МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)

Лабораторная работа 3
по дисциплине
«Статистика для анализа данных»

Выполнили:
Трифонов Василий Максимович
гр. J3111
ИСУ 467758,
Соловьев Матвей Михайлович
гр. J3111
ИСУ 467551,
Ежов Дмитрий Александрович
гр. J3111
ИСУ 471242,

Отчет сдан: 13.05.2025

Санкт-Петербург 2025

1. Ход выполнения работы

В рамках данной лабораторной работы было выполнено исследование бутстрап-оценок статистических характеристик выборки и построение доверительных интервалов с использованием бутстрап-метода. Работа включала следующие основные этапы:

1. Генерация выборки из непрерывного распределения и расчет базовых точечных оценок с их сравнением с теоретическими значениями. Визуализация распределения данных с помощью гистограммы и ядерной оценки плотности (KDE).
2. Реализация алгоритма бутстрапа и получение бутстрап-распределений для выборочного среднего, медианы, дисперсии и интерквартильного размаха (IQR). Визуализация полученных распределений.
3. Построение бутстрап-доверительных интервалов для среднего и медианы методом процентилей при различных уровнях доверия и их визуализация.
4. Исследование влияния объема исходной выборки (N) и числа бутстрап-итераций (B) на ширину доверительных интервалов среднего.
5. Эмпирическая проверка покрытия 95%-ных доверительных интервалов среднего для данных из стандартного нормального распределения $N(0, 1)$ при различных комбинациях N и B.

2. Основная часть

2.1. Генерация данных и базовые оценки

2.1.1. Создание выборки

На первом этапе была сгенерирована выборка объемом $N = 500$ из непрерывного распределения. В качестве распределения было выбрано нормальное распределение с параметрами $\mu = 10$ и $\sigma = 2$, т.е. $X \sim N(10, 2)$.

2.1.2. Расчет теоретических значений статистик

На основе сгенерированной выборки были рассчитаны точечные оценки основных статистик: выборочное среднее, медиана, выборочная дисперсия и интерквартильный размах (IQR). Для сравнения были вычислены теоретические значения этих статистик для выбранного нормального распределения.

- Теоретическое среднее $\mu = 10$.
- Теоретическая медиана также равна $\mu = 10$ для нормального распределения.
- Теоретическая дисперсия $\sigma^2 = 2^2 = 4$.
- Теоретический IQR для нормального распределения с параметрами μ и σ равен $2 \times \Phi^{(-1)}(0.75) \times \sigma$, где $\Phi^{(-1)}(0.75)$ - 75-й процентиль стандартного нормального распределения (приблизительно 0.6745). Таким образом, теоретический IQR $\approx 2 \times 0.6745 \times 2 \approx 2.698$.

2.1.3. Расчет эмпирических значений статистик

Полученные точечные оценки и их сравнение с теоретическими значениями представлены ниже :

- Выборочное среднее: 9.9745 (Разница с теоретическим: 0.0255)
- Выборочная медиана: 10.1616 (Разница с теоретической: 0.1616)
- Выборочная дисперсия: 3.7611 (Разница с теоретической: 0.2389)
- Выборочный IQR: 2.4959 (Разница с теоретическим: 0.2020)

2.1.4. Гистограммы и KDE

Далее была построена гистограмма сгенерированных данных с наложением ядерной оценки плотности (KDE) для различного числа бинов (10, 30, 506 100).

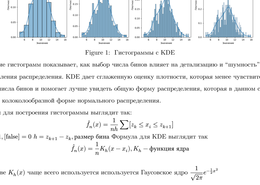

Figure 1: Гистограммы с KDE

Сравнение гистограмм показывает, как выбор числа бинов влияет на детализацию и "шумность" представления распределения. KDE дает сглаженную оценку плотности, которая менее чувствительна к выбору числа бинов и помогает лучше увидеть общую форму распределения, которая в данном случае близка к колоколообразной форме нормального распределения.

Формула для построения гистограммы выглядит так:

$$\hat{f}_n(x) = \frac{1}{nh} \sum [\hat{z}_k \leq x_i \leq \hat{z}_{k+1}]$$

$[\text{true}] = 1, [\text{false}] = 0$ $h = \hat{z}_{k+1} - \hat{z}_k$, размер бина Формула для KDE выглядит так

$$\hat{f}_n(x) = \frac{1}{n} K_h(x - x_i), K_h - \text{функция ядра}$$

В качестве $K_h(x)$ чаще всего используется используется Гауссовское ядро $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

2.2. Бутстрап для точечных оценок

Суть бутстрэп-метода заключается в многократном пересчете статистики на случайных выборках с возвращением (бутстрэп-выборках), извлеченных из исходной выборки. Это позволяет аппроксимировать выборочное распределение данной статистики и, как следствие, строить доверительные интервалы для параметров генеральной совокупности, основываясь на данных исходной выборки.

Из исходной выборки объемом $N = 500$ было сгенерировано $B = 1000$ бутстрэп-выборок путем случайного извлечения элементов с возвращением. Для каждой бутстрэп-выборки были рассчитаны те же статистики: среднее, медиана, дисперсия и IQR.

В результате мы получили $B = 1000$ бутстрэп-оценок для каждой статистики. Распределение этих бутстрэп-оценок аппроксимирует выборочное распределение соответствующей статистики.

Были построены гистограммы бутстрэп-распределений для каждой статистики. Вертикальной линией на каждом графике отмечено значение статистики, рассчитанное на исходной выборке.

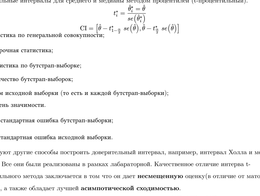

Figure 2: Гистограммы с KDE

Гистограммы бутстрэп-оценок демонстрируют разброс значений статистик, который можно ожидать при многократном извлечении выборок из генеральной совокупности (в данном случае, аппроксимируемой исходной выборкой). Ширина этих распределений связана со стандартной ошибкой статистики. Видно, что исходные выборочные оценки (красные линии) находятся в области высокой плотности бутстрэп-распределений, что ожидаемо.

2.3. Построение доверительных интервалов

С использованием бутстрэп-распределений, полученных на предыдущем шаге, были построены доверительные интервалы для среднего и медианы методом процентилей (t-процентильный).

$$t_i^* = \frac{\theta_i^* - \hat{\theta}}{se(\theta_i^*)}$$

$$\text{CI} = \left[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* \, se\left(\hat{\theta}\right), \hat{\theta} - t_{\frac{\alpha}{2}}^* \, se\left(\hat{\theta}\right)\right]$$

$\hat{\theta}$— статистика по генеральной совокупности;

$\hat{\theta}$— выборочная статистика;

θ_i^*— статистика по бутстрэп-выборке;

B—количество бутстрэп-выборок;

n—объем исходной выборки (то есть в каждой бутстрэп-выборки);

α—уровень значимости.

$se\left(\theta_i^*\right)$—стандартная ошибка бутстрэп-выборки;

$se\left(\hat{\theta}\right)$—стандартная ошибка исходной выборки.

Существуют другие способы построить доверительный интервал, например, интервал Холла и метод Эфрона. Все они были реализованы в разных лабораторной. Качественное отличие интерва t-процентильного метода заключается в том что он дает **несмещенную** оценку(в отличие от метода Эфрона), а также обладает лучшей **асимптотической сходимостью**.

Доверительные интервалы были рассчитаны для уровней доверия 90%, 95% и 99% t-процентильным методом.

Процентильный метод бутстрэпа основан на квантилях бутстрэп-распределения статистики. Для построения $(1-\alpha)$ доверительного интервала используются $\left(\frac{\alpha}{2}\right)$ и $\left(1-\frac{\alpha}{2}\right)$ процентили бутстрэп-оценок.

Полученные доверительные интервалы:

Для Среднего:

- 90% CI: [9.7683, 10.0426]
- 95% CI: [9.7353, 10.0738]
- 99% CI: [9.6723, 10.1343]

Для Медианы:

- 90% CI: [9.6388, 10.0038]
- 95% CI: [9.6027, 10.0090]
- 99% CI: [9.5360, 10.0581]

Визуализация этих интервалов:

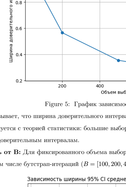

Figure 3: Визуализация CI для среднего

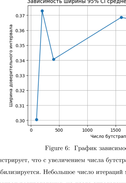

Figure 4: Визуализация CI для медианы

Как и ожидалось, с увеличением уровня доверия ширина доверительного интервала увеличивается. Это связано с тем, что для достижения более высокой уверенности в том, что интервал содержит истинное значение параметра, требуется более широкий диапазон значений.

2.4. Влияние объема выборки и числа итераций

Был исследован влияние объема исходной выборки (N) и числа бутстрэп-итераций (B) на ширину 95%-ного доверительного интервала для среднего (с использованием метода Эфрона).

Зависимость от N: Были сгенерированы выборки различных объемов ($N = [50, 100, 200, 500, 1000]$), и для каждой выборки построен 95% CI среднего при фиксированном числе бутстрэп-итераций $B = 1000$.

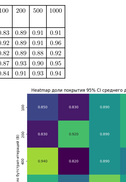

Figure 5: График зависимости ширины CI от N

График показывает, что ширина доверительного интервала уменьшается с увеличением объема выборки N. Это согласуется с теорией статистики: большие выборки дают более точные оценки, что приводит к более узким доверительным интервалам.

Зависимость от B: Для фиксированного объема выборки $N = 500$ были построены 95% CI среднего при различном числе бутстрэп-итераций ($B = [100, 200, 400, 1600, 3200]$).

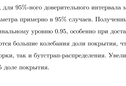

Figure 6: График зависимости ширины CI от B

График демонстрирует, что с увеличением числа бутстрэп-итераций B ширина доверительного интервала стабилизируется. Небольшое число итераций может привести к более изменчивым оценкам квантилей бутстрэп-распределения, но после определенного порога (в данном случае, около $B = 1000$) увеличение B уже не оказывает существенного влияния на ширину интервала, лишь повышая точность его границ.

2.5. Проверка покрытия интервалов

На заключительном этапе была проведена эмпирическая проверка покрытия 95%-ных доверительных интервалов среднего. Для этого были сгенерированы 100 выборок из стандартного нормального распределения $N(0, 1)$ для различных комбинаций объема выборки N и числа бутстрэп-итераций B. Для каждой выборки был построен 95% CI среднего (методом Эфрона), и было определено, содержит ли интервал истинное значение среднего $mu = 0$. Была рассчитана доля интервалов, содержащих $mu = 0$.

Результаты представлены в таблице и в виде тепловой карты:

B \ N	50	100	200	500	1000
100	0.85	0.83	0.89	0.91	0.91
200	0.83	0.92	0.92	0.91	0.96
400	0.94	0.82	0.89	0.88	0.92
1600	0.91	0.87	0.93	0.90	0.95
3200	0.83	0.84	0.91	0.93	0.94

Figure 7: Heatmap доли покрытия

Теоретически, для 95%-ного доверительного интервала мы ожидаем, что он будет покрывать истинное значение параметра примерно в 95% случаев. Полученные результаты показывают, что доля покрытия близка к номинальному уровню 0.95, особенно при достаточно больших значениях N и B. При малых N и B наблюдаются большие колебания доли покрытия, что связано с повышенной изменчивостью как исходной выборки, так и бутстрэп-распределения. Увеличение N и B приводит к более стабильной и близкой к 0.95 доле покрытия.

3. Заключение и выводы

- Бутстрап позволяет эмпирически оценить выборочное распределение статистики без строгих предположений о форме генеральной совокупности, что особенно ценно для сложных статистик или при неизвестном распределении данных.
- Процентильный метод бутстрапа является простым и интуитивно понятным способом построения доверительных интервалов на основе бутстрап-распределения.
- Ширина бутстрап-доверительного интервала существенно зависит от объема исходной выборки N, уменьшаясь с его увеличением.
- Число бутстрап-итераций B влияет на точность оценки границ доверительного интервала; увеличение B сверх определенного порога приводит к стабилизации ширины интервала.
- Эмпирическая проверка покрытия показала, что бутстрап-доверительные интервалы (методом процентилей) обеспечивают долю покрытия, близкую к номинальному уровню доверия, особенно при достаточно больших объемах выборки и числе бутстрап-итераций.

Бутстрап является мощным инструментом для статистического вывода, позволяющим получать надежные оценки и доверительные интервалы в ситуациях, когда классические аналитические методы неприменимы или сложны.