Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

Отчет

по лабораторной работе № 4

“Anova”

Автор:

Кузнецова Таисия J3114,

Долбенко Олеся J3113

Демьянов Федор J3114

Факультет: ФТИИ

Введение

Цель работы: Освоить методы проверки статистических гипотез о равенстве средних значений в нескольких группах данных.

Задачи:

1. Сгенерировать две выборки данных из нормального распределения, каждая из которых состоит из трёх подгрупп
2. Визуализировать распределения с помощью графиков плотности (KDE) и объединённых распределений для каждой выборки.
3. Реализовать парный Z-тест для проверки гипотезы о равенстве средних значений при известных дисперсиях.
4. Реализовать парный t-тест для случая неизвестных, но равных дисперсий.
5. Провести оба вида парных тестов для всех пар групп в обеих выборках.
6. Применить однофакторный дисперсионный анализ (ANOVA) для проверки гипотезы о равенстве всех средних значений в группах.
7. Сравнить полученные результаты

Ход работы:

1. Мы сгенерировали две выборки из нормального распределения. Каждая выборка включала по три группы одинакового объёма:

– в первой выборке группы имели близкие математические ожидания;

– во второй — сильно отличающиеся.

2. Для визуального анализа распределений мы построили графики плотности (KDE) как для каждой группы, так и для объединённых данных внутри каждой выборки.

3. Затем мы реализовали парные Z-тесты для всех пар групп, предполагая известную и одинаковую дисперсию. Результаты показали различия в выборке с сильно отличающимися средними.

4. Далее мы провели парные t-тесты при неизвестных, но равных дисперсиях, что также позволило выявить различия между группами.

5. Для одновременного анализа всех трёх групп мы применили однофакторный дисперсионный анализ (ANOVA), который подтвердил различия в выборках, особенно во второй.

6. Мы оценили вычислительную эффективность обоих подходов и пришли к выводу, что ANOVA эффективнее при большом числе групп. При этом парные тесты предоставляют более детальную информацию о различиях между конкретными парами.

Теоретическая часть

Нормальное распределение — это непрерывное распределение вероятностей с плотностью:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

где:

μ − математическое ожидание

σ^2 − дисперсия

Парный Z-тест (при известных дисперсиях)

Z-тест применяется для проверки гипотезы:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

При известных и равных дисперсиях σ^2, статистика рассчитывается по формуле:

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{2\sigma^2/n}}$$

где:

$\overline{X_1}, \overline{X_2}$ − среднеен значение выборок, n − объем каждой выборки

Результат сравнивается с критическим значением стандартного нормального распределения $Z_{crit} = \pm z_a / 2$

Парный t-тест (при неизвестных, но равных дисперсиях)

Если дисперсия неизвестна, но предполагается одинаковой, используется двухвыборочный t-тест:

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_P\sqrt{2/n}}$$

где S_P — объединённая (пулевая) оценка стандартного отклонения:

$$S_P = \sqrt{\frac{(n-1)\,S_1^2 + (n-1)S_2^2}{2n-2}}$$

и S_1^2, S_2^2 — выборочные дисперсии. Статистика t сравнивается с критическим значением распределения Стьюдента с 2n–2 степенями свободы.

Однофакторный дисперсионный анализ (ANOVA)

ANOVA применяется для проверки гипотезы о равенстве средних более чем в двух группах

$$H_0:\ \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$$

$$H_1:\ \text{хотя бы одно из } \mu_i \text{ отличается}$$

F-статистика рассчитывается как:

$$F = \frac{\text{МЕЖГРУППОВАЯ ДИСПЕРСИЯ } (MS\_\ between)}{\text{ВНУТРИГРУППОВАЯ ДИСПЕРСИЯ } (MS\_\ bithim)}$$

где:

MS_between= $\frac{SSbetwen}{k-1}$

MS_within= $\frac{SSwithin}{N-k}$

k— число групп, N — общее количество наблюдений.

Если F превышает критическое значение из F-распределения, гипотеза H_0 отклоняется.

Уровень значимости и p-значение

Во всех тестах применяется уровень значимости α=0.05 Если рассчитанное p-значение < α, то нулевая гипотеза отклоняется.

Распределение Стьюдента (t-распределение) — это непрерывное распределение вероятностей, используемое при проверке статистических гипотез в случае, когда: объём выборки мал (обычно менее 30 наблюдений), дисперсия неизвестна,
данные имеют нормальное распределение.
Это распределение особенно важно при использовании t-тестов, которые оценивают, различаются ли средние значения двух групп.

Плотность вероятности t-распределения определяется следующей формулой:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

где:

t — значение переменной,
ν (ню) — число степеней свободы (обычно n – 1, где n — размер выборки),
Г(x) — гамма-функция (обобщение факториала).

Основные свойства распределения Стьюдента:

Симметрично относительно нуля.

Более "толстые хвосты", чем у нормального распределения — это учитывает большую неопределённость при малом числе наблюдений.
При увеличении числа степеней свободы (ν → ∞) t-распределение стремится к стандартному нормальному распределению N(0,1).

Практическая часть

1. **Генерация выборок**

На первом этапе мы сгенерировали две выборки, каждая из которых состоит из трёх подгрупп (групп). Все группы были сгенерированы из нормального распределения с одинаковой дисперсией (стандартное отклонение σ = 1), но с разными средними значениями.

- Первая выборка: три группы с близкими математическими ожиданиями (около 5.0), что позволяет моделировать ситуацию, в которой различия между группами минимальны.

- Вторая выборка: три группы с существенно различающимися средними (например, 2.0, 5.0 и 8.0), что моделирует явно выраженные различия между группами при той же дисперсии.

Размер каждой подвыборки составил 100 элементов

```
np.random.seed(42)

# Общие параметры
n = 100  # Размер каждой подвыборки
std = 1  # Стандартное отклонение
var = std**2

# Первая выборка (близкие средние)
group1 = np.random.normal(loc=5.0, scale=std, size=n)
group2 = np.random.normal(loc=5.2, scale=std, size=n)
group3 = np.random.normal(loc=4.9, scale=std, size=n)

# Вторая выборка (сильно разные средние )
group4 = np.random.normal(loc=2.0, scale=std, size=n)
group5 = np.random.normal(loc=5.0, scale=std, size=n)
group6 = np.random.normal(loc=8.0, scale=std, size=n)
```

2. **График KDE**

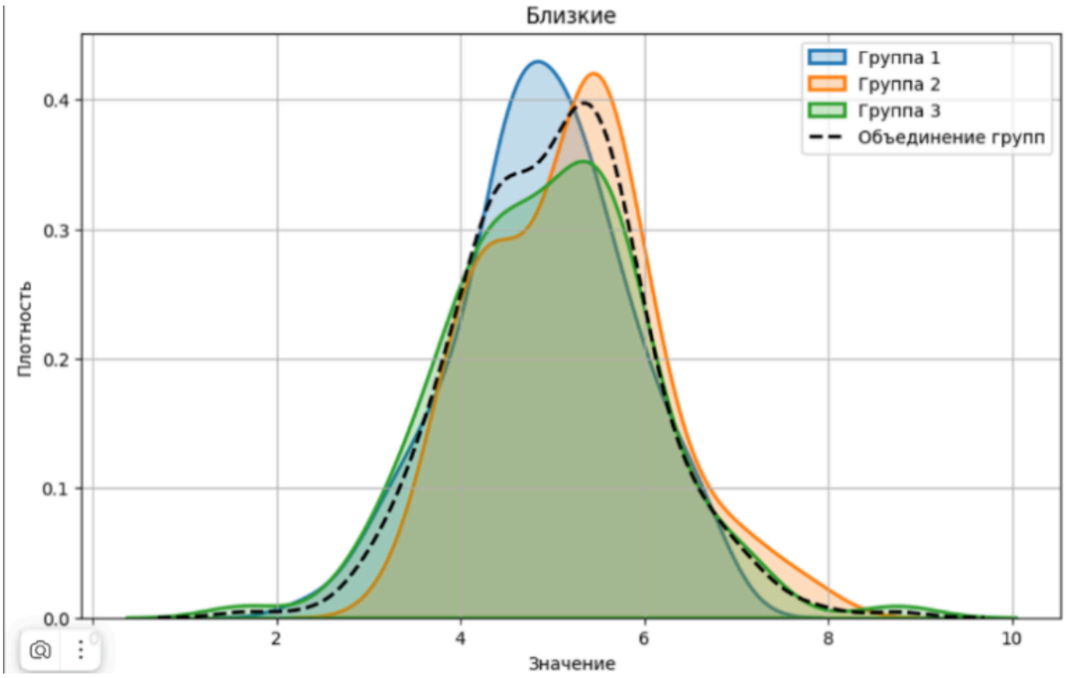
Для наглядной проверки различий между группами мы построили графики оценки плотности распределения (KDE).

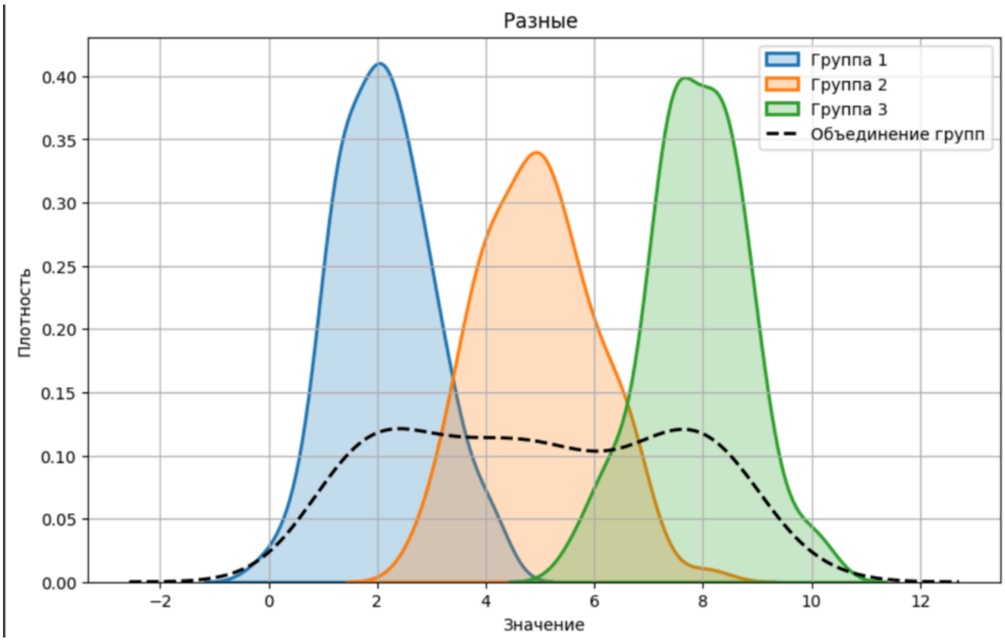
Графики строились отдельно для каждой из двух выборок:

- Для первой выборки с близкими средними значения KDE-групп частично перекрываются, что визуально подтверждает, что различия между группами минимальны.

- Для второй выборки KDE-групп показывают чётко разделённые пики, что говорит о выраженных различиях между средними значениями групп.

Также на каждый график была добавлена кривая объединённого распределения (всех трёх групп), выделенная пунктирной линией. Это позволяет сравнить индивидуальные плотности с общей структурой распределения.





3. Парные тесты

Парный тест для проверки гипотезы о равенстве математических ожиданий при известных дисперсиях. На данном этапе мы реализовали парные Z-тесты для проверки гипотезы Для каждой пары групп внутри выборок была рассчитана Z-статистика.Для каждой пары сравнивался расчётный Z с критическим значением Z на уровне значимости α=0.05. В случае превышения критического значения (в модуле) нулевая гипотеза отклонялась.

```
[ ] from scipy.stats import norm

    def z_test_known_variance(x1, x2, sigma_squared, alpha=0.05):
        n = len(x1)
        mean1, mean2 = np.mean(x1), np.mean(x2)

        z = (mean1 - mean2) / np.sqrt(2 * sigma_squared / n)
        z_crit = norm.ppf(1 - alpha/2)  # двусторонний тест

        reject = abs(z) > z_crit
        return z, z_crit, reject
```

```
    def run_z_tests(groups, label, sigma_squared=1, alpha=0.05):
        print(f"\n=== Z-тесты при известных дисперсиях для {label} ===")
        for i in range(len(groups)):
            for j in range(i + 1, len(groups)):
                z, z_crit, reject = z_test_known_variance(groups[i], groups[j], sigma_squared, alpha)
                result = "ОТКЛОНЯЕМ H₀" if reject else "НЕ отклоняем H₀"
                print(f"Группа {i+1} vs Группа {j+1}: Z = {z:.2f}, Z_crit = ±{z_crit:.2f} → {result}")
```

```
[ ] run_z_tests([group1, group2, group3], "Выборка 1 (близкие ожидания)")
    run_z_tests([group4, group5, group6], "Выборка 2 (разные ожидания)")
```

```
=== Z-тесты при известных дисперсиях для Выборка 1 (близкие ожидания) ===
Группа 1 vs Группа 2: Z = -2.31, Z_crit = ±1.96 → ОТКЛОНЯЕМ H₀
Группа 1 vs Группа 3: Z = -0.49, Z_crit = ±1.96 → НЕ отклоняем H₀
Группа 2 vs Группа 3: Z = 1.82, Z_crit = ±1.96 → НЕ отклоняем H₀

=== Z-тесты при известных дисперсиях для Выборка 2 (разные ожидания) ===
Группа 1 vs Группа 2: Z = -20.06, Z_crit = ±1.96 → ОТКЛОНЯЕМ H₀
Группа 1 vs Группа 3: Z = -40.86, Z_crit = ±1.96 → ОТКЛОНЯЕМ H₀
Группа 2 vs Группа 3: Z = -20.79, Z_crit = ±1.96 → ОТКЛОНЯЕМ H₀
```

В первой выборке (близкие средние) только одна пара (группа 1 и 2) показала статистически значимое отличие, что подтверждает наличие незначительных различий.Во второй выборке (разные средние) все три пары дали очень высокие значения Z и уверенное отклонение гипотезы H0 , что говорит о сильных различиях между группами.

Затем мы провели парные t-тесты Стьюдента, предполагая, что дисперсии выборок неизвестны, но равны.

```python
from scipy.stats import ttest_ind

def run_t_tests(groups, label, alpha=0.05):
    print(f"\n=== t-тесты при неизвестных, но равных дисперсиях для {label} ===")
    for i in range(len(groups)):
        for j in range(i + 1, len(groups)):
            stat, p = ttest_ind(groups[i], groups[j], equal_var=True)
            result = "ОТКЛОНЯЕМ H₀" if p < alpha else "НЕ отклоняем H₀"
            print(f"Группа {i+1} vs Группа {j+1}: t = {stat:.2f}, p = {p:.4f} → {result}")
```

```python
run_t_tests([group1, group2, group3], "Выборка 1 (близкие ожидания)")
run_t_tests([group4, group5, group6], "Выборка 2 (разные ожидания)")
```

```
=== t-тесты при неизвестных, но равных дисперсиях для Выборка 1 (близкие ожидания) ===
Группа 1 vs Группа 2: t = -2.48, p = 0.0141 → ОТКЛОНЯЕМ H₀
Группа 1 vs Группа 3: t = -0.49, p = 0.6275 → НЕ отклоняем H₀
Группа 2 vs Группа 3: t = 1.78, p = 0.0762 → НЕ отклоняем H₀

=== t-тесты при неизвестных, но равных дисперсиях для Выборка 2 (разные ожидания) ===
Группа 1 vs Группа 2: t = -20.51, p = 0.0000 → ОТКЛОНЯЕМ H₀
Группа 1 vs Группа 3: t = -45.19, p = 0.0000 → ОТКЛОНЯЕМ H₀
Группа 2 vs Группа 3: t = -20.88, p = 0.0000 → ОТКЛОНЯЕМ H₀
```

В первой выборке только одна пара (группа 1 и 2) показала значимое отличие ($p < 0.05$), остальные — нет. Это согласуется с результатами Z-теста.
Во второй выборке все пары дали p-значения, близкие к нулю, что подтверждает сильное расхождение средних значений.

Таким образом, t- и Z-тесты показали согласованные выводы. Это означает корректность реализации и хорошее соответствие между теоретическими ожиданиями и практическими результатами.

5. ANOVA

Мы применили однофакторный дисперсионный анализ (ANOVA) для проверки гипотезы о равенстве средних значений во всех трёх группах одновременно.

H_0: математические ожидания всех групп равны,
H_1: хотя бы одно из средних значений отличается от других.

```python
from scipy.stats import f_oneway

def run_anova(groups, label):
    f_stat, p_val = f_oneway(*groups)
    result = "ОТКЛОНЯЕМ H₀" if p_val < 0.05 else "НЕ отклоняем H₀"
    print(f"\n=== ANOVA для {label} ===")
    print(f"F = {f_stat:.2f}, p = {p_val:.4f} → {result}")
```

```python
run_anova([group1, group2, group3], "Выборка 1 (близкие ожидания)")
run_anova([group4, group5, group6], "Выборка 2 (разные ожидания)")
```

```
=== ANOVA для Выборка 1 (близкие ожидания) ===
F = 3.05, p = 0.0490 → ОТКЛОНЯЕМ H₀

=== ANOVA для Выборка 2 (разные ожидания) ===
F = 905.27, p = 0.0000 → ОТКЛОНЯЕМ H₀
```

Выборка 1 (близкие ожидания):
Несмотря на близкие средние, различие оказалось на границе статистической значимости. Это подтверждает, что одна из групп может немного отличаться.
Выборка 2 (разные ожидания):
Значение $p = 0$ это ожидаемый результат так как группы были с явно различающимися средними

ANOVA позволила быстро оценить наличие различий между всеми группами сразу. В отличие от парных тестов, она не указывает, между какими группами есть различия — она лишь проверяет сам факт наличия отличий. Результаты анализа полностью согласуются с результатами парных Z- и t-тестов, что подтверждает корректность расчётов и моделей.

6. Анализ

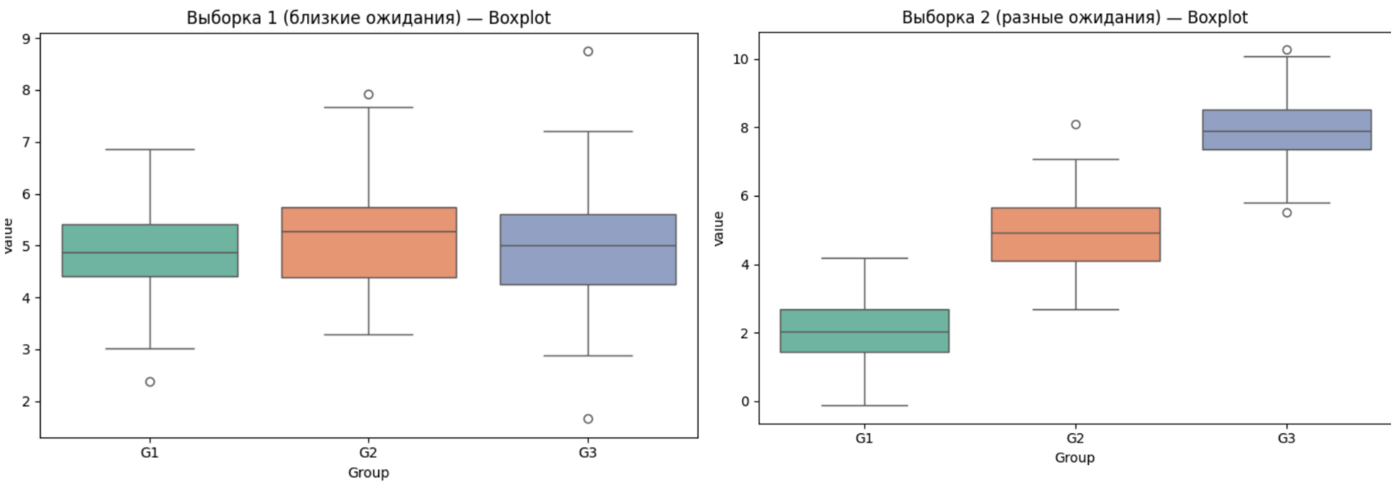
Для сравнения эффективности t-тестов и ANOVA мы замерили время выполнения каждого из методов на одной и той же выборке, состоящей из трёх групп по 100 элементов.

```python
import time

# t-тесты
start = time.time()
run_t_tests([group1, group2, group3], "Timing test")
print("t-тесты заняли:", time.time() - start, "сек")

# ANOVA
start = time.time()
run_anova([group1, group2, group3], "Timing test")
print("ANOVA заняла:", time.time() - start, "сек")
```

```
=== t-тесты при неизвестных, но равных дисперсиях для Timing t
Группа 1 vs Группа 2: t = -2.48, p = 0.0141 → ОТКЛОНЯЕМ H₀
Группа 1 vs Группа 3: t = -0.49, p = 0.6275 → НЕ отклоняем H₀
Группа 2 vs Группа 3: t = 1.78, p = 0.0762 → НЕ отклоняем H₀
t-тесты заняли: 0.011661052703857422 сек

=== ANOVA для Timing test ===
F = 3.05, p = 0.0490 → ОТКЛОНЯЕМ H₀
ANOVA заняла: 0.0014448165893554688 сек
```

Даже при сравнительно небольшом объёме данных однофакторный дисперсионный анализ (ANOVA) работает значительно быстрее, чем серия парных t-тестов. При увеличении числа групп разница в эффективности будет становиться ещё более заметной, поскольку количество t-тестов растёт квадратично, а ANOVA выполняется за постоянное время.

Для наглядной оценки различий между группами мы построили boxplot для обеих выборок



В первой выборке boxplot-графики трёх групп существенно перекрываются. Медианы близки, размеры ящиков и усы схожи. Это соответствует статистическим выводам о слабых различиях между группами.

Во второй выборке ящики отчётливо разнесены по вертикали. Медианы различны, и пересечения почти отсутствуют. Это наглядно подтверждает наличие существенных различий между группами, что согласуется с результатами t-тестов и ANOVA.

Вывод

1. По парным Z- и t-тестам**:**
   В первой выборке с близкими средними отличия были зафиксированы только между одной парой групп. В остальных случаях гипотеза о равенстве средних не отклонялась. Во второй выборке все пары показали статистически значимые различия, что соответствует ожидаемому результату. Z- и t-тесты дали согласованные выводы.

2. По ANOVA:
   Дисперсионный анализ в первой выборке выявил слабое различие между группами ($p \approx 0.049$), что согласуется с результатами парных тестов. Во второй выборке ANOVA уверенно показала наличие различий между всеми группами. Метод оказался эффективным для быстрой проверки общей гипотезы.

3. Сравнение методов:
   ANOVA работает быстрее и удобнее при большом числе групп, но не показывает, между какими именно группами есть различия. Парные тесты дают более детальную информацию, но требуют больше вычислений. Оба метода полезны в зависимости от задачи.

4. Общий вывод:
   Парные тесты и дисперсионный анализ показали себя как надёжные методы для проверки различий между группами. Их результаты совпадают, и выбор метода зависит от целей анализа: ANOVA — для общей оценки, парные тесты — для уточнения.

Приложение

Матстат4.ipynb