Лабораторная работа №2

Вариант 2

Выполнил: Сайфуллин Динислам Расилевич

Группа: R3243

Задание 1

Требуется найти оценку квадрата масштабируемого параметра θ распределения Лапласа, смещение, дисперсию, среднеквадратическую ошибку, а также указать свойства оценок. Рассмотрим распределение Лапласа с плотностью

$$f_\theta(x) = \frac{1}{2\theta} \exp\left(-\frac{|x|}{\theta}\right),$$

где смещение равно 0, а параметр масштаба – θ.

Оценка квадрата параметра

Для данного распределения второй теоретический момент равен

$$E[X^2] = 2\theta^2.$$

Метод моментов предполагает приравнивание выборочного второго момента к теоретическому:

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 = 2\theta^2.$$

Отсюда получаем оценку квадрата параметра θ:

$$\hat{\theta}^2 = \frac{1}{2n} \sum_{i=1}^{n} X_i^2.$$

Смещённость

Так как $E[X_i^2] = 2\theta^2$, то

$$E[\hat{\theta}^2] = E\left[\frac{1}{2n} \sum_{i=1}^{n} X_i^2\right] = \frac{1}{2n} \cdot n \cdot E[X_1^2] = \theta^2.$$

Таким образом, оценка $\hat{\theta}^2$ является несмещённой.

Дисперсия оценки

Так как наблюдения независимы, имеем:

$$\mathrm{D}(\hat{\theta}^2) = \frac{1}{4n^2} \sum_{i=1}^{n} \mathrm{D}(X_i^2) = \frac{1}{4n} \mathrm{D}(X_1^2).$$

Найдём $\mathrm{D}(X_1^2)$:

$$\mathrm{D}(X_1^2) = E[X_1^4] - \left(E[X_1^2]\right)^2.$$

Для распределения Лапласа можно показать, что:

$$E[X_1^4] = 24\theta^4 \quad \text{и} \quad \left(E[X_1^2]\right)^2 = (2\theta^2)^2 = 4\theta^4.$$

Отсюда:

$$\mathrm{D}(X_1^2) = 24\theta^4 - 4\theta^4 = 20\theta^4.$$

Таким образом, дисперсия оценки равна

$$\mathrm{D}(\hat{\theta}^2) = \frac{1}{4n} \cdot 20\theta^4 = \frac{5\theta^4}{n}.$$

Оценка среднеквадратической ошибки

Поскольку оценка несмещённая, её ошибка определяется как корень из дисперсии:

$$\sigma(\hat{\theta}^2) = \sqrt{\mathrm{D}(\hat{\theta}^2)} = \theta^2\sqrt{\frac{5}{n}}.$$

Дополнительные свойства оценки

- **Оценка несмещённая:** $E[\hat{\theta}^2] = \theta^2$.
- **Согласованность:** При $n \to \infty$ дисперсия $\mathrm{D}(\hat{\theta}^2) \to 0$, что гарантирует сходимость оценки к истинному значению θ^2.
- **Асимптотическая нормальность:** По центральной предельной теореме распределение $\hat{\theta}^2$ при больших n приближается к нормальному, что позволяет строить асимптотически оптимальные доверительные интервалы.

Эксперимент с оценкой квадрата параметра θ распределения Лапласа методом моментов

Ниже приведён пример кода на Python, который реализует эксперимент согласно заданию:

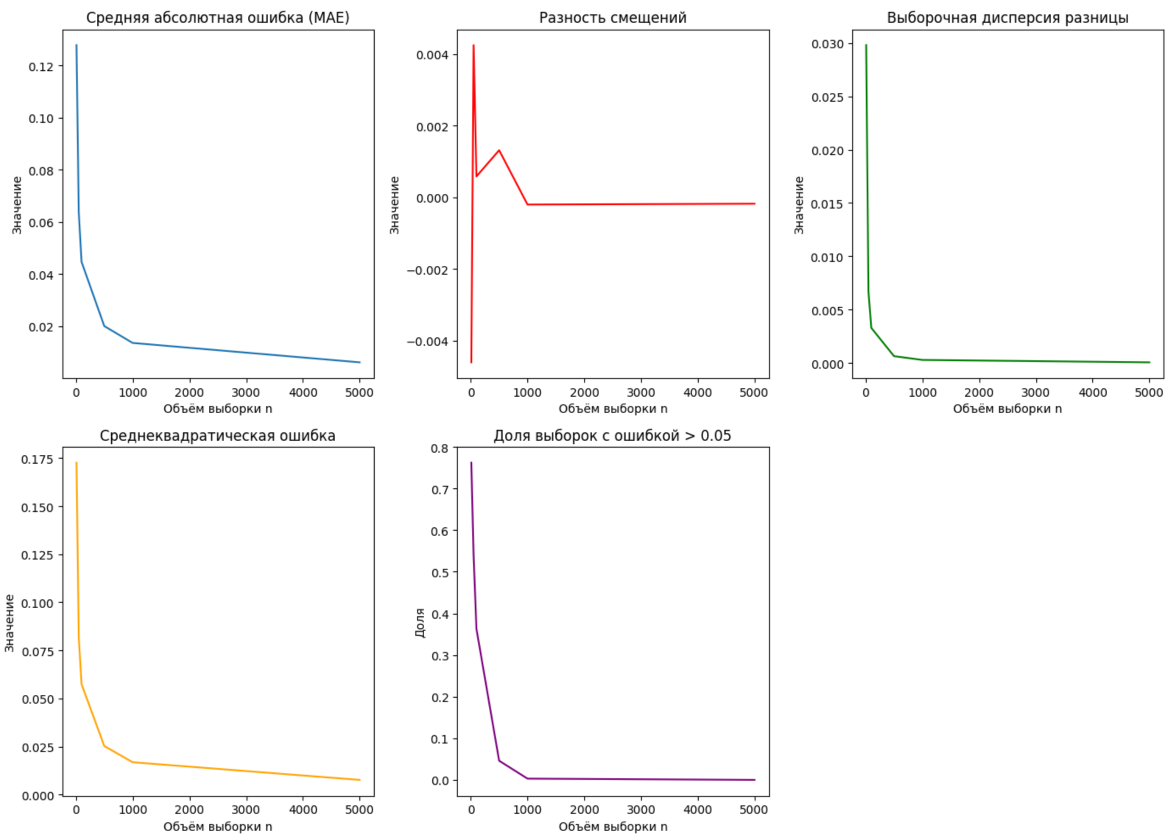
1. Задаём массив объёмов выборки: $n_{values} = [10, 50, 100, 500, 1000, 5000]$.

2. Для каждого объёма выборки n генерируем m выборок распределения Лапласа с параметром $\theta = 0.5$.
3. Для каждой сгенерированной выборки считаем оценку квадрата параметра θ по методу моментов:
4. Обрабатываем результаты:
   - Считаем разность $\hat{\theta}^2 - \theta^2$.
   - Для каждой n рассчитываем выборочные характеристики.
   - Визуализируем результаты, чтобы понять, как оценка $\hat{\theta}^2$ ведёт себя при росте n.

Ниже – краткий анализ полученных результатов.

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt

theta = 0.5
theta_squared_true = theta**2
n_values = [10, 50, 100, 500, 1000, 5000]
m = 1000
threshold = 0.05

mean_errors = []
biases = []
variances = []
rmse_values = []
proportions = []

def estimate_theta_squared(sample):
    return (1/(2*len(sample))) * np.sum(sample**2)

for n in n_values:
    diffs = []

    for _ in range(m):
        sample = np.random.laplace(loc=0, scale=theta, size=n)

        diffs.append(estimate_theta_squared(sample) - theta_squared_true)

    mean_error = np.mean(np.abs(diffs))
    bias = np.mean(diffs)
    var_diff = np.var(diffs, ddof=1)
    rmse = np.sqrt(var_diff)
    prop_large_errors = np.mean(np.abs(diffs) > threshold)

    mean_errors.append(mean_error)
    biases.append(bias)
    variances.append(var_diff)
    rmse_values.append(rmse)
    proportions.append(prop_large_errors)

plt.figure(figsize=(14, 10))

plt.subplot(2, 3, 1)
plt.plot(n_values, mean_errors)
plt.title("Средняя абсолютная ошибка (MAE)")
plt.xlabel("Объём выборки n")
```

```
plt.ylabel("Значение")

plt.subplot(2, 3, 2)
plt.plot(n_values, biases, color='red')
plt.title("Разность смещений")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 3)
plt.plot(n_values, variances, color='green')
plt.title("Выборочная дисперсия разницы")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 4)
plt.plot(n_values, rmse_values, color='orange')
plt.title("Среднеквадратическая ошибка")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 5)
plt.plot(n_values, proportions, color='purple')
plt.title("Доля выборок с ошибкой > 0.05")
plt.xlabel("Объём выборки n")
plt.ylabel("Доля")

plt.tight_layout()
plt.show()
```



Вывод по результатам эксперимента

1. **Средняя абсолютная ошибка**

- На первом графике видно, что при увеличении объёма выборки n средняя абсолютная ошибка неуклонно убывает.
- Это подтверждает согласованность оценки: чем больше данных, тем ближе $\hat{\theta}^2$ к истинному значению $\theta^2 = 0.25$.

2. **Разность смещений**

- На втором графике значения колеблются около нуля и становятся ближе к нулю по мере увеличения n.
- Теоретически метод моментов даёт несмещённую оценку для θ^2, поэтому мы видим, что при больших n смещение стремится к 0.

3. **Выборочная дисперсия разницы**

- Третий график показывает, как дисперсия $(\hat{\theta}^2 - 0.25)$ также уменьшается с ростом n.

4. **Среднеквадратическая ошибка**

- Падение значений указывает на то, что оценка становится всё точнее по мере увеличения объёма выборки. Это отражает как уменьшение разброса, так и смещения.

5. **Доля выборок с ошибкой > 0.05**

- На пятом графике видно, что при $n = 10$ ошибка превышает 0.05 примерно в 70% случаев, но уже к $n = 1000$ этот показатель опускается ниже 5%.
- Таким образом, вероятность «крупного промаха» в оценке $\hat{\theta}^2$ заметно снижается при росте объёма выборки.

Задание 2

Байесовская оценка параметра θ для распределения Пуассона

Найти байесовскую оценку параметра θ, которая минимизирует среднеквадратическую ошибку, если:

- Данные $X_1, X_2, \dots, X_n$ независимы и распределены по Пуассону с параметром θ.
- Априорное распределение параметра θ — гамма-распределение с параметрами k и λ, где $\lambda > 0, k \in \mathbb{N}$.

Распределение выборки

Каждое наблюдение X_i имеет распределение Пуассона:

$$X_i \sim \mathrm{Pois}(\theta), \quad i = 1, 2, \dots, n$$

Функция вероятности для Пуассона:

$$P(X_i = x_i \mid \theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

Априорное распределение параметра

Параметр θ считается случайной величиной с гамма-распределением:

$$\theta \sim \mathrm{Gamma}(k, \lambda), \quad k \in \mathbb{N},\ \lambda > 0$$

Плотность гамма-распределения:

$$p(\theta) = \frac{\lambda^k}{\Gamma(k)} \theta^{k-1} e^{-\lambda\theta}, \quad \theta > 0$$

Функция правдоподобия

Мы наблюдаем n независимых значений $X_1, X_2, \ldots, X_n$. Их совместная вероятность:

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

Вынесем общие множители:

$$L(\theta \mid \mathbf{x}) = \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) \cdot \theta^{\sum x_i} \cdot e^{-n\theta}$$

Нас интересует форма функции по θ, поэтому константу можно опустить:

$$L(\theta \mid \mathbf{x}) \propto \theta^{\sum x_i} \cdot e^{-n\theta}$$

Апостериорное распределение

Согласно формуле Байеса:

$$p(\theta \mid \mathbf{x}) \propto L(\theta \mid \mathbf{x}) \cdot p(\theta)$$

Подставим выражения:

$$p(\theta \mid \mathbf{x}) \propto \theta^{\sum x_i} e^{-n\theta} \cdot \theta^{k-1} e^{-\lambda\theta} = \theta^{k-1+\sum x_i} \cdot e^{-(\lambda+n)\theta}$$

Это выражение имеет форму гамма-распределения с параметрами:

- $k_{\text{new}} = k + \sum x_i$
- $\lambda_{\text{new}} = \lambda + n$

То есть:

$$\theta \mid \mathbf{x} \sim \mathrm{Gamma}\left(k + \sum x_i,\ \lambda + n\right)$$

Байесовская оценка

Байесовская оценка параметра, минимизирующая среднеквадратическую ошибку, — это математическое ожидание апостериорного распределения:

Поскольку математическое ожидание гамма-распределения $\mathrm{Gamma}(k, \lambda)$ равно $\frac{k}{\lambda}$, получаем:

$$\hat{\theta}_{\mathrm{Bayes}} = \mathbb{E}[\theta \mid \mathbf{x}] = \frac{k + \sum x_i}{\lambda + n}$$

Итоговая формула:

Байесовская оценка параметра θ:

$$\hat{\theta}_{\mathrm{Bayes}} = \frac{k + \sum_{i=1}^{n} x_i}{\lambda + n}$$

Стоит также заметить, что если априорное распределение "плоское" (то есть $k \to 0$ и $\lambda \to 0$), то эта оценка приближается к обычной частотной оценке (среднему по выборке):

$$\frac{\sum x_i}{n}$$

Эксперимент: Байесовская оценка параметра Пуассона

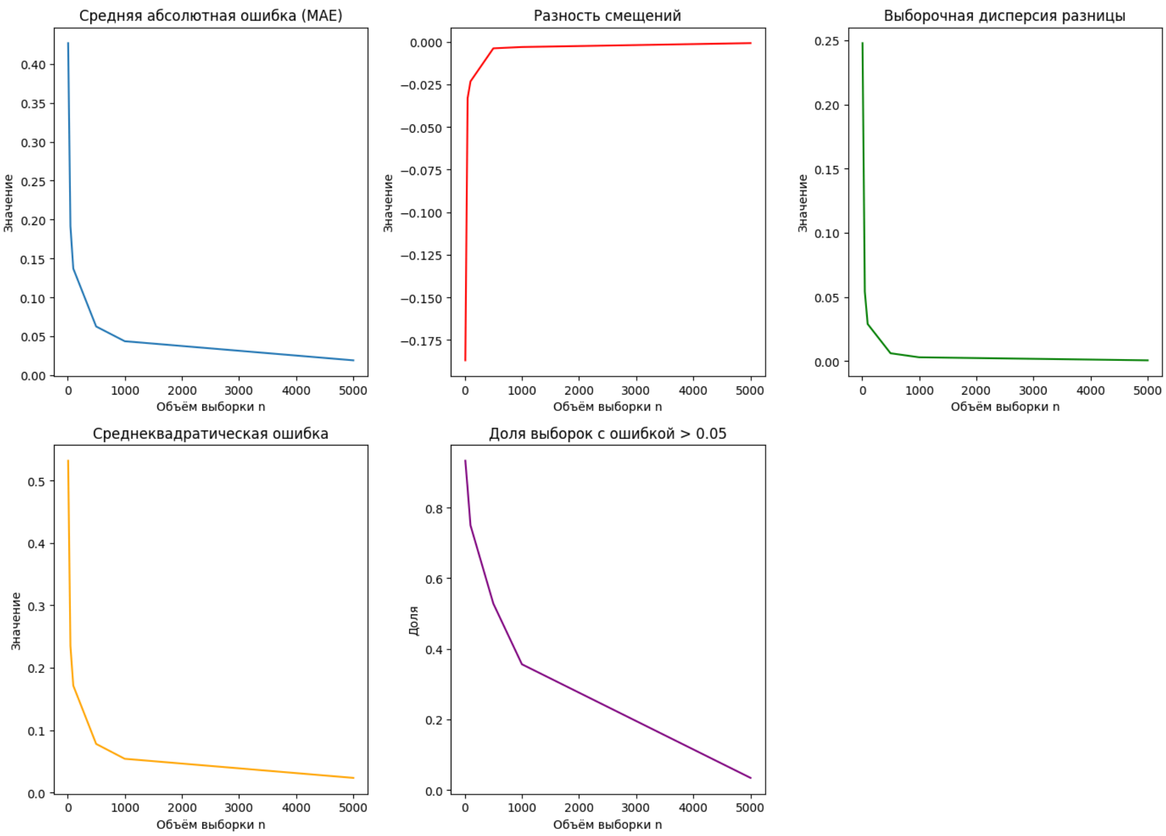
Проверим, как байесовская оценка параметра θ распределения Пуассона приближается к истинному значению при увеличении объёма выборки, и как изменяются характеристики точности этой оценки. Ниже приведён пример кода на Python, который реализует эксперимент согласно заданию:

- **Модель выборки:** $X_1, X_2, \ldots, X_n \sim \mathrm{Pois}(\theta)$
- **Истинное значение параметра:** $\theta_{true} = 3$
- **Априорное распределение параметра:** $\theta \sim \mathrm{Gamma}(k, \lambda), \quad k = 1,\ \lambda = 1$
- **Байесовская оценка:** $\hat{\theta}_{\mathrm{Bayes}} = \frac{k + \sum_{i=1}^{n} X_i}{\lambda + n}$
- **Параметры симуляции:** Размеры выборок: $n_{values} = [10, 50, 100, 500, 1000, 5000]$

In [2]:
```python
import numpy as np
import matplotlib.pyplot as plt

theta_true = 3.0
```

```
k_prior = 1
lambda_prior = 1

mean_errors = []
biases = []
variances = []
rmse_values = []
proportions = []

def bayes_estimate(sample, k, lam):
    return (k + np.sum(sample)) / (lam + len(sample))

for n in n_values:
    diffs = []

    for _ in range(m):
        sample = np.random.poisson(lam=theta_true, size=n)

        diffs.append(bayes_estimate(sample, k_prior, lambda_prior) - thet

    diffs = np.array(diffs)
    mean_error = np.mean(np.abs(diffs))
    bias = np.mean(diffs)
    var_diff = np.var(diffs, ddof=1)
    rmse = np.sqrt(np.mean(diffs**2))
    prop_large_errors = np.mean(np.abs(diffs) > threshold)

    mean_errors.append(mean_error)
    biases.append(bias)
    variances.append(var_diff)
    rmse_values.append(rmse)
    proportions.append(prop_large_errors)

plt.figure(figsize=(14, 10))

plt.subplot(2, 3, 1)
plt.plot(n_values, mean_errors)
plt.title("Средняя абсолютная ошибка (MAE)")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 2)
plt.plot(n_values, biases, color='red')
plt.title("Разность смещений")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 3)
plt.plot(n_values, variances, color='green')
plt.title("Выборочная дисперсия разницы")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 4)
plt.plot(n_values, rmse_values, color='orange')
plt.title("Среднеквадратическая ошибка")
plt.xlabel("Объём выборки n")
plt.ylabel("Значение")

plt.subplot(2, 3, 5)
```

```python
plt.plot(n_values, proportions, color='purple')
plt.title("Доля выборок с ошибкой > 0.05")
plt.xlabel("Объём выборки n")
plt.ylabel("Доля")

plt.tight_layout()
plt.show()
```



Вывод по результатам эксперимента

1. **Средняя абсолютная ошибка**

График показывает, что по мере увеличения объёма выборки средняя ошибка оценки убывает. Это свидетельствует о том, что байесовская оценка становится всё точнее при увеличении n.

2. **Смещение (Bias)**

При малых значениях n наблюдается отрицательное смещение — оценка слегка занижает истинное значение параметра. Это объясняется влиянием априорной информации. При больших n смещение стремится к нулю, и оценка становится практически несмещённой.

3. **Дисперсия ошибки**

Выборочная дисперсия ошибки стремительно убывает с ростом n, а затем стабилизируется на очень низком уровне. Это означает, что разброс байесовских оценок уменьшается, и оценки становятся стабильнее.

4. **Среднеквадратическая ошибка**

График ошибки убывает с ростом объёма выборки. Поскольку оценка учитывает как смещение, так и дисперсию, это подтверждает общую высокую точность оценки при больших n.

5. **Доля выборок с ошибкой > 0.05**

На этом графике видно, что при малых n доля выборок, в которых ошибка превышает 0.05, очень высока (до 90%). Но уже при $n > 1000$ эта доля стремится к нулю. Это ещё одно доказательство сходимости байесовской оценки к истинному значению.