ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Университет ИТМО

Отчёт по лабораторной работе № 3

«Бутстрап-оценки»

Выполнили работу:

Демьянов Фёдор Александрович 471882

Кузнецова Таисия Павловна 466397

Долбенко Олеся Дмитриевна 465781

Академическая группа: № J3114

Санкт-Петербург 2025

Цель работы

- Ознакомиться с методами точечного оценивания параметров распределения (среднее, медиана, дисперсия, IQR).

- Изучить и реализовать непараметрический бутстрап для оценки разброса статистик и построения доверительных интервалов.

- Проанализировать влияние объёма выборки (N) и числа бутстрап-итераций (B) на ширину доверительного интервала.

- Оценить эмпирическое покрытие (95%)-го ДИ для среднего.

План работы

1. Сгенерировать выборку $(\mathrm{N} = 500)$

2. Вычислить точечные оценки и сравнить с теоретическими значениями.

3. Построить гистограмму и KDE.

4. Реализовать бутстрап $(\mathrm{B} = 1000)$, получить распределения статистик.

5. Построить доверительные интервалы для среднего и медианы при уровнях доверия $(90\%), (95\%), (99\%)$.

6. Исследовать зависимость ширины (95%) ДИ от (N) и (B).

7. Провести симуляцию покрытия (95%)-CI: $(k = 100)$ повторений $(N \in \{50,100,200,500,1000\})$, $(B \in \{100,200,400,1600,3200\})$.

8. Сделать выводы.

Реализация

Теория

- Выборочная медиана — значение, делящее упорядоченную выборку пополам.

- Выборочная дисперсия (несмещённая):

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

- Интерквартильный размах (IQR):

$$\mathrm{IQR} = Q_3 - Q_1,$$

Теоретические значения для ($N(0,1)$):

$$\mu = 0, \quad \mathrm{median} = 0, \quad \sigma^2 = 1,$$

$$Q_1 = \Phi^{-1}(0.25) \approx -0.6745,$$

$$Q_3 = \Phi^{-1}(0.75) \approx +0.6745,$$

$$\mathrm{IQR} \approx 1.349.$$

1. Генерация данных и базовые оценки

```
## 2.1 Генерация выборки
"""

# Установим «семя» генерации
np.random.seed(52)

# Количество наблюдений
N = 500

# Генерация из стандартного нормального распределения
data = np.random.randn(N)

# Выборочное среднее
sample_mean = np.mean(data)

# Медиана
sample_median = np.median(data)

# Выборочная несмещённая дисперсия
sample_var = np.var(data, ddof=1)

# Выборочная смещённая дисперсия
sample_var_a = np.var(data)


# Квартили
Q1 = np.percentile(data, 25)
Q2 = np.percentile(data, 50)
Q3 = np.percentile(data, 75)

# IQR
sample_iqr = Q3 - Q1
```

```
Выборочное среднее: 0.07253250605445645
Выборочная медиана: 0.08333931014438042
Выборочная несмещённая дисперсия: 0.8874789809141594
Выборочная смещённая дисперсия: 0.8857040229523311
Q1 (25%): -0.609642227547311
Q2 (50%): 0.08333931014438042
Q3 (75%): 0.7036743951454055
Выборочный IQR: 1.3133166226927164
```

```
"""## 2.3 Вычисление теоритический оценок"""

# Теоретические квантили для N(0,1)
q1_theor = stats.norm.ppf(0.25, loc=0, scale=1)
q2_theor = stats.norm.ppf(0.50, loc=0, scale=1)
q3_theor = stats.norm.ppf(0.75, loc=0, scale=1)

# Теоритический IQR
iqr_theor = q3_theor - q1_theor
```

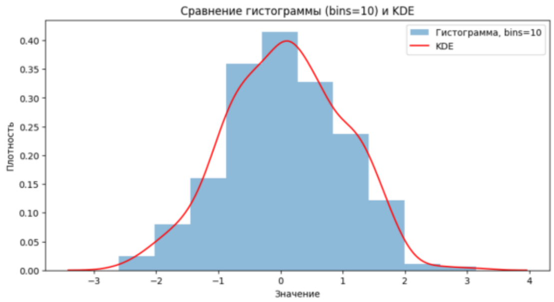
```
Теоретический Q1 (25%): -0.6744897501960817
Теоретический Q2 (50%): 0.0 (медиана)
Теоретический Q3 (75%): 0.6744897501960817
Теоретический IQR: 1.3489795003921634
```

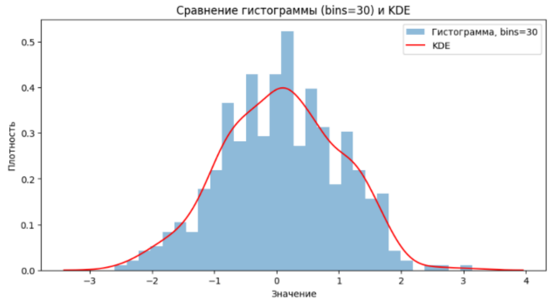
Сравнение оценок:

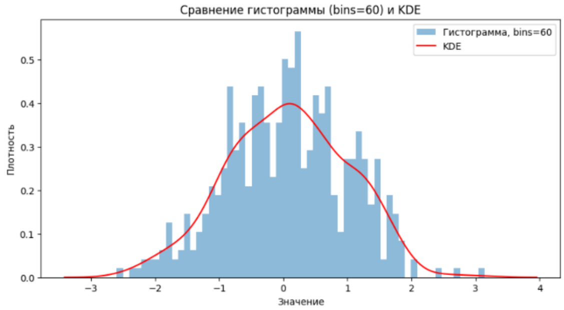
```
Разница в Q1: 0.0648475226487707
Разница в Q2: 0.08333931014438042
Разница в Q3: 0.029184644949323824
Разница в IQR: 0.03566287769944698
Разница в дисперсии: 0.11252101908584056
```

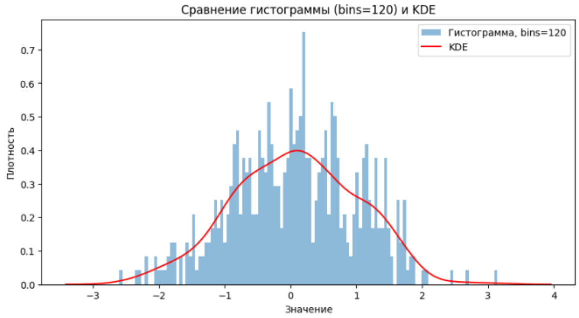
Результаты близки к теории, а небольшие отклонения обусловлены случайностью выборки.

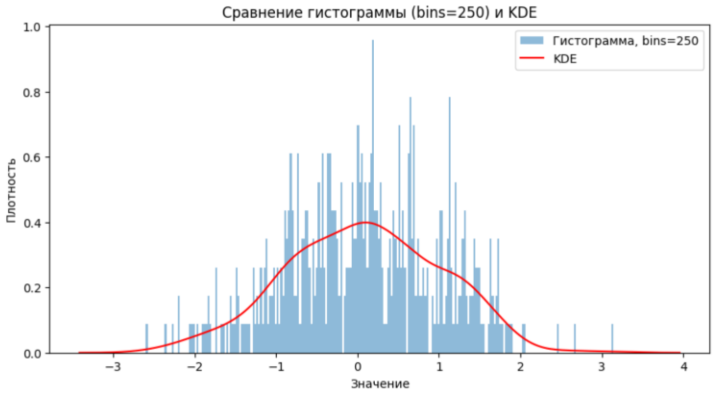
Визуализация распределения











KDE выдаёт более гладкую оценку плотности нежели гистограмма.

Гистограмма чувствительна к количеству бинов: при небольшом их количестве, гистограмма можешь ”проглотить” важные детали, при большом их количестве, появляется много шума, и гистограмма становится трудна в восприятии.

Бутстрап-оценки статистик

Идея бутстрапа и алгоритм

1. Есть исходная выборка $\{x_i\}_{i=1}^{N}$.
2. Повторно (B) раз (например, $(B = 1000)$) берём бутстрап-выборку *с возвращением*:

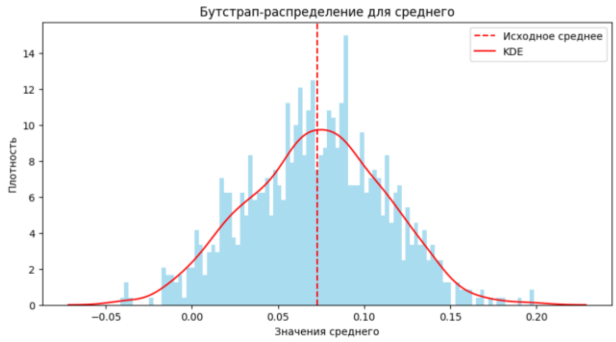
   $\{x_1^{*(b)}, x_2^{*(b)}, \dots, x_N^{*(b)}\} \sim \text{sample}\{x_i\} \text{ with replacement}, b = 1, \dots, B.$
3. Для каждой (b)-й выборки вычисляем статистику

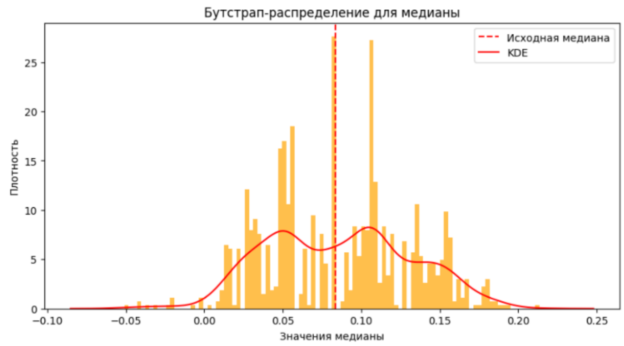
   $T^{(b)} = T\left(x_1^{*(b)}, x_2^{*(b)}, \dots, x_N^{*(b)}\right)$ (например, среднее, медиану и т. д.)
4. Массив $\{T^{(1)}, T^{(2)}, \dots, T^{(B)}\}$ эмулирует распределение статистики (T)

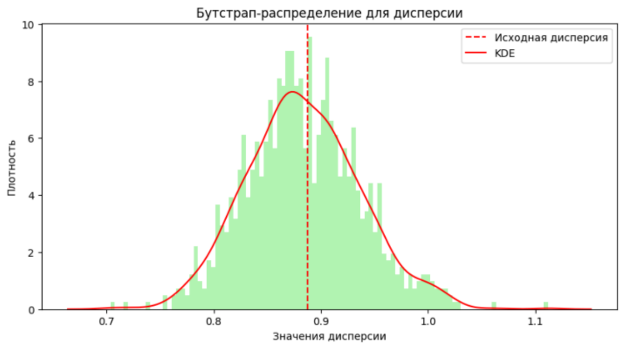
Код:

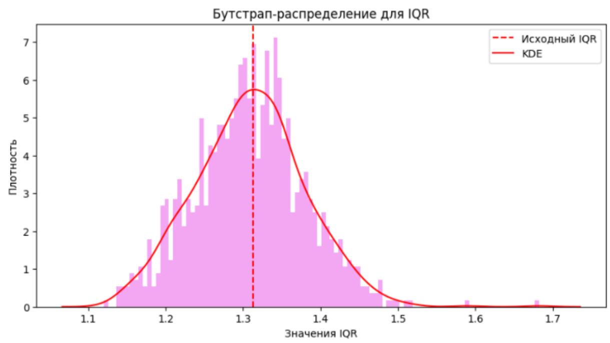
```
"""Основной цикл бутстрапа

"""

for b in range(B):
    # Берём одну бутстрап-выборку из data с возвращением
    sample_b = np.random.choice(data, size=N, replace=True)

    # Вычисляем статистики для этой выборки
    bootstrap_means[b] = np.mean(sample_b)
    bootstrap_medians[b] = np.median(sample_b)
    bootstrap_vars[b] = np.var(sample_b, ddof=1)
    Q1_b = np.percentile(sample_b, 25)
    Q3_b = np.percentile(sample_b, 75)
    bootstrap_iqrs[b] = Q3_b - Q1_b
```

Визуализации:









В целом, бустрап-оценки лежат близко к исходным значениям.

Доверительные интервалы

Доверительный интервал (ДИ) – это диапазон значений, который с заданной вероятностью (уровнем доверия) содержит истинное значение параметра распределения (например, истинного среднего μ).

Формулы:

Процентильный $(100(1-\alpha)\%)$-CI для статистики (T):

$$\mathrm{CI}_{\text{perc}} = \left[T_*^{(\alpha/2)},\ T_*^{(1-\alpha/2)}\right],$$

где

– $(T^*(p))$— (p)-й перцентиль массива $(\{T^{(b)}\}_{b=1}^{B})$,

– (α) — уровень значимости (например, для $(95\backslash\%)$ CI – $(\alpha = 0.05)$).

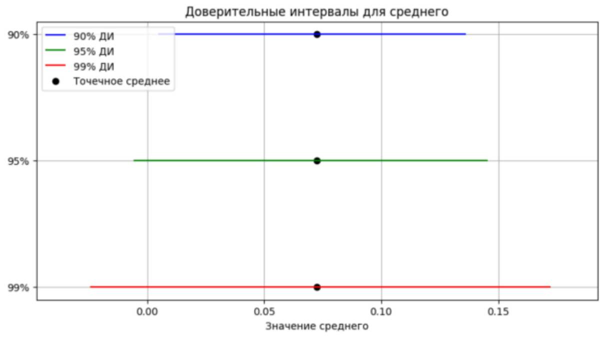
Вычисление интервалов:

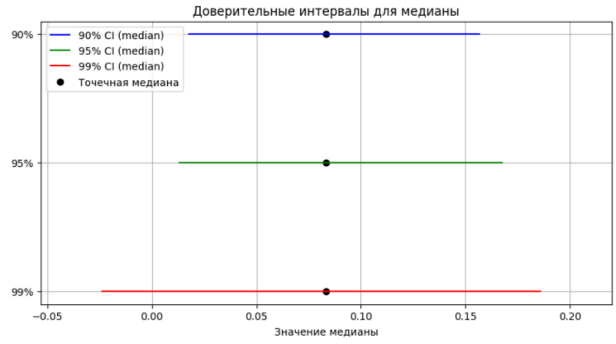
```
Функция для нахождения пределов интервала
"""

def bootstrap_percentile_interval(stat_array, alpha=0.05):
    lower = np.percentile(stat_array, 100 * (alpha/2))
    upper = np.percentile(stat_array, 100 * (1 - alpha/2))
    return lower, upper

"""## 4.2 Вычисление интервалов

Для среднего значения
"""

ci_mean_90 = bootstrap_percentile_interval(bootstrap_means, alpha=0.10)
ci_mean_95 = bootstrap_percentile_interval(bootstrap_means, alpha=0.05)
ci_mean_99 = bootstrap_percentile_interval(bootstrap_means, alpha=0.01)

"""Для медианы"""

ci_med_90 = bootstrap_percentile_interval(bootstrap_medians, alpha=0.10)
ci_med_95 = bootstrap_percentile_interval(bootstrap_medians, alpha=0.05)
ci_med_99 = bootstrap_percentile_interval(bootstrap_medians, alpha=0.01)
```

Вывод:

```
ДИ для среднего:
  90%: (np.float64(0.004644940603959617), np.float64(0.13611638986062763))
  95%: (np.float64(-0.005895783482000985), np.float64(0.1454045678298851))
  99%: (np.float64(-0.024283569191524104), np.float64(0.17258365246645893))

ДИ для медианы:
  90%: (np.float64(0.017344099626918144), np.float64(0.15695296072099782))
  95%: (np.float64(0.01283268875603577), np.float64(0.16789464133835882))
  99%: (np.float64(-0.024273462770936716), np.float64(0.1862089090765946))
```

Визуализация:

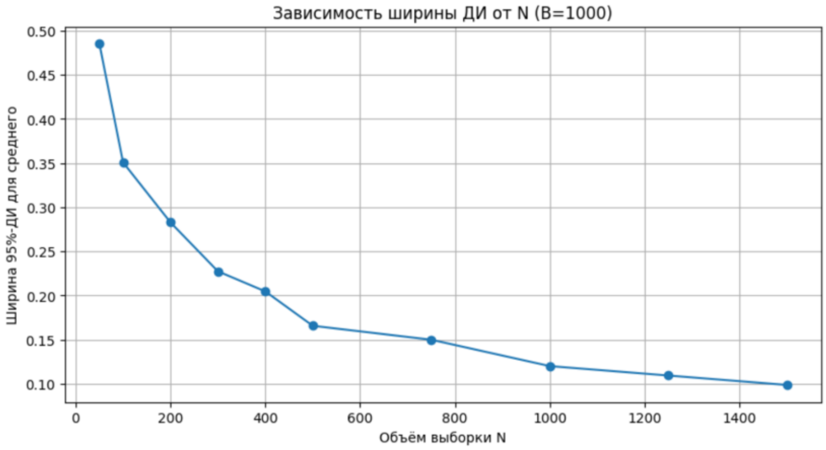




Чем процент “доверия” больше, тем, закономерно, шире границы интервала

Зависимость ширины ДИ от N:

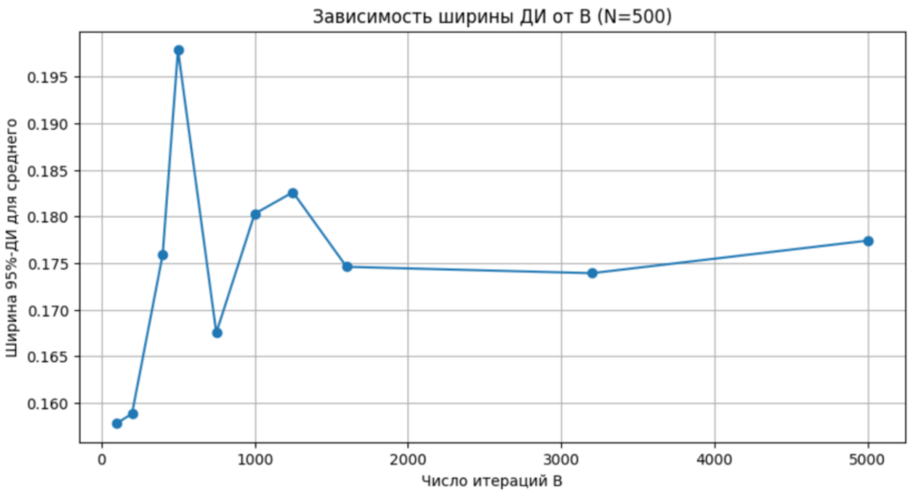
1. Сгенерировали выборки разных размеров N∈{50, 100, 200, 500, 1000}.
2. Для каждой выборки построили B=1000 бутстрап-оценок среднего.
3. Посчитали ширину CI
4. Построили график



- При небольших N (50–200) ширина CI велика, интервал ненадёжно широк.
- С ростом N CI сужается примерно пропорционально $1/\sqrt{N}$
- После N ≈ 500 дальнейшее сужение становится менее выраженным

Зависимость ширины ДИ от числа итераций B:

1. Зафиксировали N = 500, варьировали B∈{100, 200, 400, 1600, 3200}
2. Для каждого B сгенерировали соответствующее число бутстрап-выборок и построили ДИ
3. Вычислили ширину
4. Построили график



- При B <1500 ширина CI «шумит» — сильно колеблется между запусками.
- При B ≥ 1500 ширина стабилизируется, дальнейшее увеличение B даёт незначительные изменения.

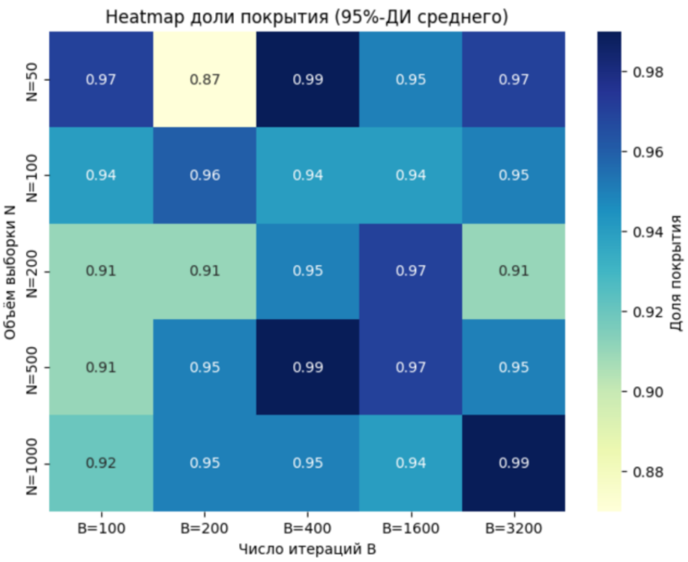
Проверка покрытия 95%-CI среднего

Провели симуляцию:

- k=100 независимых «полных» экспериментов для каждой пары (N,B).
- В каждом эксперименте сгенерировали выборку размера N, построили B бутстрап-оценок и 95%-CI.
- Подсчитали долю случаев, когда истинное среднее μ=0 лежало в CI.
- Построили Heatmap

```
## 5.1 Эксперимент

Параметры эксперимента
"""

Ns = [50, 100, 200, 500, 1000]
Bs = [100, 200, 400, 1600, 3200]
k = 100  # число независимых экспериментов для каждой пары (N, B)

"""Матрица для хранения долей покрытия (размер len(Ns) x len(Bs))"""

coverage_matrix = np.zeros((len(Ns), len(Bs)))

"""Цикл по всем N и B"""

for i, N_current in enumerate(Ns):
    for j, B_current in enumerate(Bs):

        # Счётчик того, сколько раз ДИ содержит 0
        count_contains_mu = 0

        # k экспериментов: каждый раз генерируем новую выборку
        for experiment in range(k):
            # Генерируем выборку из N(0,1)
            data_N = np.random.randn(N_current)

            # Строим бутстрап-распределение средних (B_current итераций)
            boot_means_N = np.zeros(B_current)
            for b in range(B_current):
                sample_b = np.random.choice(data_N, size=N_current, replace=True)
                boot_means_N[b] = np.mean(sample_b)

            # Вычисляем 95%-й ДИ
            lower, upper = np.percentile(boot_means_N, [2.5, 97.5])

            # Проверяем, содержится ли true_mean=0 в интервале
            if lower <= 0 <= upper:
                count_contains_mu += 1

        # Вычисляем долю покрытия
        coverage_matrix[i, j] = count_contains_mu / k
```

Heatmap:



- Идеально для 95%-го CI доля покрытия должна быть ≈ 0.95.
- При N≥50 и B≥100 фактическое покрытие близко к номинальному.
- При многократных запусках значения сильно колебались, однако почти всегда были ≈ 0.95

Итоговое заключение

1. Точечные оценки (среднее, медиана, дисперсия, IQR) для N=500 показали хорошее согласие с теоретическими значениями нормального распределения, что подтверждает корректность генерации и расчётов.

2. Ядерная оценка плотности (KDE) оказалась более информативной и стабильной по сравнению с гистограммой, особенно при сравнении нескольких выборок или при небольшом числе наблюдений, однако может сгладить необходимые детали.

3. Бутстрап (с выборкой с возвращением, B=1000) надёжно оценил распределение любых статистик, дал эмпирические стандартные ошибки и позволил построить доверительные интервалы без сложных аналитических формул.

4. Процентильный метод для CI прост в реализации и даёт асимметричные границы, адекватно отражающие форму бутстрап-распределения статистики.

5. Объём выборки N — главный фактор, влияющий на ширину, тогда как число итераций B лишь «успевает» стабилизировать оценку перцентилей, поэтому рекомендуется B ≥ 1500 и N ≥ 600.

6. Симуляционная проверка покрытия не показала конкретных результатов, при многократном запуске, значения сильно варьировались и не удалось выявить какой-либо тенденции.

Приложение

Код в colab: ST_lab_3.ipynb