Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени
Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Лабораторная работа № 3 по дисциплине «Методы машинного обучения»

Тема Проверка гипотезы о математическом ожидании - Две выборки

Студент Сапожков А.М,

Группа ИУ7-23М

Преподаватель Солодовников В.И.

Москва, 2025

Содержание

ВВЕДЕНИЕ

Математическая статистика является фундаментальным компонентом машинного обучения, поскольку она обеспечивает необходимую базу для оценки достоверности выводов и принятия обоснованных решений на основе данных. В частности, проверка гипотез о математическом ожидании двух выборок имеет важное значение в статистике, поскольку позволяет исследователям оценить вероятность того, что две группы данных имеют схожие характеристики.

Целью данной лабораторной работы является освоение практических навыков проверки гипотез о математическом ожидании для двух случайных выборок с помощью статистических методов.

Задачи данной лабораторной работы:

1) сгенерировать две независимые выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ с нормальные законом распределения и с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответствено;

2) осуществить проверку гипотезы H_0 о соответствии выборок нормальному закону распределения;

3) осуществить проверку гипотезы H_0 о принадлежности выборок одной генеральной совокупности;

4) осуществить проверку гипотезы $H_0 : a_1 = a_2$ против альтернативы $H_1 : a_1 \neq a_2$;

5) производить сдвиг вправо всех элементов второй выборки на величину $\delta = 0.01$ и осуществлять проверку гипотезы $H_0 : a_1 = a_2$ до тех пор, пока гипотеза H_0 не будет отвергнута;

6) для второй выборки назначить a_2 равным середине пройденного отрезка из пункта 5 и постепенно увеличивать число элементов в выборках и осуществлять проверку гипотезы $H_0 : a_1 = a_2$ до тех пор, пока гипотеза H_0 не будет отвергнута

7) рассчитать 95% доверительные интервалы для математических ожиданий двух выборок в момент, когда гипотеза H_0 была отвергнута в пунктах 5 и 6.

1 Аналитическая часть

1.1 Критерий Шапиро-Уилка

Критерий Шапиро-Уилка используется для проверки гипотезы H_0: «случайная величина X распределена нормально» и является одним наиболее эффективных критериев проверки нормальности. Критерии, проверяющие нормальность выборки, являются частным случаем критериев согласия.

Критерий Шапиро-Уилка основан на оптимальной линейной несмещённой оценке дисперсии к её обычной оценке методом максимального правдоподобия. Статистика критерия имеет вид:

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2, \tag{1.1}$$

где $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Числитель является квадратом оценки среднеквадратического отклонения Ллойда. Коэффициенты a_{n-i+1} берутся из таблиц. Критические значения статистики $W(\alpha)$ также находятся таблично.

Если $W < W(\alpha)$, то нулевая гипотеза о нормальности распределения отклоняется при уровне значимости α. Приближённая вероятность получения эмпирического значения W при H_0 вычисляется по формуле

$$z = \gamma + \eta \ln\left(\frac{W - \epsilon}{1 - W}\right), \tag{1.2}$$

где $\gamma,\ \eta,\ \epsilon$ — табличные коэффициенты.

Критерий Шапиро-Уилка является очень мощным критерием для проверки нормальности, но, к сожалению, имеет ограниченную применимость. При больших значениях $n (n > 100)$ таблицы коэффициентов a_{n-i+1} становятся неудобными.

1.2 Критерий Стьюдента

Рассмотрим специальный случай двухвыборочных критериев согласия. Проверяется гипотеза сдвига, согласно которой распределения двух выборок имеют одинаковую форму и отличаются только сдвигом на константу.

Критерий Стьюдента. Рассмотрим теперь задачу сравнения средних значений двух нормальных выборок.

Пусть $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ — нормальные независимые выборки из законов распределения с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответственно.

Рассмотрим проверку гипотезы:

$$H_0 : a_1 = a_2 H_1 : a_1 \neq a_2 \tag{1.3}$$

Относительно параметров σ_1^2 и σ_2^2 выделим следующие четыре варианта предположений:

1) обе дисперсии известны и равны между собой;
2) обе дисперсии известны, но не равны между собой;
3) обе дисперсии неизвестны, но предполагается, что они равны между собой;
4) обе дисперсии неизвестны, их равенство не предполагается.

Для построения критерия проверки гипотезы H_0 проведем следующие рассуждения.

От выборок $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ перейдем к выборочным средним $\overline{x}$ и $\overline{y}$. Согласно свойствам нормального распределения и выдвинутой гипотезе, величины $\overline{x}$ и $\overline{y}$ имеют нормальные распределения с одними тем же средним и дисперсиями σ_1^2/n и σ_2^2/m.

Далее перейдем к статистике, основанной на выборочных средних $\overline{x}$ и $\overline{y}$ и дисперсиях σ_1^2 и σ_2^2 (если они известны) или их оценках s_1^2 и $s2^2$ (если дисперсии неизвестны).

$\frac{1}{m} \sum_{i=1}^{m} x_i$ — выборочное среднее,
$\frac{1}{m-1} \sum_{i=1}^{m} (x_i - \overline{x})^2$ — выборочная дисперсия.

Далее рассмотрим случай, когда обе дисперсии известны и равны между собой.

$$\frac{\overline{x} - \overline{y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{1.4}$$

Статистика имеет стандартное нормальное распределение, так как является линейной комбинацией независимых нормальных величин. Гипотеза H принимается на уровне значимости α, если

$$\left| \frac{\overline{x} - \overline{y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < z_{1-\alpha/2} \tag{1.5}$$

в противном случае гипотеза отвергается в пользу альтернативы $a_1 \neq a_2$.

2 Технологическая часть

2.1 Средства реализации

В качестве языка программирования для реализации алгоритмов был выбран язык программирования Python ввиду наличия библиотек для обучения регрессионных моделей, таких как sklearn и numpy.

2.2 Реализация алгоритмов

На листинге 2.1 представлена реализация алгоритма проверки статистических гипотез для двух выборок.

Листинг 2.1 — Проверка статистических гипотез для двух выборок

```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
import seaborn as sns
import math

a = 0
sigma = 1
n = m = 30
delta = 0.1
alpha = 0.05

x = np.random.normal(a, sigma, n)
y = np.random.normal(a, sigma, n)
x_0, y_0 = x, y

plt.hist(x, alpha=0.5, label='x')
plt.hist(y, alpha=0.5, label='y')
plt.legend(loc='upper right')
plt.show()

shapiro_test1 = stats.shapiro(x)
shapiro_test2 = stats.shapiro(y)

print("Тест Шапиро-Уилка для выборки 1: Statistic =", shapiro_test1.
   statistic, "p-value =", shapiro_test1.pvalue)
print("Тест Шапиро-Уилка для выборки 2: Statistic =", shapiro_test2.
   statistic, "p-value =", shapiro_test2.pvalue)
```

```
t_stat, p_value = stats.ttest_ind(x, y, equal_var=True)
print("Тест Стьюдента для
   выборок: Statistic =", t_stat, "p-value =", p_value)

a_new = a
t_stats, p_values = [], []
rejected = False

while not rejected:
  t_stat, p_value = stats.ttest_ind(x, y, equal_var=True)
  rejected = p_value < alpha
  t_stats.append(t_stat)
  p_values.append(p_value)
  if not rejected:
    a_new += delta
    y += delta

print('Итоговый сдвиг второй выборки:', a_new - a)
x_4, y_4 = x, y

plt.plot(t_stats, label="t-statistic")
plt.plot(p_values, label="P-value")
plt.axhline(alpha, color="red", linestyle="--")
plt.legend()
plt.show()

plt.hist(x_4, alpha=0.5, label='x')
plt.hist(y_4, alpha=0.5, label='y')
plt.legend(loc='upper right')
plt.show()

a_new = (a + a_new) / 2
y -= (a_new - a) / 2
t_stats, p_values = [], []
rejected = False

while not rejected:
  t_stat, p_value = stats.ttest_ind(x, y, equal_var=True)
  rejected = p_value < alpha
  t_stats.append(t_stat)
  p_values.append(p_value)
```

```
  if not rejected:
    x = np.hstack((x, np.random.normal(a, sigma, int(n*delta))))
    y = np.hstack((y, np.random.normal(a_new, sigma, int(n*delta))))

print('Размеры выборок: len(x)=', len(x), 'len(y)=', len(y))
x_5, y_5 = x, y

plt.plot(t_stats, label="t-statistic")
plt.plot(p_values, label="P-value")
plt.axhline(alpha, color="red", linestyle="--")
plt.legend()
plt.show()

plt.hist(x_5, alpha=0.5, label='x')
plt.hist(y_5, alpha=0.5, label='y')
plt.legend(loc='upper right')
plt.show()

conf_int_x = stats.t.interval(0.95, len(x_4)-1, loc=np.mean(x_4),
   scale=stats.sem(x_4))
conf_int_y = stats.t.interval(0.95, len(y_4)-1, loc=np.mean(y_4),
   scale=stats.sem(y_4))

print(f"95% доверительный интервал для x: {conf_int_x}")
print(f"Ширина {conf_int_x[1] - conf_int_x[0]}")
print(f"95% доверительный интервал для y: {conf_int_y}")

conf_int_x = stats.t.interval(0.95, len(x_5)-1, loc=np.mean(x_5),
   scale=stats.sem(x_5))
conf_int_y = stats.t.interval(0.95, len(y_5)-1, loc=np.mean(y_5),
   scale=stats.sem(y_5))

print(f"95% доверительный интервал для x: {conf_int_x}")
print(f"Ширина {conf_int_x[1] - conf_int_x[0]}")
print(f"95% доверительный интервал для y: {conf_int_y}")

t_dist = np.linspace(stats.t.ppf(0.001, n+m-2), stats.t.ppf(0.999, n+
   m-2), 1000)
pdf_values = stats.t.pdf(t_dist, n+m-2)

plt.plot(t_dist, pdf_values, label='t-распределение')
```

```
plt.axvline(t_stat, color='r', linestyle='--', label='Критическое
    значение')
plt.fill_between(t_dist, pdf_values, where=((t_dist < stats.t.ppf(
    alpha/2, n+m-2)) | (t_dist > stats.t.ppf(1 - alpha/2, n+m-2))),
    color='gray', alpha=0.5, label='Область отклонения
    гипотезы')
plt.legend()
plt.show()

print(f'Критическое значение: {t_stat:.4f}')

ci_low_a1, ci_high_a1 = stats.norm.interval(0.95, loc=np.mean(x_0),
    scale=stats.sem(x))
ci_low_a2, ci_high_a2 = stats.norm.interval(0.95, loc=np.mean(y_0),
    scale=stats.sem(y))

print(f'Доверительный интервал для первой
    выборки: [{ci_low_a1:.4f}, {ci_high_a1:.4f}]')
print(f'Доверительный интервал для второй
    выборки: [{ci_low_a2:.4f}, {ci_high_a2:.4f}]')
```

3 Исследовательская часть

3.1 Среда для тестирования

Для тестирования разработанного алгоритма применялась облачная платформа Google Colab, не требующая установки ПО на локальный компьютер.

3.2 Проверка статистических гипотез

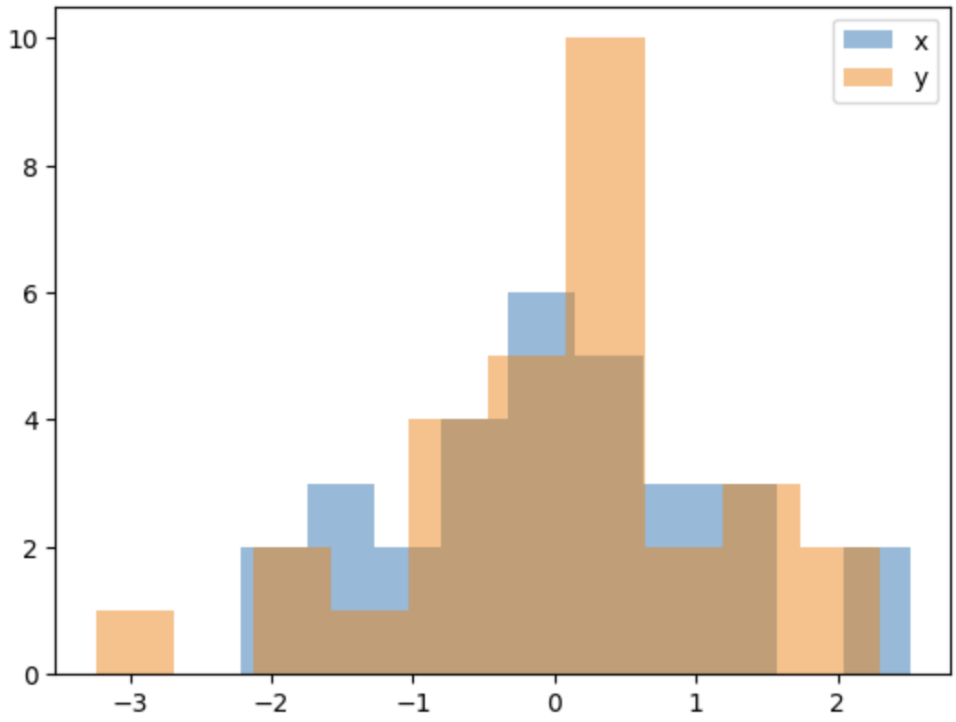

Рисунок 3.1 — Исходные данные

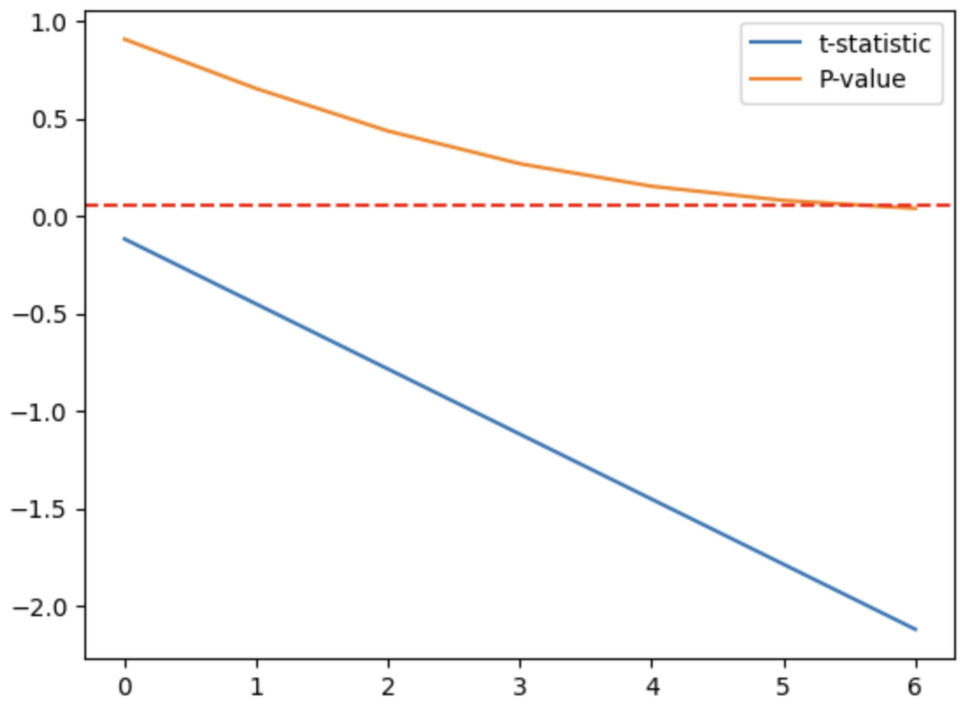

Рисунок 3.2 — Динамика изменения значений статистики критерия и P-value для всех итераций проверки гипотезы о мат. ожидании при смещении второй выборки

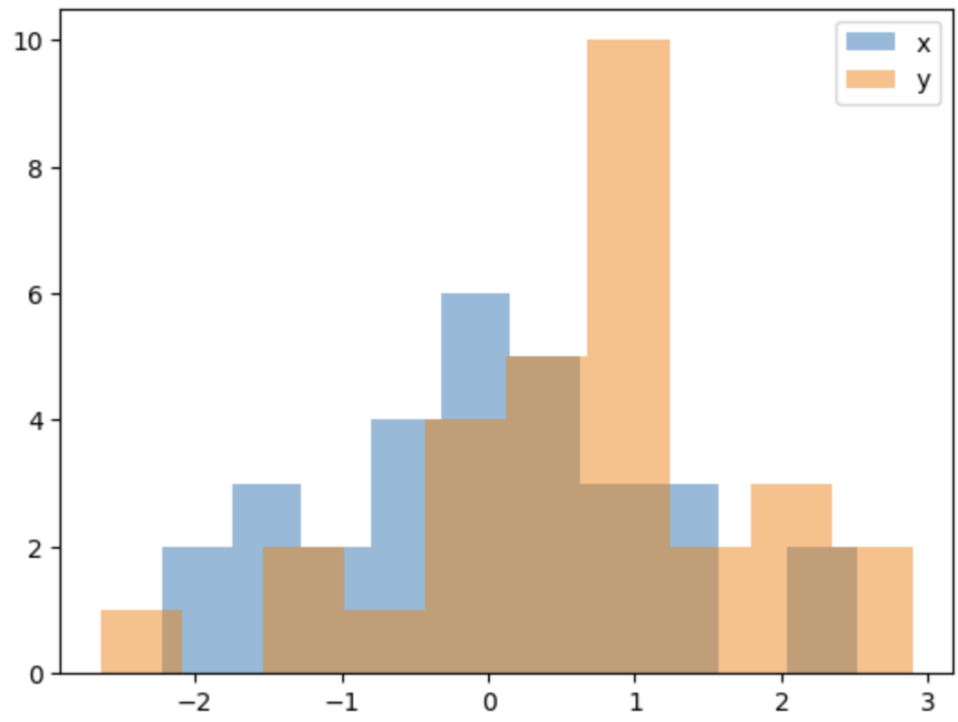

Рисунок 3.3 — Выборки в момент, когда гипотеза была отвергнута

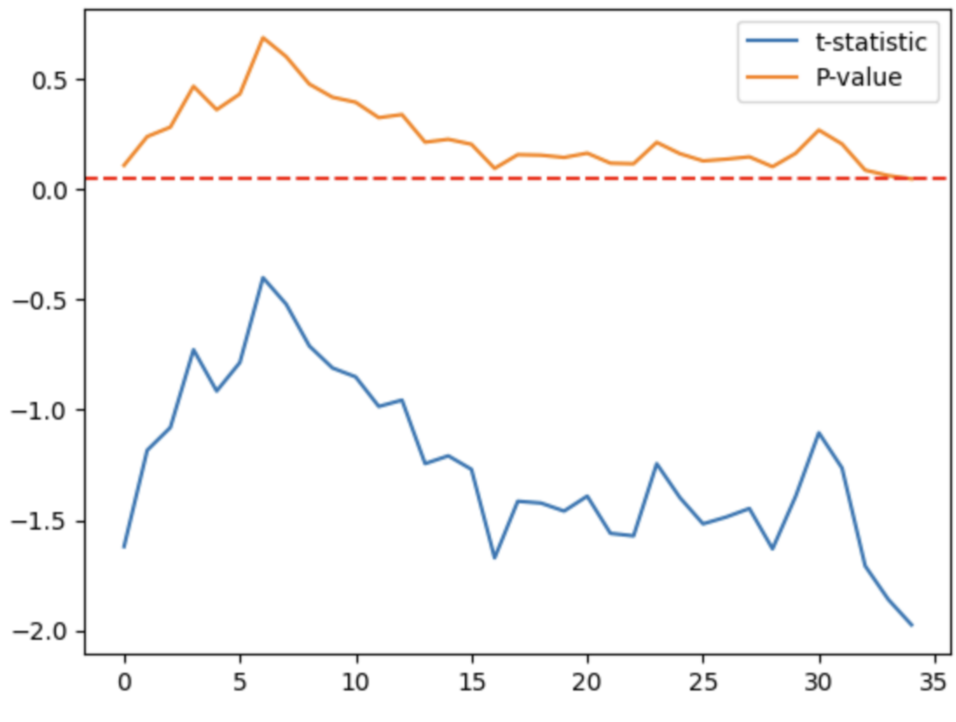

Рисунок 3.4 — Динамика изменения значений статистики критерия и P-value для всех итераций проверки гипотезы о мат. ожидании при увеличении объёмов выборок

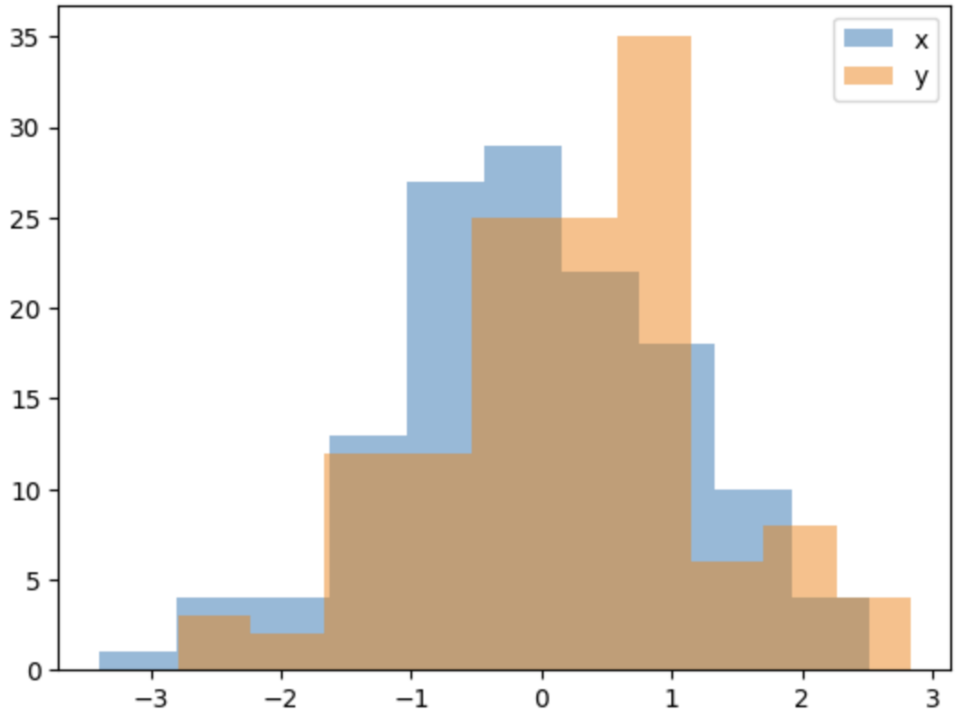

Рисунок 3.5 — Выборки в момент, когда гипотеза была отвергнута

ЗАКЛЮЧЕНИЕ

В рамках лабораторной работы была изучена модель полиномиальной регрессии и регуляризация. Все поставленные задачи были выполнены.

1. Сгенерировать две независимые выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ с нормальные законом распределения и с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответствено;
2. Осуществлена проверка гипотезы H_0 о соответствии выборок нормальному закону распределения;
3. Осуществлена проверка гипотезы H_0 о принадлежности выборок одной генеральной совокупности;
4. Осуществлена проверка гипотезы $H_0 : a_1 = a_2$ против альтернативы $H_1 : a_1 \neq a_2$;
5. Произведён сдвиг вправо всех элементов второй выборки на величину $\delta = 0.01$ и осуществлена проверку гипотезы $H_0 : a_1 = a_2$ до тех пор, пока гипотеза H_0 не будет отвергнута;
6. Для второй выборки назначено a_2 равным середине пройденного отрезка из пункта 5 и постепенно увеличивалось число элементов в выборках и осуществлялась проверку гипотезы $H_0 : a_1 = a_2$ до тех пор, пока гипотеза H_0 не будет отвергнута
7. Рассчитаны 95% доверительные интервалы для математических ожиданий двух выборок в момент, когда гипотеза H_0 была отвергнута в пунктах 5 и 6.