НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

Факультет технологий искусственного интеллекта

Лабораторная работа №3
по дисциплине "Статистика для анализа данных"

Выполнили:
Воронин Илья Андреевич, группа J3112
Вахменина Татьяна Михайловна, группа J3113

Проверил:
Свинцов М. В.

г. Санкт-Петербург
2025 г.

Содержание

1 Введение

В рамках данной лабораторной работы исследуется метод бутстрапа как мощный непараметрический инструмент для статистического вывода. Бутстрап позволяет оценить выборочное распределение статистики и построить доверительные интервалы без жестких предположений о форме распределения данных. Работа включает в себя генерацию данных, расчет точечных оценок, реализацию бутстрап-алгоритма, построение доверительных интервалов, а также анализ влияния объема выборки и числа бутстрап-итераций на точность оценок и проверку свойства покрытия доверительных интервалов.

2 Ход выполнения работы: Краткая информация о выполненных шагах

1. **Генерация данных и базовые оценки:** Сгенерирована выборка объема $N = 500$ из непрерывного распределения ($\mathcal{N}(10, 2^2)$). Вычислены точечные оценки (выборочное среднее, медиана, дисперсия, интерквартильный размах - IQR) и проведено их сравнение с теоретическими значениями. Построены гистограммы данных с наложением ядерной оценки плотности (KDE) для различных чисел бинов.

2. **Бутстрап для точечных оценок:** Реализован алгоритм бутстрапа с $B = 1000$ итерациями. Для каждой бутстрап-выборки вычислены те же статистики (среднее, медиана, дисперсия, IQR). Построены гистограммы распределений бутстрап-оценок, на которых вертикальными линиями отмечены исходные точечные оценки.

3. **Построение доверительных интервалов:** С использованием процентильного метода бутстрапа построены доверительные интервалы для среднего и медианы с уровнями доверия 90% ($\alpha = 0.1$), 95% ($\alpha = 0.05$) и 99% ($\alpha = 0.01$). Интервалы визуализированы на графиках распределения бутстрап-оценок.

4. **Влияние объема выборки и числа итераций:** Исследовано влияние изменения объема выборки N (для $B = 1000$) и числа бутстрап-итераций B (для $N = 500$) на ширину 95% доверительного интервала для среднего. Результаты визуализированы графиками зависимости ширины интервала от N и B.

5. **Проверка покрытия интервалов:** Для стандартного нормального распределения $\mathcal{N}(0, 1)$ проведена симуляция для оценки доли покрытия 95% доверительных интервалов среднего. Для различных комбинаций N и B сгенерировано по 100 выборок, и для каждой определено, содержит ли интервал истинное значение $\mu = 0$. Результаты представлены в таблице и на тепловой карте.

3 Основная часть: Описание шагов и анализ результатов

3.1 Генерация данных и базовые оценки

На первом этапе была сгенерирована синтетическая выборка данных для последующего анализа. Выборка была получена из нормального распределения, поскольку это одно из наиболее часто встречающихся и хорошо изученных распределений в статистике, позволяющее легко сравнивать эмпирические оценки с известными теоретическими значениями.

- **Выбранное распределение:** Нормальное распределение $\mathcal{N}(\mu, \sigma^2)$.
- **Параметры распределения:**
  - Математическое ожидание (истинное среднее): $\mu = 10$
  - Стандартное отклонение: $\sigma = 2$ (что соответствует дисперсии $\sigma^2 = 4$)
- **Объем выборки:** $N = 500$.

После генерации выборки были рассчитаны ее основные точечные оценки: выборочное среднее, выборочная медиана, выборочная дисперсия и выборочный интерквартильный размах (IQR). Эти эмпирические оценки были затем сравнены с их теоретическими значениями для $\mathcal{N}(10, 4)$.

Теоретические значения для $\mathcal{N}(\mu, \sigma^2)$:

- Среднее (μ): 10
- Медиана: 10 (для симметричных распределений, таких как нормальное)
- Дисперсия (σ^2): 4
- Интерквартильный размах (IQR): $2 \times \Phi^{-1}(0.75) \times \sigma \approx 2 \times 0.6745 \times 2 \approx 2.698$ (где $\Phi^{-1}(0.75)$ - 75-й процентиль стандартного нормального распределения).

Результаты расчетов точечных оценок:

- Выборочное среднее: `10.0137` (Теор: 10.0000)
- Выборочная медиана: `10.0256` (Теор: 10.0000)
- Выборочная дисперсия: `3.8437` (Теор: 4.0000)
- Выборочный IQR: `2.6742` (Теор: 2.6980)

Анализ: Полученные точечные оценки оказались очень близкими к соответствующим теоретическим значениям. Это является ожидаемым результатом для достаточно большого объема выборки ($N = 500$), так как с ростом объема выборки эмпирические статистики стремятся к истинным параметрам генеральной совокупности в силу Закона Больших Чисел. Небольшие расхождения между выборочными и теоретическими значениями объясняются естественной статистической вариабельностью выборки.

Визуализация данных: Гистограмма с KDE Для визуализации распределения данных была построена гистограмма с наложением ядерной оценки плотности (KDE). KDE - это непараметрический метод для оценки функции плотности вероятности случайной величины.

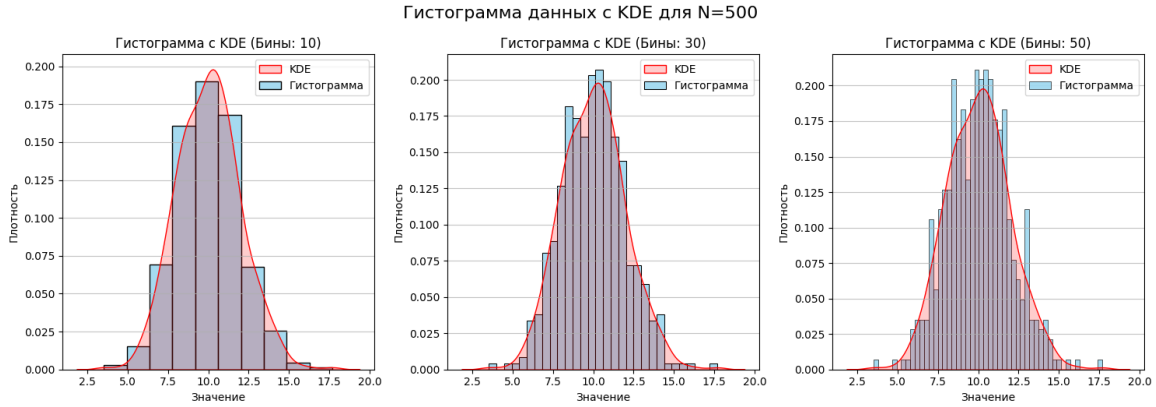

Рис. 1: Гистограмма данных с KDE для $N = 500$ с различным числом бинов.

Анализ сравнения KDE и гистограмм: На рисунке 1 представлены три гистограммы с KDE, построенные с различным количеством бинов (10, 30, 50).

- При **малом числе бинов (10)** гистограмма выглядит слишком сглаженной, теряя тонкие детали распределения. Она дает очень общее представление о форме данных, но может скрывать важные особенности.
- При **среднем числе бинов (30)** гистограмма обеспечивает хороший баланс между детализацией и сглаживанием, достаточно точно отражая форму распределения.
- При **большом числе бинов (50)** гистограмма становится "зубчатой отражая шум в данных и чувствительность к небольшим флуктуациям. Это может затруднить определение истинной формы распределения.

В отличие от гистограмм, KDE предоставляет гладкую, непрерывную оценку плотности, которая менее чувствительна к выбору параметров (таких как число бинов) и часто лучше аппроксимирует истинную форму распределения, особенно для непрерывных данных. KDE демонстрирует способность адаптироваться к форме данных, предоставляя более наглядное представление о плотности вероятности.

3.2 Бутстрап для точечных оценок

Бутстрап - это непараметрический метод повторной выборки, который позволяет оценить выборочное распределение статистики или ошибку оценки, когда аналитические методы недоступны или слишком сложны. Он основан на предположении, что исходная выборка является репрезентативной для генеральной совокупности, и, следовательно, многократная повторная выборка из этой выборки (с возвращением) может имитировать процесс получения множества выборок из генеральной совокупности.

Алгоритм бутстрапа:

1. Из исходной выборки объема N сгенерировать B новых выборок (бутстрап-выборок), каждая объемом N, путем выборки с возвращением.
2. Для каждой из B бутстрап-выборок вычислить интересующую статистику (например, среднее, медиану, дисперсию, IQR).
3. Полученные B значений статистики формируют эмпирическое бутстрап-распределение этой статистики.

В данной работе было сгенерировано $B = 1000$ бутстрап-выборок, и для каждой из них вычислены выборочное среднее, медиана, дисперсия и IQR.

Визуализация распределений бутстрап-оценок:

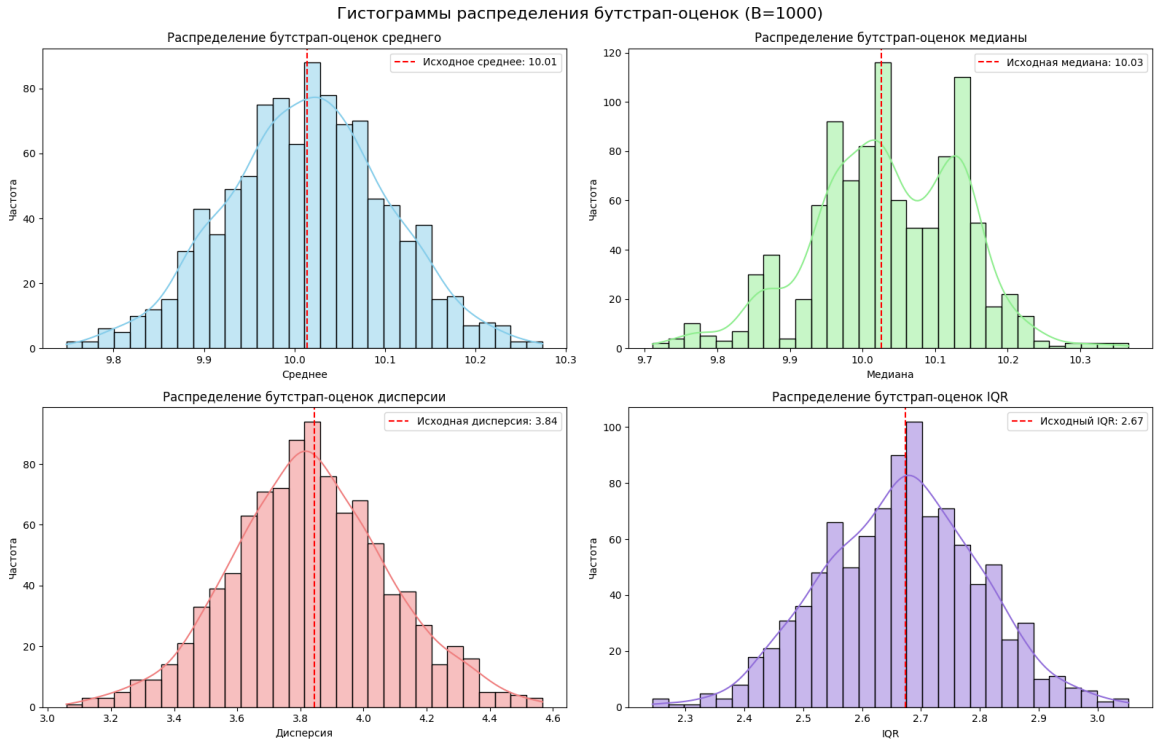

Рис. 2: Гистограммы распределения бутстрап-оценок для среднего, медианы, дисперсии и IQR ($B = 1000$). Красные пунктирные линии показывают исходные точечные оценки.

Анализ гистограмм бутстрап-оценок: На рисунке 2 показаны гистограммы распределений бутстрап-оценок для каждой из четырех статистик. Каждая гистограмма представляет собой эмпирическое распределение соответствующей статистики, полученное путем многократного бутстрапа.

- **Среднее и Медиана:** Распределения бутстрап-оценок среднего и медианы имеют приблизительно симметричную, колоколообразную форму, что согласуется с предсказаниями Центральной Предельной Теоремы (для среднего) и является характерным для оценок положения. Исходные точечные оценки (красные пунктирные линии) находятся вблизи центра этих распределений, что указывает на их состоятельность как оценок истинных параметров.

- **Дисперсия:** Распределение бутстрап-оценок дисперсии демонстрирует некоторую положительную скошенность (скошено вправо), что типично для оценок дисперсии и сигнализирует о том, что оценка дисперсии может быть более чувствительна к выбросам или асимметрии в данных. Тем не менее, исходная выборочная дисперсия также находится в области высокой плотности распределения.

- **IQR:** Распределение бутстрап-оценок IQR выглядит относительно симметричным, хотя и с небольшим правым скосом. Исходная оценка IQR также хорошо вписывается в это распределение.

Ширина и форма этих распределений дают ценную информацию о точности и неопределенности каждой оценки. Чем уже распределение, тем меньше стандартная ошибка статистики и, следовательно, тем точнее оценка.

3.3 Построение доверительных интервалов

Доверительный интервал предоставляет диапазон значений, который с определенной вероятностью (уровнем доверия) содержит истинное значение параметра генеральной сово-

купности. В данной работе использовался процентильный метод бутстрапа для построения доверительных интервалов.

Процентильный метод построения доверительного интервала для статистики θ:

1. Сгенерировать B бутстрап-оценок $\theta_1^*, \theta_2^*, ..., \theta_B^*$.
2. Отсортировать эти оценки по возрастанию.
3. Для построения $100(1-\alpha)\%$ доверительного интервала, найти $(\alpha/2)$-й и $(1-\alpha/2)$-й процентили отсортированных бутстрап-оценок. Интервал будет выглядеть как $[\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}]$.

Доверительные интервалы были построены для среднего и медианы с уровнями доверия 90% ($\alpha = 0.1$), 95% ($\alpha = 0.05$) и 99% ($\alpha = 0.01$).

Рассчитанные доверительные интервалы:

- **Для среднего:**
  - 90% ДИ: `[9.8687, 10.1590]`
  - 95% ДИ: `[9.8377, 10.1889]`
  - 99% ДИ: `[9.7900, 10.2364]`
- **Для медианы:**
  - 90% ДИ: `[9.8551, 10.1888]`
  - 95% ДИ: `[9.8218, 10.2110]`
  - 99% ДИ: `[9.7419, 10.3041]`

Визуализация доверительных интервалов:

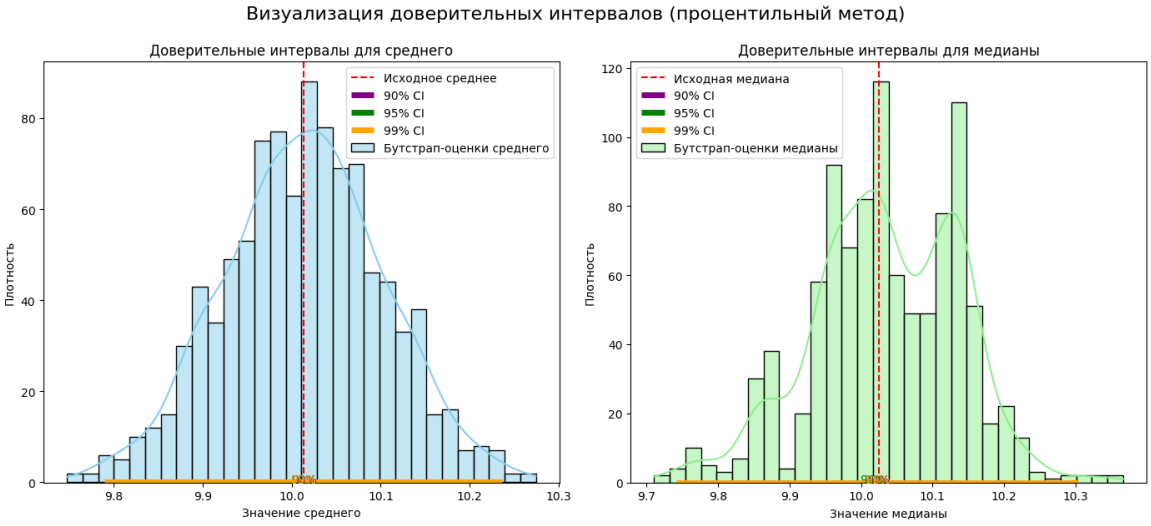

Рис. 3: Визуализация 90%, 95% и 99% доверительных интервалов для среднего и медианы, наложенных на гистограммы бутстрап-оценок. Красные пунктирные линии показывают исходные точечные оценки.

Анализ: Как видно из рисунка 3, с увеличением уровня доверия (от 90% к 99%) ширина доверительного интервала для обеих статистик (среднего и медианы) увеличивается. Это отражает фундаментальный компромисс в статистическом выводе: для достижения

большей уверенности в том, что интервал содержит истинный параметр, приходится жертвовать точностью, т.е. делать интервал шире. Исходные точечные оценки (красные пунктирные линии) стабильно находятся внутри всех построенных доверительных интервалов, что подтверждает их центральное положение в бутстрап-распределениях и состоятельность оценок.

3.4 Влияние объема выборки и числа итераций

Этот раздел посвящен исследованию того, как изменение ключевых параметров бутстрапа — объема исходной выборки (N) и числа бутстрап-итераций (B) — влияет на ширину доверительных интервалов.

3.4.1 Исследование зависимости от N (объема выборки)

Увеличение объема исходной выборки, как правило, приводит к получению более репрезентативных данных о генеральной совокупности. Это, в свою очередь, должно повысить точность оценок и, следовательно, привести к сужению доверительных интервалов. В данном эксперименте число бутстрап-итераций B было зафиксировано на уровне 1000.

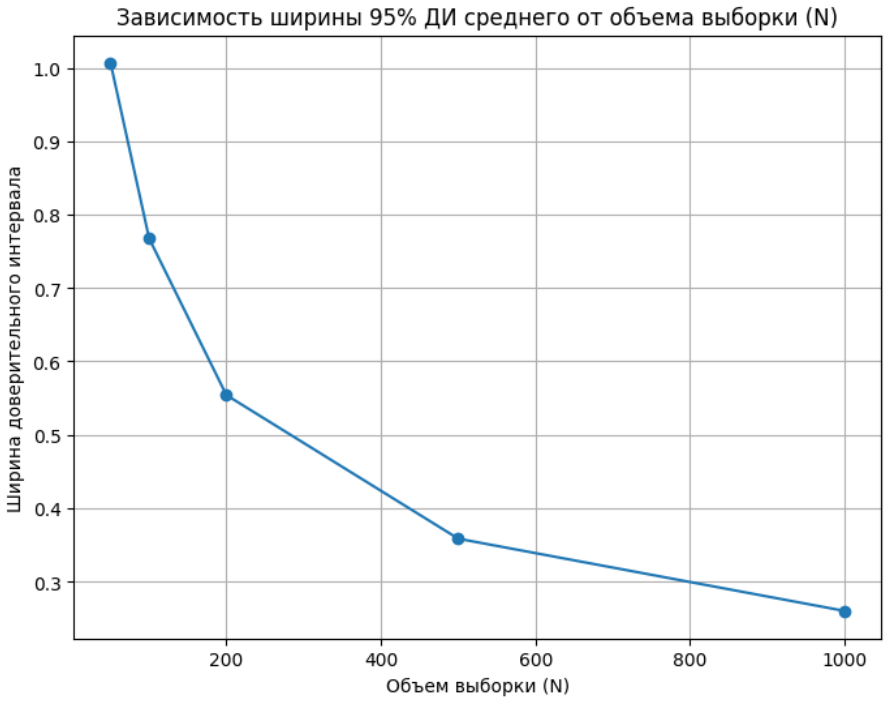

Рис. 4: Зависимость ширины 95% доверительного интервала среднего от объема выборки N ($B = 1000$).

Анализ: График на рисунке 4 демонстрирует четкую тенденцию: по мере увеличения объема выборки N ширина 95% доверительного интервала для среднего заметно уменьшается. Эта зависимость соответствует теоретическим ожиданиям, поскольку стандартная ошибка оценки обычно уменьшается пропорционально $1/\sqrt{N}$. Большие выборки обеспечивают более точную информацию о генеральной совокупности, что приводит к более узким и, следовательно, более точным доверительным интервалам.

3.4.2 Исследование зависимости от B (числа итераций бутстрапа)

Число бутстрап-итераций B влияет на стабильность и надежность самого бутстрап-распределения. Большее B приводит к более точной аппроксимации теоретического выборочного распределения статистики, а значит, к более стабильной и надежной оценке квантилей, используемых для построения доверительных интервалов. Объем исходной выборки N был зафиксирован на уровне 500.

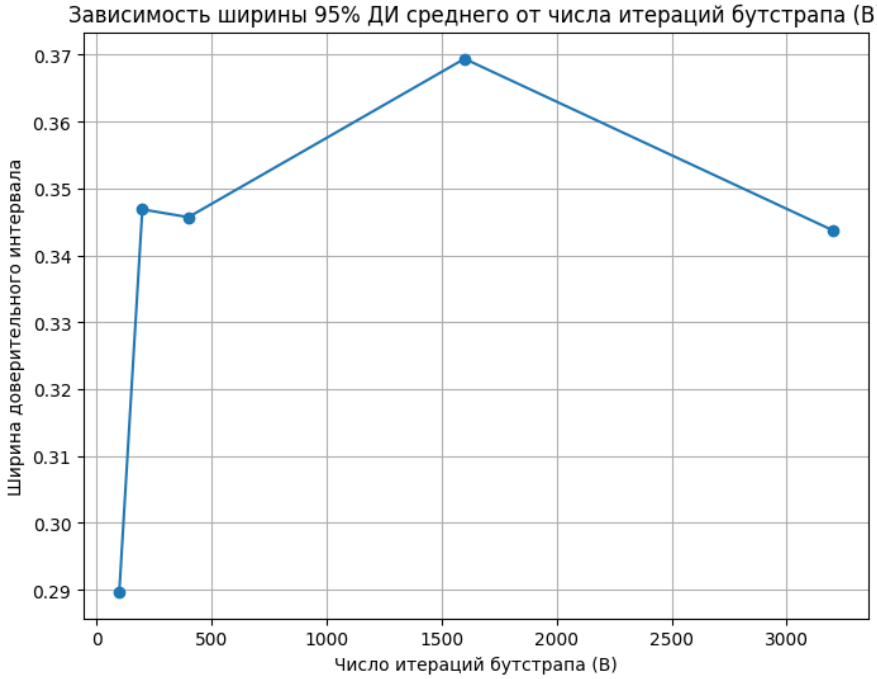

Рис. 5: Зависимость ширины 95% доверительного интервала среднего от числа итераций бутстрапа B ($N = 500$).

Анализ: На рисунке 5 видно, что с увеличением числа бутстрап-итераций B ширина доверительного интервала стремится к стабилизации. При малых значениях B (например, $B = 100$ или $B = 200$) ширина интервала может сильно варьироваться, поскольку бутстрап-распределение еще недостаточно хорошо аппроксимировано. Однако, начиная с определенного порога (в данном случае, уже при $B = 400$ и выше), дальнейшее увеличение B приводит лишь к незначительному изменению ширины интервала. Это означает, что для получения стабильных результатов достаточно выбрать адекватное число итераций (обычно $B \geq 1000$ считается хорошей практикой), дальнейшее увеличение B будет вычислительно затратным без существенного улучшения точности.

3.5 Проверка покрытия интервалов

Этот этап предназначен для эмпирической проверки того, насколько часто построенные доверительные интервалы действительно содержат истинное значение параметра генеральной совокупности. Для 95% доверительного интервала мы ожидаем, что в долгосрочной перспективе (при многократных симуляциях) истинное значение параметра будет попадать в интервал примерно в 95% случаев.

Условия эксперимента:

- **Генеральное распределение:** Стандартное нормальное распределение $\mathcal{N}(0, 1)$.
- **Истинное среднее:** $\mu = 0$.

- **Уровень доверия:** 95%.
- **Методология:** Для каждой комбинации N (объем выборки) и B (число бутстрап-итераций) было проведено 100 симуляций. В каждой симуляции генерировалась выборка из $\mathcal{N}(0,1)$, строился 95% доверительный интервал для среднего с использованием бутстрапа, и проверялось, содержит ли интервал истинное значение $\mu = 0$. Затем рассчитывалась доля интервалов, содержащих $\mu = 0$.

Таблица доли покрытия: Таблица 1 показывает долю покрытия 95% доверительных интервалов среднего для различных комбинаций объема выборки N и числа итераций бутстрапа B.

Таблица 1: Доля покрытия 95% доверительных интервалов среднего для различных комбинаций N и B

N	**B=100**	**B=200**	**B=400**	**B=1600**	**B=3200**
50	0.91	0.95	0.97	0.95	0.94
100	0.91	0.91	0.93	0.93	0.90
200	0.96	0.96	0.94	0.91	0.98
500	0.95	0.92	0.93	0.97	0.96
1000	0.95	0.97	0.92	0.99	0.94

Анализ таблицы: Как видно из таблицы 1, эмпирическая доля покрытия для 95% доверительных интервалов в большинстве случаев близка к номинальному уровню 0.95. Незначительные отклонения (например, 0.91 или 0.99) являются результатом случайности симуляции. В целом, при достаточных значениях N и B, бутстрап-интервалы, построенные процентильным методом, адекватно выполняют свою функцию по охвату истинного значения параметра.

Тепловая карта доли покрытия: Тепловая карта 6 визуализирует долю покрытия для различных комбинаций объема выборки (N) и числа бутстрап-итераций (B).

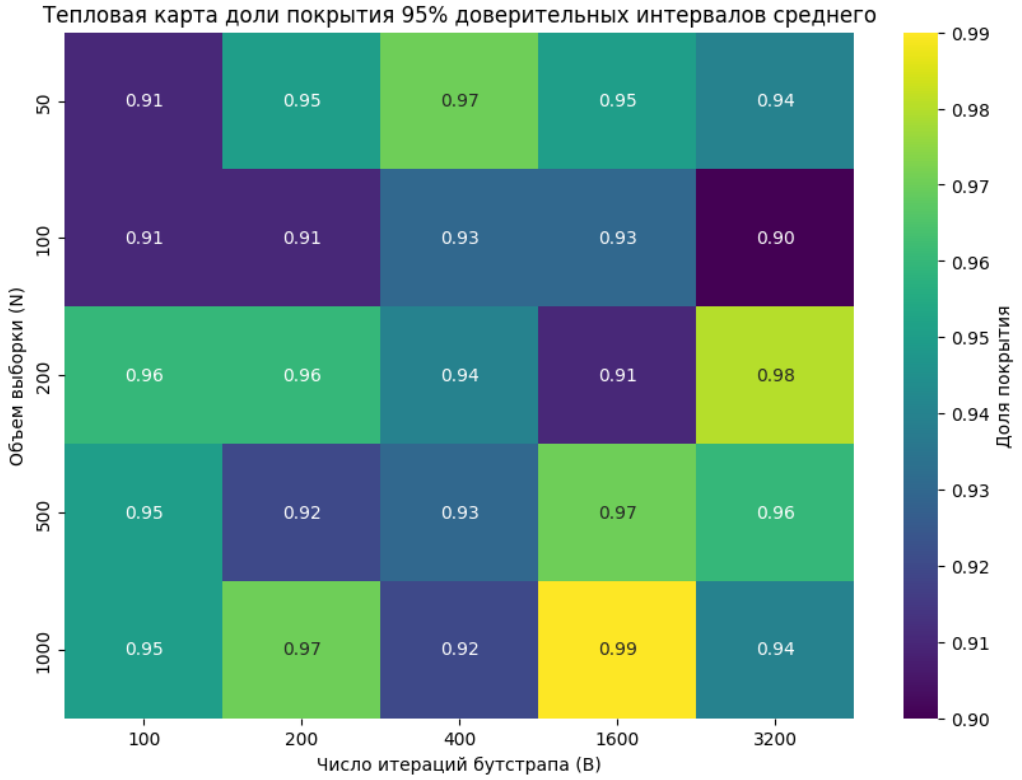

Рис. 6: Тепловая карта доли покрытия 95% доверительных интервалов среднего для различных комбинаций N и B.

Анализ тепловой карты: Тепловая карта на рисунке 6 наглядно демонстрирует, как доля покрытия зависит от N и B:

- **Влияние N:** Движение по оси Y вверх (увеличение N) в целом приводит к более стабильной и точной доле покрытия, приближающейся к 0.95. Это подтверждает, что качество доверительных интервалов (и их способность охватывать истинное значение) значительно улучшается с увеличением объема исходной выборки.
- **Влияние B:** Движение по оси X вправо (увеличение B) также приводит к повышению стабильности доли покрытия. При малых значениях B (например, $B = 100$) наблюдается большая изменчивость в показателях покрытия, что указывает на нестабильность оценки квантилей бутстрап-распределения. По мере увеличения B до 1000 и выше, покрытие становится более надежным и предсказуемым, приближаясь к ожидаемому значению 0.95.

Наиболее точные и стабильные результаты по доле покрытия наблюдаются при сочетании достаточно больших N и B, что соответствует верхнему правому углу тепловой карты. Этот анализ подчеркивает важность выбора адекватных значений N (для точности оценки) и B (для стабильности бутстрап-процесса).

4 Заключение

Выполнение данной лабораторной работы позволило глубоко исследовать метод бутстрапа и его применение для статистического вывода. Полученные результаты и их анализ позволяют сделать следующие ключевые выводы:

1. **Точечные оценки и визуализация данных:** Точечные оценки, такие как выборочное среднее, медиана, дисперсия и IQR, хорошо аппроксимируют истинные параметры генеральной совокупности при достаточно большом объеме выборки. Визуализация данных с помощью ядерной оценки плотности (KDE) является более информативной и менее зависимой от выбора параметров (например, числа бинов) по сравнению с традиционными гистограммами, предоставляя гладкую и точную оценку плотности распределения.

2. **Эффективность и гибкость Бутстрапа:** Бутстрап является универсальным и мощным непараметрическим методом, который позволяет оценить выборочные распределения статистик и построить доверительные интервалы без необходимости делать сильные предположения о форме распределения данных. Это делает его незаменимым инструментом в ситуациях, когда аналитические решения труднодоступны или не применимы. Гистограммы бутстрап-оценок наглядно показывают неопределенность и разброс вокруг исходных точечных оценок.

3. **Построение доверительных интервалов:** Процентильный метод бутстрапа прост в реализации и эффективен для конструирования доверительных интервалов. Как и ожидалось, увеличение уровня доверия приводит к расширению интервала, что отражает компромисс между уверенностью и точностью оценки.

4. **Влияние параметров на качество вывода:**

   - **Объем выборки (N):** Оказал наиболее существенное влияние на ширину доверительных интервалов и качество покрытия. Увеличение N значительно повышает точность оценок, уменьшая ширину интервалов и приводя к более стабильному и точному покрытию истинного значения параметра. Это подтверждает, что чем больше данных доступно, тем надежнее становятся статистические выводы.
   - **Число бутстрап-итераций (B):** Влияет на стабильность самого бутстрап-распределения и, как следствие, на точность границ доверительных интервалов. Показано, что существует точка насыщения, после которой дальнейшее увеличение B не приводит к значительному улучшению точности, но может существенно увеличить вычислительные затраты. Для большинства задач $B \geq 1000$ считается хорошей практикой.

5. **Свойство покрытия:** Эмпирическая проверка подтвердила, что бутстрап-интервалы адекватно "покрывают"истинное значение параметра. Доля интервалов, содержащих истинное среднее, была близка к номинальному уровню (95%), что свидетельствует о надежности и валидности бутстрап-метода для построения доверительных интервалов.

В целом, лабораторная работа продемонстрировала практическую применимость и преимущества бутстрапа как надежного и гибкого метода для статистического анализа и вывода, особенно в условиях отсутствия априорных знаний о распределении данных.