп-метод был предложен Эфроном как некоторое обобщение алгоритма складного ножа. Основная идея бутстрепа состоит в том, чтобы методом статистических испытаний многократно извлекать повторные выборки из исходных наблюдений. Пусть имеется исходная выборка. Требуется сгенерировать *M* ре-выборок. На основе генератора псевдослучайных величин, равномерно распределенных на, генерируются случайные числа, соответствующие индексам элементов исходной выборки, которые помещаются в ре-выборку. Таким образом, можно сформировать любое, сколь угодно большое, число ре-выборок. При этом в одной ре-выборке какие-то элементы исходной выборки могут повторяться, тогда, как другие элементы могут отсутствовать.

В наивном бутстреп-методе применяется простая схема выборки с возвращением. Это означает, что каждый элемент в выборке может быть выбран несколько раз или вовсе отсутствовать. На каждой итерации метода бутстрепа создается ре-выборка, состоящая из случайных выбранных элементов из исходной выборки. Затем на этой ре-выборке снова оцениваются параметры или статистики интересующего нас распределения.

Поскольку выборка создается путем случайного выбора с возвращением, на каждой итерации некоторые элементы могут повторяться, а некоторые могут быть пропущены. Это приводит к тому, что на каждой итерации оценки параметров или статистик будут немного различаться.

Наивный бутстреп-метод прост в реализации, но может дать неточные оценки, особенно если исходные данные содержат выбросы или имеют нестандартное распределение. Он широко применяется в статистике и машинном обучении для оценки доверительных интервалов, стандартных ошибок и других статистических характеристик

**2. 6. Выбросы**

Выброс (в статистике) — это наблюдение, которое существенно отличается от остальных значений в выборке. Оно может быть значительно меньше или больше остальных значений или иметь непривычно высокое или низкое значение в контексте остальных данных.

Выбросы могут возникать по разным причинам, таким как ошибки измерений, случайные флуктуации или наличие редких событий. Они могут искажать статистические характеристики выборок.

**2. 7. Описание модели выбросов Тьюки**

Модель выбросов Тьюки, также известная как метод Тьюки или критерий Тьюки, является статистическим подходом к определению и обработке выбросов в данных. Она была разработана американским статистиком Джоном Тьюки в 1970-х годах и является одним из наиболее популярных методов для обнаружения и устранения выбросов.

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

**2. 8. Генератор псевдослучайных величин модели выбросов Тьюки**

Мы хотим создать генератор псевдослучайных чисел, который будет генерировать числа согласно распределению модели Тьюки. У нас есть два генератора чисел: который генерирует числа с распределением и, который генерирует числа с распределением. Наш новый генератор должен выбирать, из какого распределения сгенерировать число на каждом шаге его работы.

Для этого мы используем дополнительный генератор случайных чисел u, который генерирует числа из равномерного распределения. Если u генерирует число из диапазона, то мы используем генератор для создания случайного числа. В противном случае, если u генерирует число вне этого диапазона, мы используем генератор.

Таким образом, на каждом шаге генерации мы используем случайное число u, чтобы определить, какой генератор использовать. Это позволяет нам создавать псевдослучайные числа, соответствующие распределению модели Тьюки [1].

**3 ИССЛЕДОВАНИЕ ОЦЕНОК**

Для исследования оценок было реализовано следующие:

1. Генератор псевдослучайных величин;
2. Методы, рассчитывающие оценку Ходжеса-Лемана и оценку на порядковых статистиках;
3. «Наивный» бутстреп-метод;
4. Генератор псевдослучайных величин модели выбросов Тьюки.

Результаты исследование представлены в таблицах и рисунках 1-12. Параметры равномерного распределения: a = 0, b = 1. Распределение у симметричных и асимметричных выбросов так же равномерное. Параметры распределения симметричных выбросов: a = -1, b = 2, а у асимметричных: a = 5, b = 6. Процент выбросов – 10%.

Таблица 1 – Исследование оптимального порядка на основе генератора псевдослучайных величин, N = 50

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 0.0000001337 | 0.0001397620 | 0.0001398957 |
| 0.1 | 0.0000052045 | 0.0010531092 | 0.0010583138 |
| 0.2 | 0.0000034076 | 0.0023992702 | 0.0024026777 |
| 0.3 | 0.0000110025 | 0.0030479304 | 0.0030589329 |
| 0.4 | 0.0000320013 | 0.0038048777 | 0.0038368790 |
| 0.5 | 0.0000611481 | 0.0041704891 | 0.0042316372 |

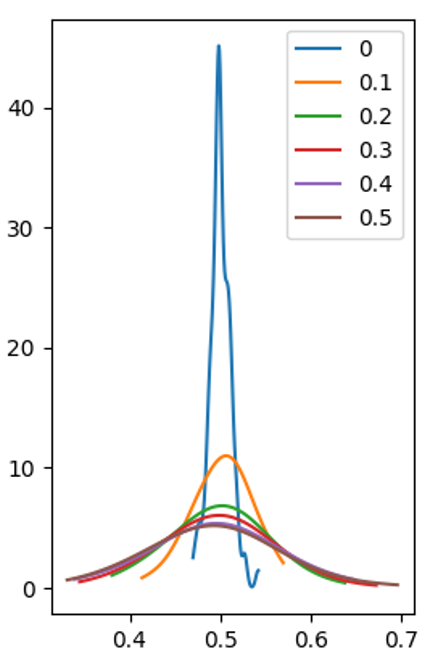


Рисунок 1 – Оценка плотности порядковых статистик при N = 50, без выбросов

Из таблицы 1 и рисунка 1 можно сделать вывод, что самым оптимальной порядковой статистикой является 0-я порядковая статистика, которая является оценкой min-max, потому что параметр сдвига зависит от a –минимального значения в выборке и от b – самого максимального значения в выборке.

Таблица 2 – Исследование оценок на основе генератора псевдослучайных величин, N = 50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0000002698 | 0.0016638104 | 0.0016640801 | 9.693425797 |
| На порядковых статистиках  (r = 0) | 0.0000000010 | 0.0001716700 | 0.0001716710 |

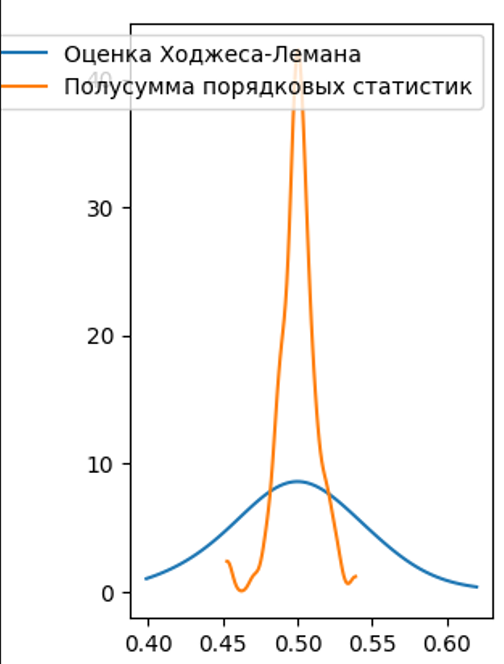


Рисунок 2 – Оценка плотности оценок при N = 50, без выбросов

Из таблицы 2 и рисунка 2 видно, что наиболее эффективной для оценки параметра сдвига будет оценка на порядковых статистиках с порядком r = 0, засчет наименьшей дисперсии. Доверительный интервал оценки Ходжеса-Лемана – [0.420605745, 0.559302088], оценки на порядковых статистиках – [0.4809684817, 0.522240421].

Таблица 3 – Исследование оптимального порядка на основе генератора псевдослучайных величин, N = 500

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 0.0000000010 | 0.0000018616 | 0.0000018626 |
| 0.1 | 0.0000000017 | 0.0001065665 | 0.0001065682 |
| 0.2 | 0.0000000254 | 0.0002042663 | 0.0002042917 |
| 0.3 | 0.0000001196 | 0.0003101792 | 0.0003102988 |
| 0.4 | 0.0000000001 | 0.0004070373 | 0.0004070375 |
| 0.5 | 0.0000002710 | 0.0005002432 | 0.0005005142 |

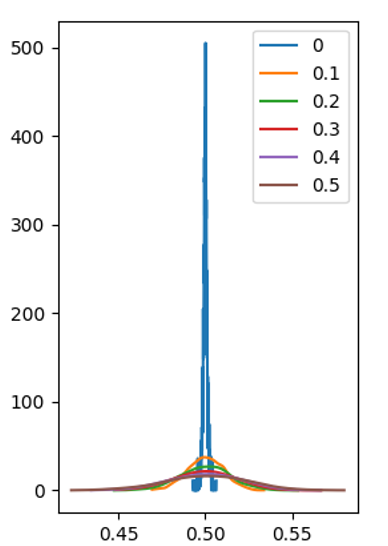


Рисунок 3 – Оценка плотности порядковых статистик при N = 500, без выбросов

Из таблицы 3 и рисунка 3 видно, что самым оптимальной порядковой статистикой является 0-я порядковая статистика, которая является оценкой min-max, потому что параметр сдвига зависит от a –минимального значения в выборке и от b – самого максимального значения в выборке. Можно сделать вывод, что для равномерного распределения без выбросов на любых объемах выборке наиболее эффективной будет оценка на порядковых статистиках с порядком r = 0, то есть оценка min-max.

Таблица 4 – Исследование оценок на основе генератора псевдослучайных величин, N = 500

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0000000215 | 0.0001821160 | 0.0001821375 | 85.5828869 |
| На порядковых статистиках  (r = 0) | 0.0000000052 | 0.0000021230 | 0.0000021282 |

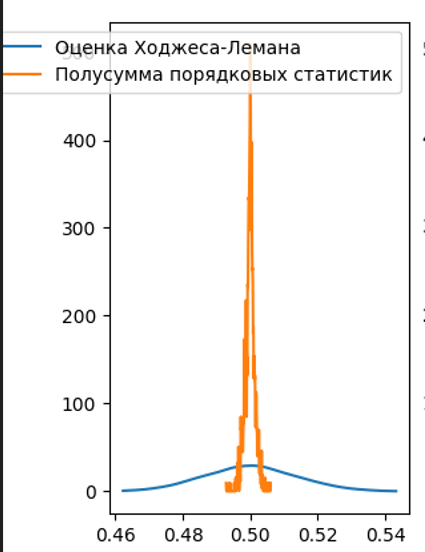


Рисунок 4 – Оценка плотности оценок при N = 500, без выбросов

Из таблицы 4 и рисунка 4 видно, что наиболее эффективной для оценки параметра сдвига будет оценка на порядковых статистиках с порядком r = (N-1)/N, потому что дисперсия и сдвиг у данной оценки меньше. Можно сделать вывод, что что для равномерного распределения без выбросов на любых объемах выборке наиболее эффективной будет оценка на порядковых статистиках, чем оценка Ходжеса-Лемана. Доверительный интервал оценки Ходжеса-Лемана – [0.4790328041, 0.520526225], оценки на порядковых статистиках – [0.497797646, 0.50224012].

Таблица 5 – Исследование оптимального порядка на основе генератора псевдослучайных величин модели Тьюки, симметричные выбросы, N = 50

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 0.0000198715 | 0.0094799837 | 0.0094998552 |
| 0.1 | 0.0000001511 | 0.0002927749 | 0.0002929261 |
| 0.2 | 0.0000007999 | 0.0005598641 | 0.0005606640 |
| 0.3 | 0.0000018177 | 0.0007992999 | 0.0008011176 |
| 0.4 | 0.0000012499 | 0.0010956142 | 0.0010968641 |
| 0.5 | 0.0000000433 | 0.0013056238 | 0.0013056671 |

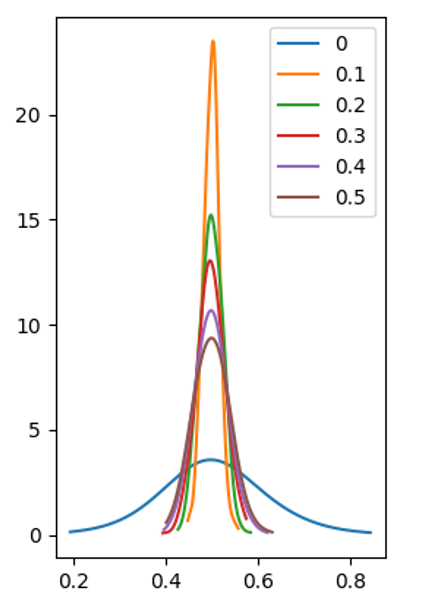


Рисунок 5 – Оценка плотности порядковых статистик при N = 50, симметричные выбросы

Из таблицы 5 и рисунка 5 видно, что оценка min-max не устойчива на симметричных выбросах. Наиболее оптимальной оценкой на порядковых статистиках является оценка с порядком большим, чем половина от процента выбросов, потому что выбросы распределяются равномерно на два противоположных конца. Поэтому наиболее оптимальная оценка на порядковых статистиках будет с порядком r = 0.1, в данном случае.

Таблица 6 – Исследование оценок на основе генератора псевдослучайных величин модели Тьюки, симметричные выбросы, N = 50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0000003534 | 0.0006624898 | 0.0006628432 | 1.0402965804 |
| На порядковых статистиках  (r = 0.2) | 0.0000000169 | 0.0003006806 | 0.0003006975 |

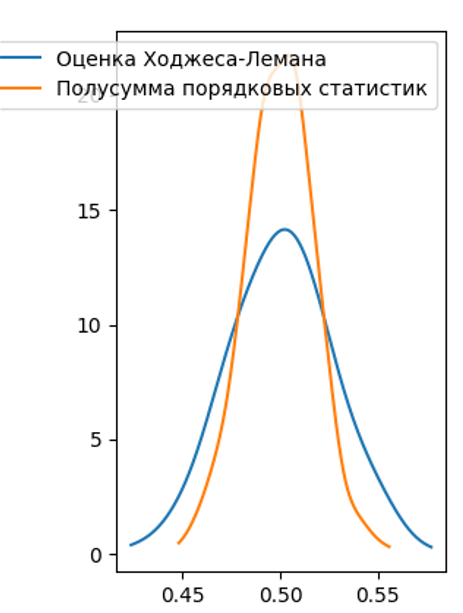


Рисунок 6 – Оценка плотности оценок при N = 50, симметричные выбросы

Из таблицы 6 и рисунка 6 видно, что оценка Ходжеса-Лемана не сильно уступает оценке на порядковых статистиках с порядком r = 0.1, потому что эти две оценки устойчивы к выбросам. Доверительный интервал оценки Ходжеса-Лемана – [0.4563606, 0.54455711], оценки на порядковых статистиках – [0.469036327, 0.526620248].

Таблица 7 – Исследование оптимального порядка на основе генератора псевдослучайных величин модели Тьюки, симметричные выбросы, N = 500

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 0.0000002353 | 0.0016710309 | 0.0016712661 |
| 0.1 | 0.0000001141 | 0.0001194552 | 0.0001195694 |
| 0.2 | 0.0000003195 | 0.0002396354 | 0.0002399548 |
| 0.3 | 0.0000002222 | 0.0003636708 | 0.0003638930 |
| 0.4 | 0.0000000824 | 0.0004551960 | 0.0004552784 |
| 0.5 | 0.0000002811 | 0.0005611688 | 0.0005614499 |

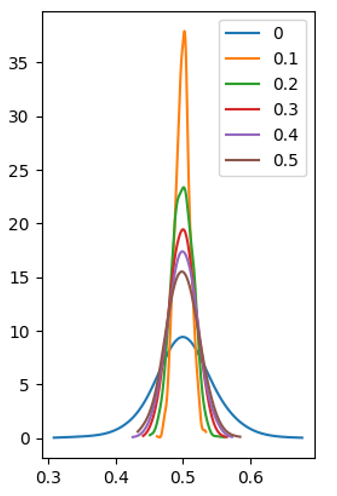


Рисунок 7 – Оценка плотности порядковых статистик при N = 500, симметричные выбросы

Из таблицы 7 и рисунка 7 видно, что так же, как и на малых объемах выборки, наиболее эффективной оценкой на порядковых статистиках является оценка с порядком r = 0.1, то есть большим, чем половина процента выбросов в выборке.

Таблица 8 – Исследование оценок на основе генератора псевдослучайных величин модели Тьюки, симметричные выбросы, N = 500

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0000000390 | 0.0002175796 | 0.0002176186 | 1.859762319 |
| На порядковых статистиках  (r = 0.1) | 0.0000000060 | 0.0001170081 | 0.0001170142 |

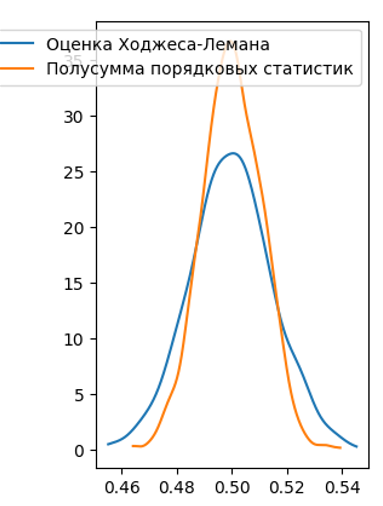


Рисунок 8 – Оценка плотности оценок при N = 500, симметричные выбросы

Из таблицы 8 и рисунка 8 видно, что хоть СКО обоих оценок мал, оценка на порядковых статистиках с порядком r = 0.1 на 86%, чем оценка Ходжеса-Лемана, засчет меньшей дисперсии. Доверительный интервал оценки Ходжеса-Лемана – [0.4756571234, 0.525541654], оценки на порядковых статистиках – [0.482429551, 0.51727751]].

Таблица 9 – Исследование оптимального порядка на основе генератора псевдослучайных величин модели Тьюки, асимметричные выбросы, N = 50

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 5.8969505441 | 0.0055298658 | 5.9024804099 |
| 0.1 | 0.0030679222 | 0.0008621290 | 0.0039300511 |
| 0.2 | 0.0029040906 | 0.0020598634 | 0.0049639541 |
| 0.3 | 0.0024481491 | 0.0031901707 | 0.0056383198 |
| 0.4 | 0.0022963943 | 0.0042290498 | 0.0065254441 |
| 0.5 | 0.0024835198 | 0.0049357190 | 0.0074192389 |

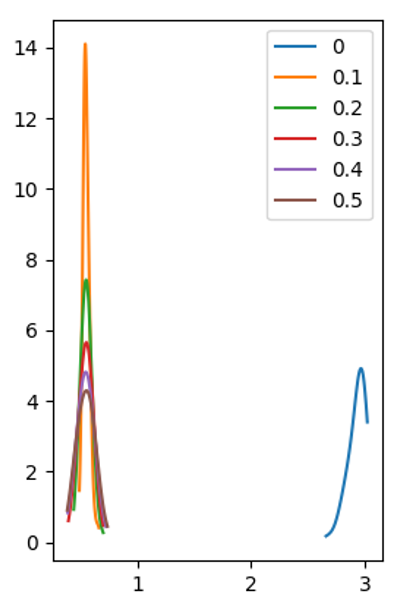


Рисунок 9 – Оценка плотности порядковых статистик при N = 50, асимметричные выбросы

Из таблицы 9 и рисунка 9 видно, что наиболее эффективной оценкой на порядковых статистиках для параметра сдвига с асимметричными выбросами будет оценка с порядком большим или равным, чем процент выбросов. В данном случае, порядок r = 0.1.

Таблица 10 – Исследование оценок на основе генератора псевдослучайных величин модели Тьюки, асимметричные выбросы, N = 50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0053697040 | 0.0017600987 | 0.0071298028 | 1.83042819 |
| На порядковых статистиках  (r = 0.2) | 0.0031389190 | 0.0007562367 | 0.0038951557 |

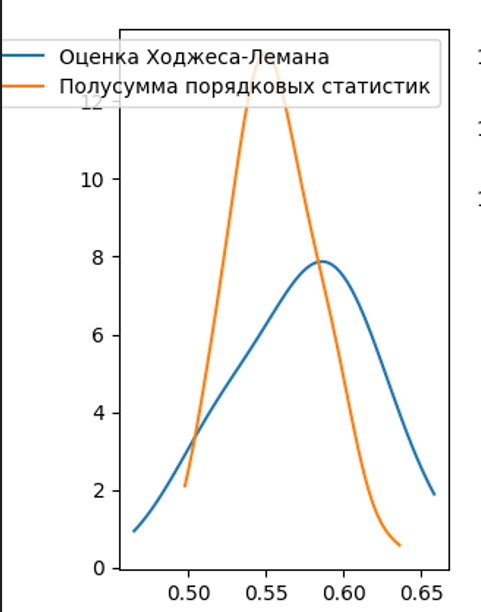


Рисунок 10 – Оценка плотности оценок при N = 50, асимметричные выбросы

Из таблицы 10 и рисунка 10 можно сделать вывод, что из-за того, что оценка Ходжеса-Лемана так же устойчива к выбросам, как и оценка на порядковых статистиках, соотношение СКО данных оценок приблизительно равно. Доверительный интервал оценки Ходжеса-Лемана – [0.482562978, 0.622698005], оценки на порядковых статистиках – [0.520220255, 0.592576732].

Таблица 11 – Исследование оптимального порядка на основе генератора псевдослучайных величин модели Тьюки, асимметричные выбросы, N = 500

|  |  |  |  |
| --- | --- | --- | --- |
| r-я порядковая статистика | Смещение | Дисперсия | СКО |
| 0 | 6.2058889569 | 0.0000931753 | 6.2059821323 |
| 0.1 | 0.0030650855 | 0.0000598876 | 0.0031249731 |
| 0.2 | 0.0030969930 | 0.0001861389 | 0.0032831319 |
| 0.3 | 0.0031798571 | 0.0003281538 | 0.0035080109 |
| 0.4 | 0.0032201342 | 0.0004635883 | 0.0036837226 |
| 0.5 | 0.0031837894 | 0.0006006697 | 0.0037844591 |

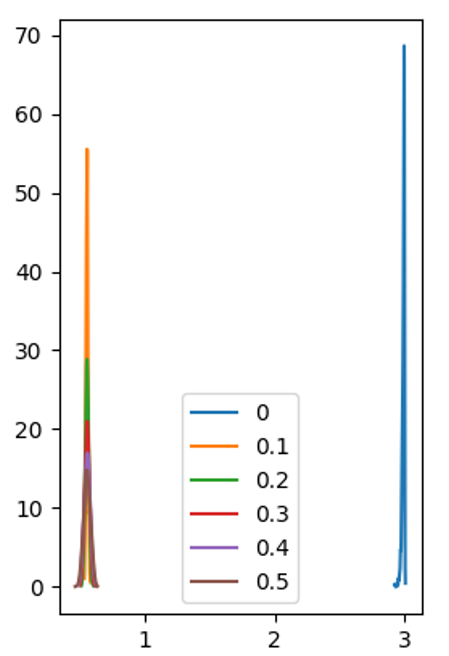


Рисунок 11 – Оценка плотности порядковых статистик при N = 500, асимметричные выбросы

Из таблицы 11 и рисунка 11 видно, что наиболее эффективной оценкой на порядковых статистиках для параметра сдвига с асимметричными выбросами будет так же оценка на порядковых статистиках с порядком большим или равным, чем процент выбросов, как и в случае, когда объем выборки мал.

Таблица 12 – Исследование оценок на основе генератора псевдослучайных величин модели Тьюки, асимметричные выбросы, N = 500

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Оценка | Смещение | Дисперсия | СКО | СКО1/СКО2 |
| Ходжеса-Лемана | 0.0040222370 | 0.0002238541 | 0.0042460911 | 1.338677134 |
| На порядковых статистиках  (r = 0.2) | 0.0031164530 | 0.0000554030 | 0.0031718560 |

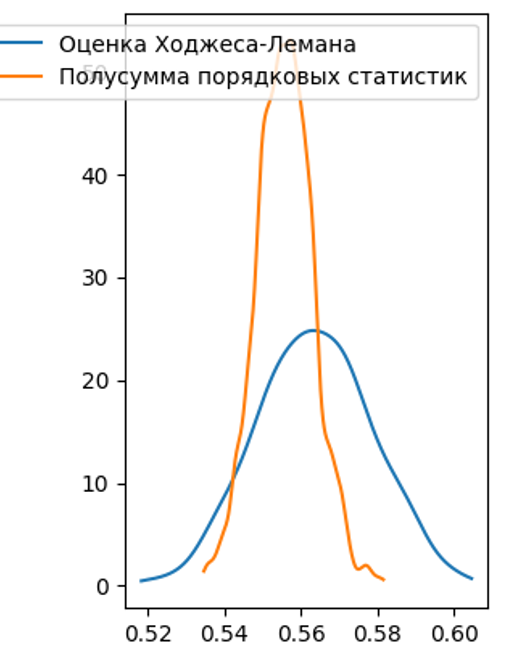


Рисунок 12 – Оценка плотности оценок при N = 500, асимметричные выбросы

Из таблицы 12 и рисунка 12 можно сделать вывод, что из-за того, что оценка Ходжеса-Лемана так же устойчива к выбросам, как и оценка на порядковых статистиках, соотношение СКО данных оценок приблизительно равно. Однако, процент различия стал меньшим на большом объеме выборки, из-за того, что в данном случае СКО дает более точные значения. Доверительный интервал оценки Ходжеса-Лемана – [0.539220928, 0.58738968], оценки на порядковых статистиках – [0.543297116, 0.569015236].

**ЗАКЛЮЧЕНИЕ**

В ходе исследования, мы сравнили две оценки для параметра сдвига равномерного распределении и пришли к выводу, что для данных без выбросов лучше использовать порядковую статистику с порядком r = 0, то есть оценку min-max. На данных с симметричными выбросами лучше использовать порядковую статистику с порядком r = 0.1 Это объясняется тем, что процент выбросов делится поровну на противоположные концы выборки при симметричных выбросах и порядок r надо выбирать так, чтобы он был больше чем половина процента выбросов. При асимметричных выбросах предпочтительнее использовать, как в случае с симметричными выбросами, оценку на порядковых статистиках с порядком r = 0.1, чем оценку Ходжеса-Лемана.

**СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ**

1. Математическая статистика [Электронный ресурс] / Мудл сайт: <https://moodle.ocherepanov.ru/course/view.php?id=2>