**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»

# Отчёт по лабораторной работе №3 по дисциплине "Методы машинного обучения"

**Тема** Проверка гипотезы о математическом ожидании - Две выборки

**Студент** Варламова Е. А.

**Группа** ИУ7-23М

**Оценка (баллы)** _____

**Преподаватели** Солодовников Владимир Игоревич

Москва — 2024 г.

# СОДЕРЖАНИЕ

# 1 | Теоретическая часть

Статистическая гипотеза – гипотеза о виде распределения и свойствах случайной величины, которые можно подтвердить или опровергнуть применением статистических методов к данным выборки. Нулевая гипотеза – принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями. Нулевая гипотеза H0 считается верной, пока нельзя доказать обратное. В случае, если нулевая гипотеза отвергается, принимается альтернативная.

Проверка статистических гипотез является важным инструментом в области статистики и исследований, она позволяет делать выводы о наличии или отсутствии статистически значимых различий между наборами данных и принимать обоснованные решения на основе этих данных.

Целью данной лабораторной работы является применение методики проверки статистических гипотез при исследовании двух выборок.

Для этого необходимо решить следующие задачи:

— формализовать задачу;

— описать методику проверки статистических гиптоез;

— привести особенности реализации ПО, решающего поставленную задачу;

— провести исследование зависимости изменения значений статистики критерия и $P - value$ для всех итераций проверки гипотезы при изменяющемся математическом ожидании одной из выборок и при изменении размеров выборки.

## 1.1 Постановка задачи

Сгенерировать две независимые выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ с нормальным законом распределения и с параметрами $(a_1, \sigma_1^2)$ и $(a_2, \sigma_2^2)$ соответственно. Изначально $a_1 = a_2$ и $\sigma_1^2 = \sigma_2^2$, число элементов $n = m = 30$. Для полученных

выборок предполагаем, что обе дисперсии неизвестны, но они равны между собой.

1. Осуществить проверку гипотезы $H_0$ о соответствии выборок нормальному закону распределения.

2. Осуществить проверку гипотезы $H_0 : a_1 = a_2$ против альтернативы $H_1 : a_1 \neq a_2$.

3. Производить сдвиг вправо математического ожидания второй выборки $a_2$ на величину $\delta = 0.01(a_2 = a_2+)$ и осуществлять проверку гипотезы $H_0 : a_1 = a_2$ до тех пор, пока гипотеза $H_0$ не будет отвергнута.

4. Для второй выборки назначить $a_2$ равным середине пройденного отрезка из пункта 3. 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.

5. Рассчитать 95% доверительные интервалы для математических ожиданий двух выборок в момент, когда гипотеза $H0$ была отвергнута в пунктах 3 и 4.

Дополнительное представление результатов:

— Вывести на экран гистограммы двух выборок;

— Отобразить в виде графиков динамику изменения значений статистики критерия и $P - value$ для всех итераций проверки гипотезы из пунктов 3 и 4.

## 1.2 Методика проверки статистических гипотез

Пусть задана случайная выборка $X^m = x_1, \ldots, x_m$ – последовательность $m$ объектов из множества $X$. Предполагается, что на множестве $X$ существует некоторая неизвестная вероятностная мера .

1. Формулируются нулевая $H_0$ и альтернативная $H_1$ гипотезы о распределении вероятностей на множестве $X$.

2. Задаётся некоторая статистика (произвольная измеримая функция выборки, которая не зависит от неизвестных параметров распределения) $T : X^M \longrightarrow R$ , для которой в условиях справедливости гипотезы $H_0$ выводится функция распределения и/или плотность распределения.

3. Фиксируется уровень значимости – допустимая для данной задачи вероятность того, что гипотеза на самом деле верна, но будет отвергнута процедурой проверки. Это должно быть достаточно малое число $\alpha$.

4. На множестве допустимых значений статистики $T$ выделяется критическое множество $\omega$ наименее вероятных значений статистики $T$, такое, что $P\{T \in \omega_\alpha | H_0\} = \alpha$.

5. Собственно статистический тест (статистический критерий) заключается в проверке условия:

   — Если $T(X^m) \in \omega_\alpha$ , то делается вывод «данные противоречат нулевой гипотезе при уровне значимости $\alpha$». Гипотеза отвергается.

   — Если $T(X^m) \in \omega_\alpha$ , то делается вывод «данные не противоречат нулевой гипотезе при уровне значимости $\alpha$». Гипотеза принимается.

## 1.3 P-value

P-value или p-значение – одна из ключевых величин, используемых в статистике при тестировании гипотез. Она показывает вероятность получения наблюдаемых результатов при условии, что нулевая гипотеза верна, или вероятность ошибки в случае отклонения нулевой гипотезы.

## 1.4 Доверительный интервал

В математической статистике – интервал, в пределах которого с заданной вероятностью лежат выборочные оценки статистических характеристик генеральной совокупности.

Если оценку среднего требуется связать с определённой вероятностью, то интересующий параметр генеральной совокупности нужно оценивать не одним числом, а интервалом. Доверительным интервалом называют интервал, в котором с определённой вероятностью $P$ находится значение оцениваемого показателя генеральной совокупности.

# 2 | Практическая часть

## 2.1 Выбор средств разработки

В качестве языка программирования был использован язык Python, поскольку этот язык кроссплатформенный и для него разработано огромное количество библиотек и модулей, решающих разнообразные задачи.

В частности, имеются библиотеки, включающие в себя функции проверки статистических гипотез в библиотеке [1].

Для создания графиков была выбрана библиотека matplotlib [2], доступная на языке Python, так как она предоставляет удобный интерфейс для работы с данными и их визуализации.

## 2.2 Исследование ПО

В листинге 2.1 представлен код, сдвигающий математическое ожидание одной из выборок на $\delta = 0.01$ относительно другой до тех пор, пока гипотеза $H_0$ о равенстве математических ожиданий не будет отвергнута. Кроме того, считаются доверительные интервалы для каждой из выборок после отвержения гипотезы $H_0$.

Листинг 2.1: код сдвига математическое ожидание одной из выборок относительно другой

```
fig, ax = plt.subplots()

delta = 0.01
interval = 0
y_s = [y]
_, p_value_equal_means = stats.ttest_ind(x, y_s[-1], equal_var=True)
means_diff = [np.mean(x) - np.mean(y_s[-1])]
values = [p_value_equal_means]
ax.hist(x, alpha=0.5, label='                X')
_, _, h = ax.hist(y_s[-1], alpha=0.5, label='                Y')

while p_value_equal_means >= 0.05:
    interval += delta
    y_s.append(np.random.normal(loc=a_2 + interval, scale=sigma_2, size=m))
    _, p_value_equal_means = stats.ttest_ind(x, y_s[-1], equal_var=True)
    values.append(p_value_equal_means)
    means_diff.append(np.mean(x) - np.mean(y_s[-1]))

def update(iternum):
    global h
    h.remove()
    plt.title("p_value = {:.3f}, interval = {}".format(values[iternum],
        delta * iternum))
    _, _, h = ax.hist(y_s[iternum], alpha=0.5)

ani = animation.FuncAnimation(fig, update, frames=len(y_s), interval=400)
ani.save('animated_plot_3.gif', writer='pillow')
print("                = {:.3f}, p-value = {:.3f}".format(interval,
    p_value_equal_means))
ci_x = stats.norm.interval(0.95, loc=np.mean(x), scale=stats.sem(x))
ci_y = stats.norm.interval(0.95, loc=np.mean(y_s[-1]), scale=stats.sem(y_s
    [-1]))
print("                                    X = {};\n
                Y = {}".format(ci_x, ci_y))

plt.clf()
plt.plot([i * delta for i in range(len(y_s))], means_diff, label='
                                ')
plt.plot([i * delta for i in range(len(y_s))], values, label='P-value')
plt.legend()
plt.savefig("diffs_3.png")
```

В листинге 2.2 представлен код, в котором увеличивается размер выборок на 200 до тех пор, пока гипотеза $H_0$ о равенстве математических ожиданий не будет отвергнута. При этом математическое ожидание выборок отличается на половину интервала, вычисленного в предыдущем пункте. Кроме того, считаются доверительные интервалы для каждой из выборок после отвержения гипотезы

$H_0$.

Листинг 2.2: код увеличения размера выборок

```
fig, ax = plt.subplots()
y_s = [y]
x_s = [x]
n_ = n
m_ = m
_, p_value_equal_means = stats.ttest_ind(x_s[-1], y_s[-1], equal_var=True)
_, _, h1 = ax.hist(x_s[-1], alpha=0.5)
_, _, h2 = ax.hist(y_s[-1], alpha=0.5)
values = [p_value_equal_means]
sizes = [n_]
means_diff = [np.mean(x) - np.mean(y)]
while p_value_equal_means >= 0.05:
    n_ += 200
    m_ += 200
    x_s.append(np.random.normal(loc=a_1, scale=sigma_1, size=n_))
    y_s.append(np.random.normal(loc=a_2 + interval / 2, scale=sigma_2, size=
        m_))
    _, p_value_equal_means = stats.ttest_ind(x_s[-1], y_s[-1], equal_var=
        True)
    values.append(p_value_equal_means)
    sizes.append(n_)
    means_diff.append(np.mean(x_s[-1]) - np.mean(y_s[-1]))
def update(iternum):
    global h1
    global h2
    h1.remove()
    h2.remove()
    plt.title("p_value = {:.3f}, size = {}".format(values[iternum], sizes[
        iternum]))
    _, _, h1 = ax.hist(x_s[iternum], alpha=0.5)
    _, _, h2 = ax.hist(y_s[iternum], alpha=0.5)

ani = animation.FuncAnimation(fig, update, frames=len(y_s), interval=400)
ani.save('animated_plot_4.gif', writer='pillow')
ci_x = stats.norm.interval(0.95, loc=np.mean(x_s[-1]), scale=stats.sem(x_s
    [-1]))
ci_y = stats.norm.interval(0.95, loc=np.mean(y_s[-1]), scale=stats.sem(y_s
    [-1]))
print("                                    X = {};\n
                  Y = {}".format(ci_x, ci_y))
plt.clf()
plt.plot(sizes, means_diff, label='                                    ')
plt.plot(sizes, values, label='P-value')
plt.legend()
plt.savefig("diffs_4.png")
```

Были сгенерированы выборки, представленные на рисунке 2.1.

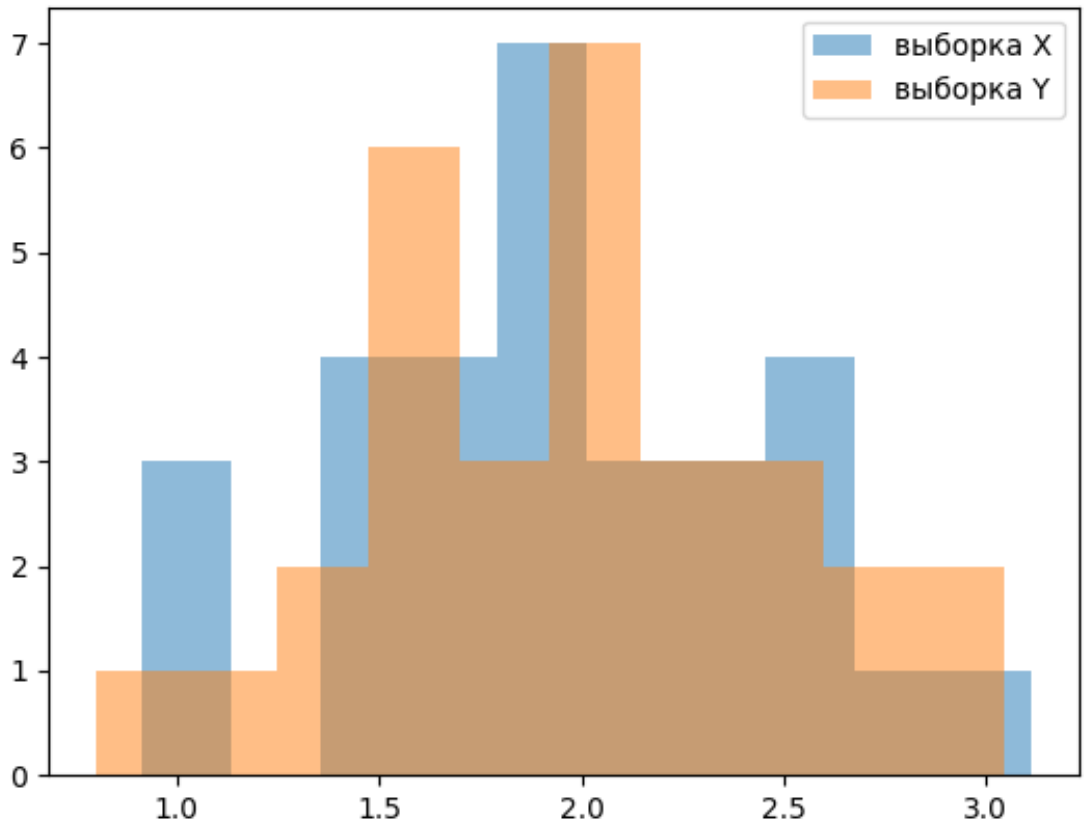


Рис. 2.1:

Для них:

1. математическое ожидание $a_1 = a_2 = 2$, дисперсия $\sigma_1 = \sigma_2 = 0.5$;
2. размер выборок равен 30;
3. p-value для гипотезы о нормальном распределении для выборки $X$ составляет 0.9976;
4. p-value для гипотезы о нормальном распределении для выборки  составляет 0.9915;
5. p-value для гипотезы о равенстве математических ожиданий составляет 0.9042;

Был проведен сдвиг математического ожидания выборки $Y$ с шагом 0.01 и проверялась гипотеза о равенстве математических ожиданий. Результаты такого сдвига приведены на рисунке 2.2.

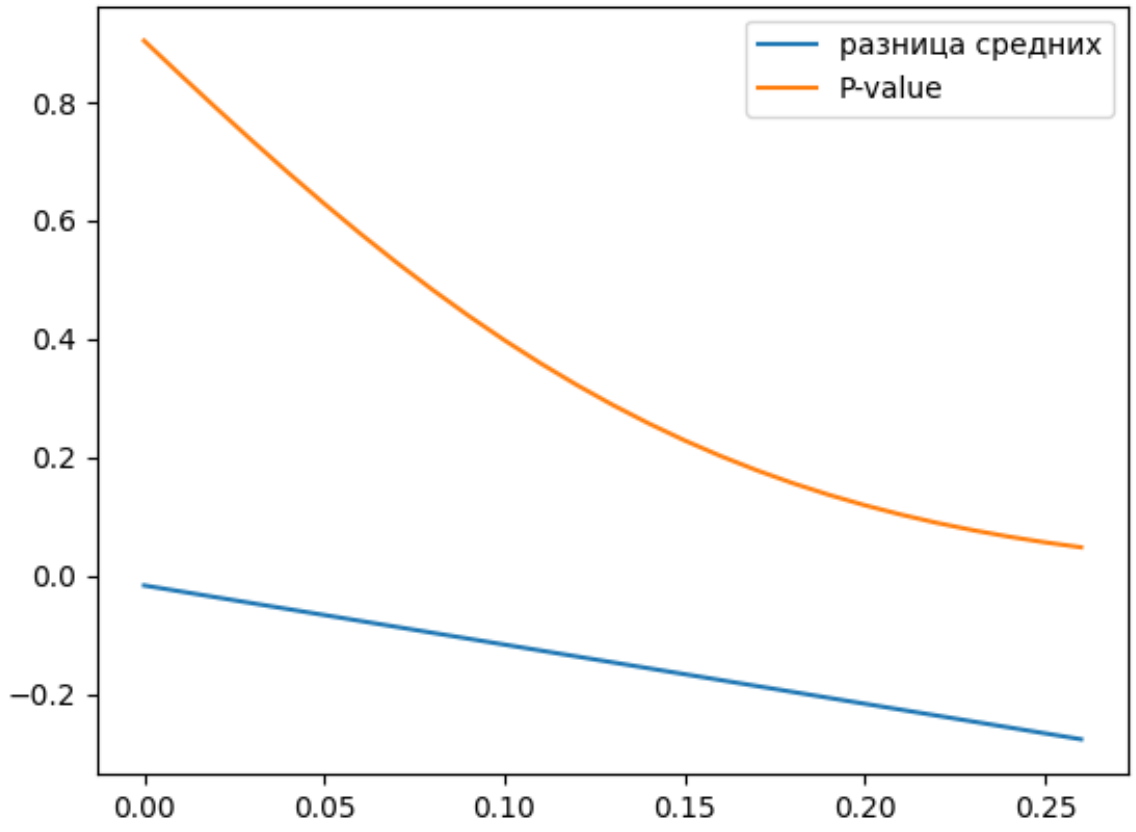


Рис. 2.2: Заивисмость p-value от расстояния между выборками

Видно, что p-value убывает с увеличением расстояния между математическими ожиданиями выборок. Когда оно достигает критического значения, равного 0.05, гипотеза $H_0$ о равенстве отвергается.

Доверительные интервалы после отвержения гипотезы $H_0$ следующие:

— интервал выборки $X = (1.76, 2.14)$;

— интервал выборки $Y = (2.04, 2.42)$.

Далее был увеличен размер выборок на 200 до тех пор, пока гипотеза $H0$ о равенстве математических ожиданий не будет отвергнута. При этом математическое ожидание выборок отличается на половину интервала, вычисленного в предыдущем пункте. Результаты приведены на рисунке 2.3.

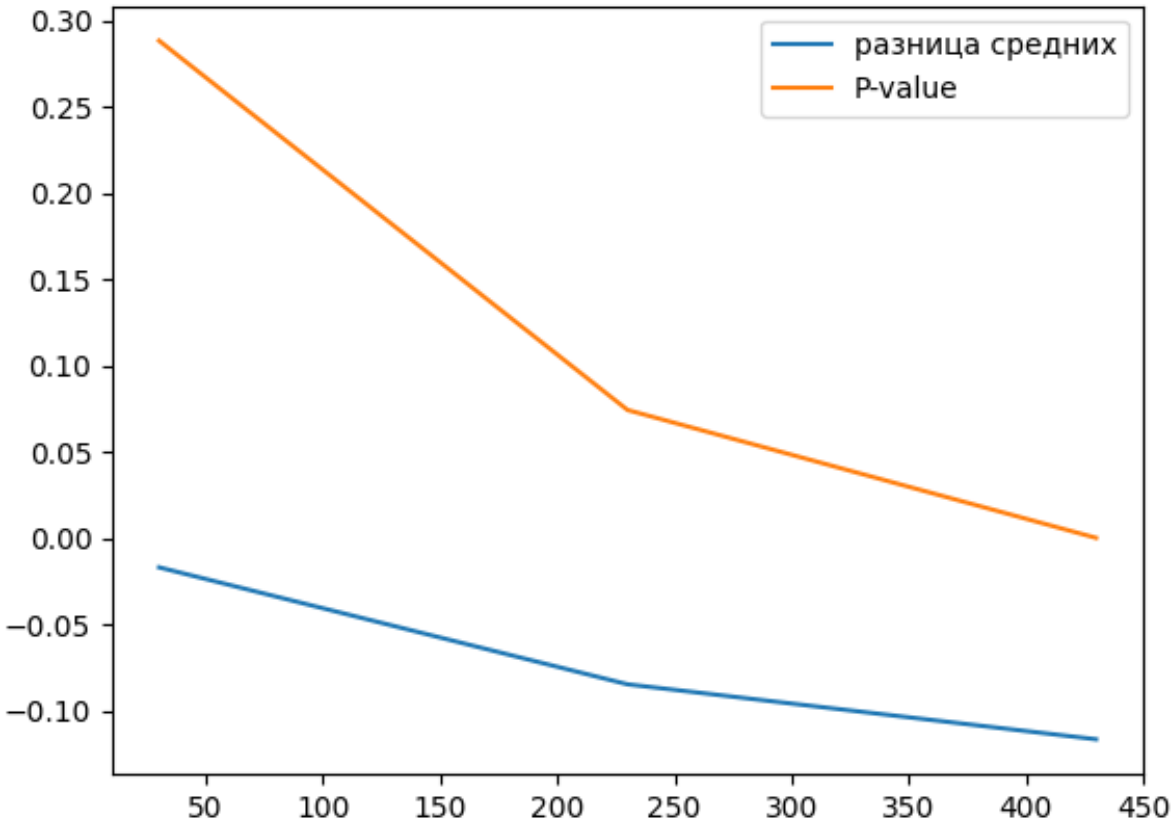


Рис. 2.3: Увеличение размера выборок

Видно, что p-value убывает с увеличением размера выборок. Когда оно достигает критического значения, равного 0.05, гипотеза $H_0$ о равенстве отвергается.

Доверительные интервалы после отвержения гипотезы $H_0$ следующие:

— интервал выборки $X = (1.96, 2.05)$;

— интервал выборки $Y = (2.07, 2.17)$.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. *Virtanen P.*, *Gommers R.*, *Oliphant T. E.* SciPy: Fundamental Algorithms for Scientific Computing in Python. — 2020. — DOI: 10.1038/s41592-019-0686-2.
2. Библиотека визуализации данных matplotlib [Электронный ресурс]. — Режим доступа: URL: https://matplotlib.org (дата обращения: 13.12.2023).