



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 лабораторным работам №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	3
1.1	Постановка задачи	3
1.2	Методика проверки статистических гипотез	4
1.3	P-value	5
1.4	Доверительный интервал	5
2	Практическая часть	6
2.1	Выбор средств разработки	6
2.2	Исследование ПО	6
	СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	12

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 H_0 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 гипотез;
- 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, \dots, x_n и y_1, \dots, y_m с нормальным законом распределения и с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответственно. Изначально $a_1 = a_2$ и $\sigma_1^2 = \sigma_2^2$, число элементов $n = m = 30$. Для полученных

выборки 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 H_0 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 $X^m = x_1, \dots, x_m$ — 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) $T : X^M \rightarrow R$, 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 α .
4. На множестве допустимых значений статистики T выделяется критическое множество ω 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 α ». Гипотеза отвергается.
 - Если $T(X^m) \in \omega_\alpha$, то делается вывод «данные не противоречат нулевой гипотезе при уровне значимости α ». Гипотеза принимается.

1.3 P-value

P-value или р-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

```

1 fig , ax = plt.subplots()
2
3 delta = 0.01
4 interval = 0
5 y_s = [y]
6 _, p_value_equal_means = stats.ttest_ind(x, y_s[-1], equal_var=True)
7 means_diff = [np.mean(x) - np.mean(y_s[-1])]
8 values = [p_value_equal_means]
9 ax.hist(x, alpha=0.5, label='X')
10 _, _, h = ax.hist(y_s[-1], alpha=0.5, label='Y')
11
12 while p_value_equal_means >= 0.05:
13     interval += delta
14     y_s.append(np.random.normal(loc=a_2 + interval, scale=sigma_2, size=m))
15     _, p_value_equal_means = stats.ttest_ind(x, y_s[-1], equal_var=True)
16     values.append(p_value_equal_means)
17     means_diff.append(np.mean(x) - np.mean(y_s[-1]))
18
19 def update(iternum):
20     global h
21     h.remove()
22     plt.title("p_value = {:.3f}, interval = {}".format(values[iternum],
23                                                         delta * iternum))
24     _, _, h = ax.hist(y_s[iternum], alpha=0.5)
25
26 ani = animation.FuncAnimation(fig, update, frames=len(y_s), interval=400)
27 ani.save('animated_plot_3.gif', writer='pillow')
28 print("interval = {:.3f}, p-value = {:.3f}".format(interval,
29                                                         p_value_equal_means))
30 ci_x = stats.norm.interval(0.95, loc=np.mean(x), scale=stats.sem(x))
31 ci_y = stats.norm.interval(0.95, loc=np.mean(y_s[-1]), scale=stats.sem(y_s
32 [-1]))
33 print("X = {};\n Y = {}".format(ci_x, ci_y))
34
35 plt.clf()
36 plt.plot([i * delta for i in range(len(y_s))], means_diff, label='
37 ')
38 plt.plot([i * delta for i in range(len(y_s))], values, label='P-value')
39 plt.legend()
40 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

```
1 fig, ax = plt.subplots()
2 y_s = [y]
3 x_s = [x]
4 n_ = n
5 m_ = m
6 _, p_value_equal_means = stats.ttest_ind(x_s[-1], y_s[-1], equal_var=True)
7 _, _, h1 = ax.hist(x_s[-1], alpha=0.5)
8 _, _, h2 = ax.hist(y_s[-1], alpha=0.5)
9 values = [p_value_equal_means]
10 sizes = [n_]
11 means_diff = [np.mean(x) - np.mean(y)]
12 while p_value_equal_means >= 0.05:
13     n_ += 200
14     m_ += 200
15     x_s.append(np.random.normal(loc=a_1, scale=sigma_1, size=n_))
16     y_s.append(np.random.normal(loc=a_2 + interval / 2, scale=sigma_2, size=
17         m_))
18     _, p_value_equal_means = stats.ttest_ind(x_s[-1], y_s[-1], equal_var=
19         True)
20     values.append(p_value_equal_means)
21     sizes.append(n_)
22     means_diff.append(np.mean(x_s[-1]) - np.mean(y_s[-1]))
23 def update(iternum):
24     global h1
25     global h2
26     h1.remove()
27     h2.remove()
28     plt.title("p_value = {:.3f}, size = {}".format(values[iternum], sizes[
29         iternum]))
30     _, _, h1 = ax.hist(x_s[iternum], alpha=0.5)
31     _, _, h2 = ax.hist(y_s[iternum], alpha=0.5)
32 ani = animation.FuncAnimation(fig, update, frames=len(y_s), interval=400)
33 ani.save('animated_plot_4.gif', writer='pillow')
34 ci_x = stats.norm.interval(0.95, loc=np.mean(x_s[-1]), scale=stats.sem(x_s
35     [-1]))
36 ci_y = stats.norm.interval(0.95, loc=np.mean(y_s[-1]), scale=stats.sem(y_s
37     [-1]))
38 print("
39     X = {};\ n
40     Y = {}".format(ci_x, ci_y))
41 plt.clf()
42 plt.plot(sizes, means_diff, label='
43     ')
44 plt.plot(sizes, values, label='P-value')
45 plt.legend()
46 plt.savefig("diffs_4.png")
```


Были сгенерированы выборки, представленные на рисунке 2.1.

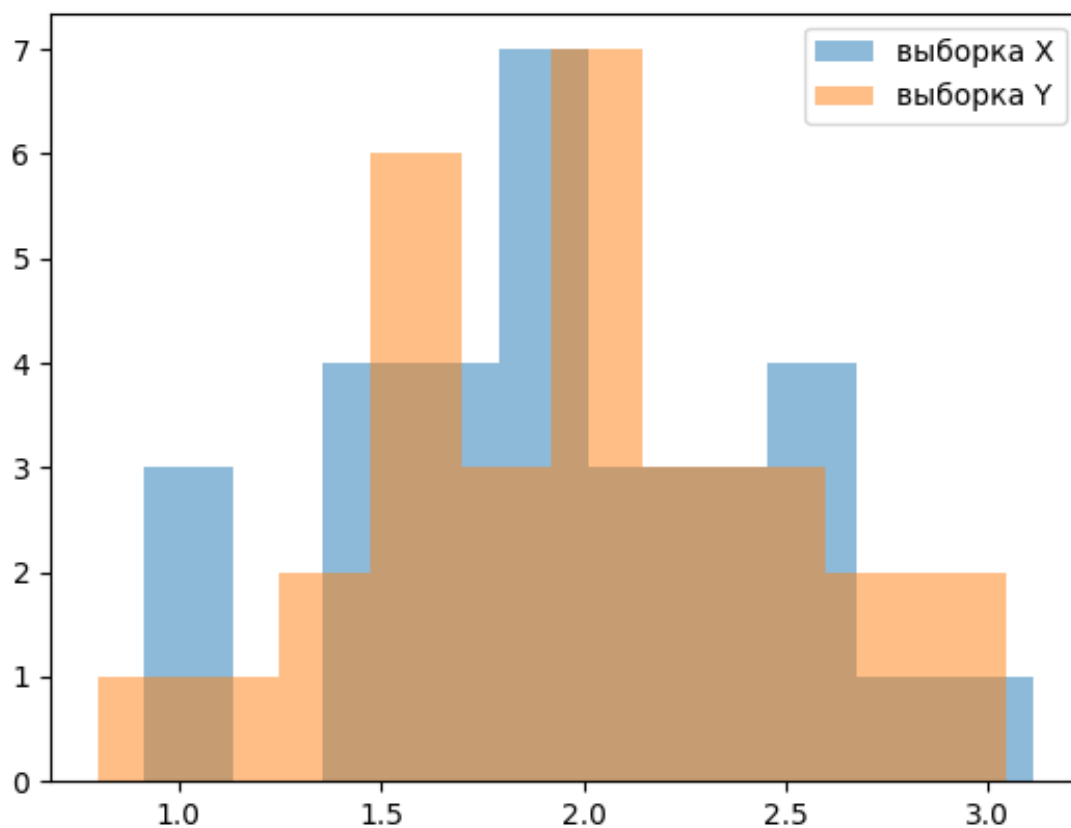


Рис. 2.1:

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 Y 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.

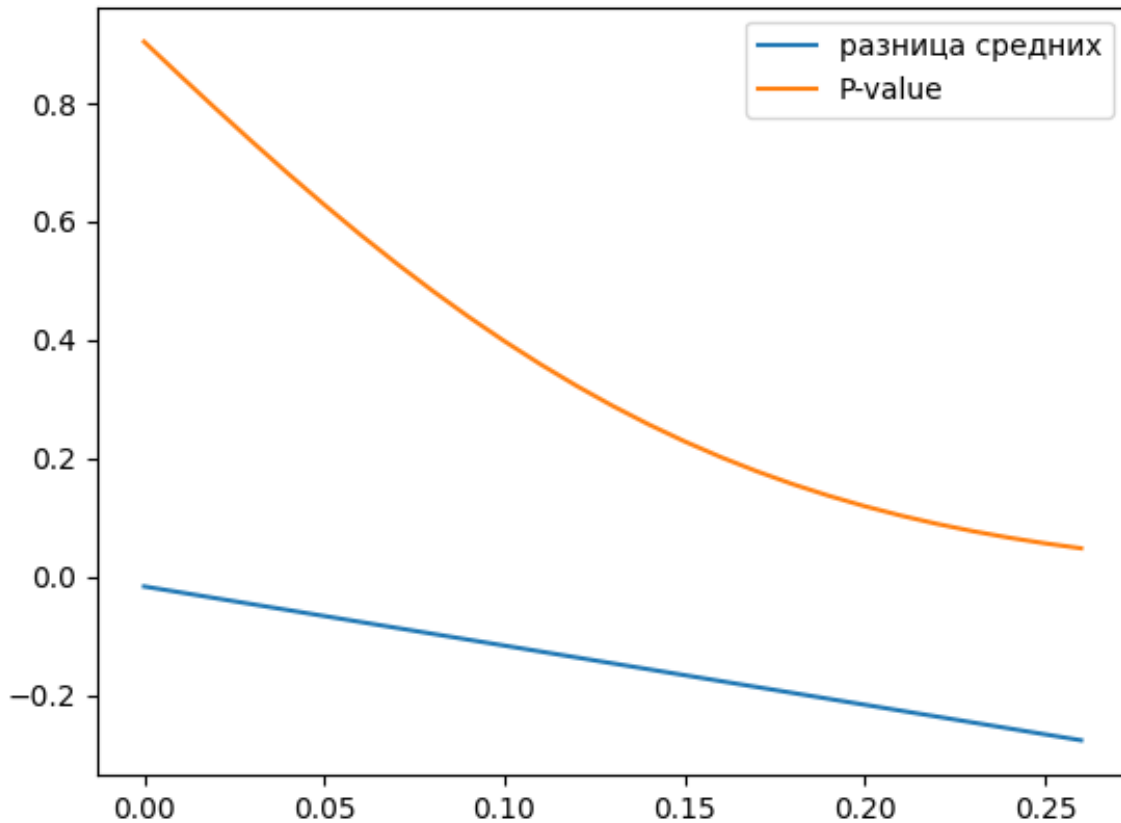


Рис. 2.2: Заивисмость p-value от расстояния между выборками

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 H_0 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.

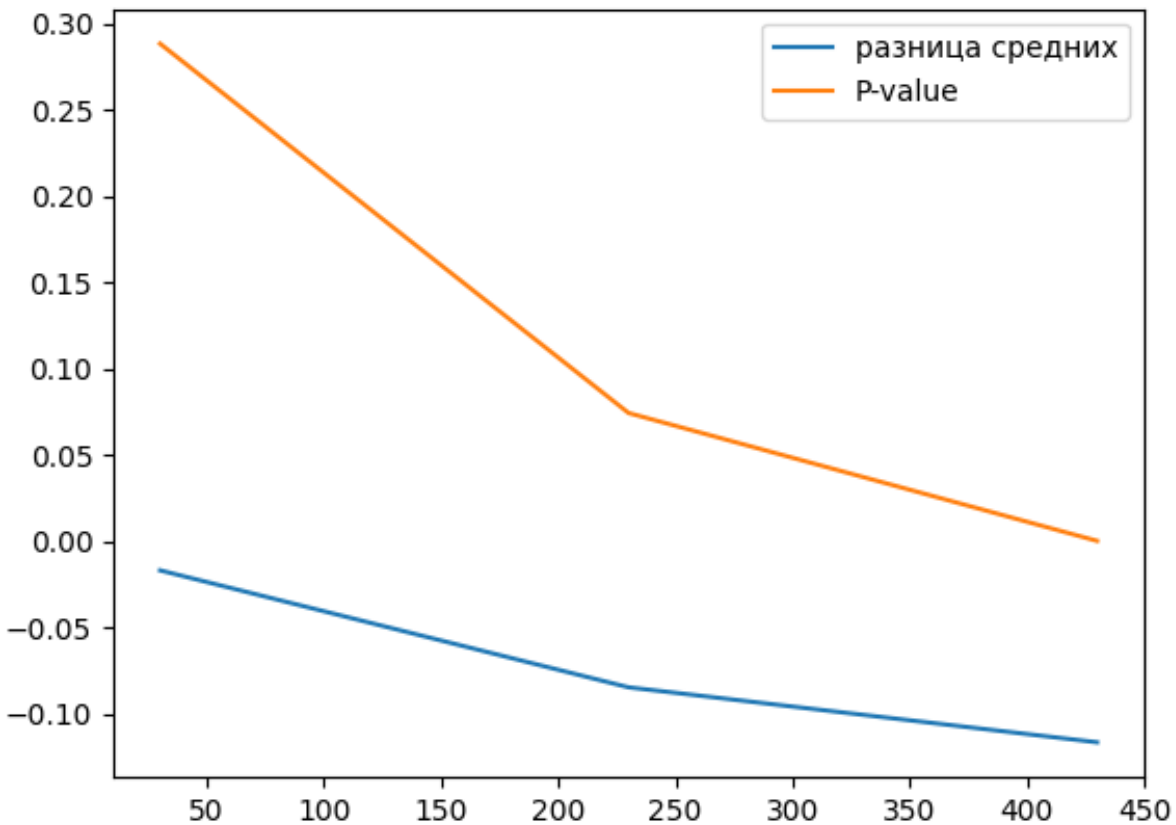


Рис. 2.3: Увеличение размера выборок

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).