МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)

Лабораторная работа 4
по дисциплине
«Статистика и анализ данных»

Выполнили:
Трифонов Василий Максимович
гр. J3111
ИСУ 467758,
Соловьев Матвей Михайлович
гр. J3111
ИСУ 467551,
Ежов Дмитрий Александрович
гр. J3111
ИСУ 471242,

Отчет сдан: 17.06.2025

Санкт-Петербург 2025

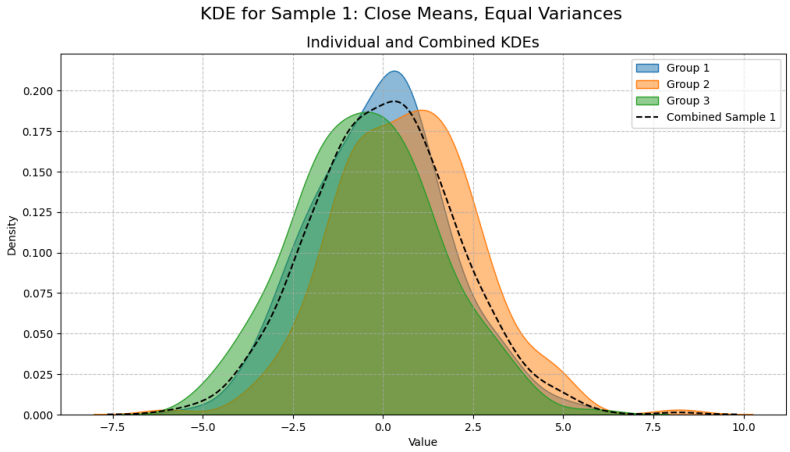
Ход выполнения работы

Данная лабораторная работа посвящена изучению и применению методов дисперсионного анализа и парных тестов для сравнения средних значений групп. Основные этапы работы включали генерацию данных, их визуализацию, реализацию и применение парных тестов (Z-тест и t-тест), а также однофакторного дисперсионного анализа (ANOVA), включая самостоятельную реализацию расчета F-статистики.

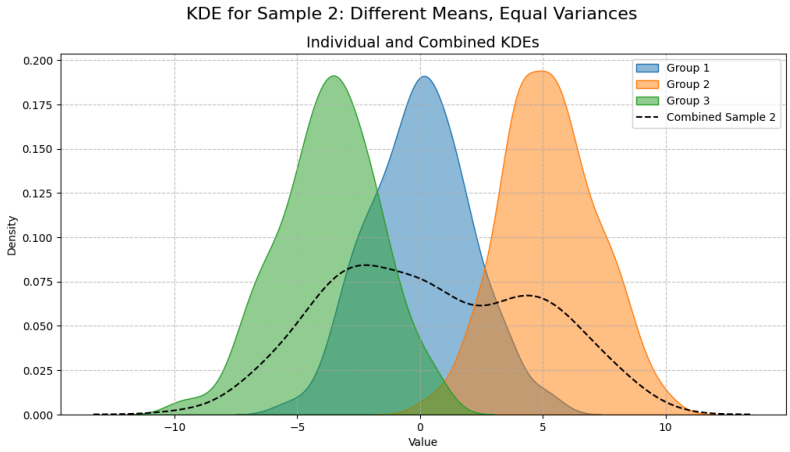
Основная часть: Описание шагов и выводы

Генерация и визуализация данных

- Сгенерированы две основные выборки, каждая из которых состоит из трех подвыборок (групп) из **нормального распределения**.
  - **Выборка 1:** Группы с одинаковыми дисперсиями и близкими математическими ожиданиями.



  - **Выборка 2:** Группы с одинаковыми дисперсиями и заметно отличающимися математическими ожиданиями.



- Для каждой выборки построены графики **оценки плотности ядра** (KDE) для каждой группы отдельно и для объединения групп.

Визуализация с помощью KDE наглядно продемонстрировала различия: для **Выборки 1** кривые сильно перекрывались, подтверждая схожесть распределений, а для **Выборки 2** кривые были четко разделены, что указывало на существенные различия в средних.

Парные тесты

Реализованы и применены два типа парных тестов для проверки гипотезы о равенстве математических ожиданий между всеми парами групп для обеих выборок.

- **Z-тест:** Применяется при **известных дисперсиях** генеральных совокупностей.
  Z-статистика:

$$Z = \frac{(|\{x\}_1 - |\{x\}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- **t-тест Стьюдента (с объединенной дисперсией):** Применяется при **неизвестных, но предполагаемо равных дисперсиях** генеральных совокупностей.

Объединенная дисперсия (для t-теста):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

t-статистика:

$$t = \frac{|\{x\}_1 - |\{x\}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- **P-значение** рассчитывается исходя из Z- или t-распределения, соответствующего статистике.

Выводы:

- Для **Выборки 1** (близкие средние) оба теста, как правило, не отвергали нулевую гипотезу(о равенстве математических ожиданий), что согласуется с данными.
- Для **Выборки 2** (различные средние) оба теста отвергали нулевую гипотезу , подтверждая статистически значимые различия между группами.
- **P-значение** является вероятностью получить наблюдаемый результат, если нулевая гипотеза верна. Чем оно меньше, тем сильнее доказательства против H_0.

Дисперсионный анализ (ANOVA)

Использован **однофакторный дисперсионный анализ** (ANOVA) и **F-тест** для проверки гипотезы о равенстве всех средних групп в рамках выборок.

- **Общая сумма квадратов (Total Sum of Squares, SS_{total}):** $SS_{\text{total}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - |(|(x))\right)^2$
- **Межгрупповая сумма квадратов (Between-Group Sum of Squares, SS_{between}):** $SS_{\text{between}} = \sum_{i=1}^{k} n_i (|(x)_i - |(|(x)))^2$
- **Внутригрупповая сумма квадратов (Within-Group Sum of Squares, SS_{within}):** $SS_{\text{within}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - |(x)_i\right)^2$
- **Средние квадраты:** $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$, $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$
- **F-статистика:** $F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$
- **P-значение для ANOVA:** рассчитывается на основе **F-распределения** с df_{between} и df_{within} степенями свободы: $P\left(F_{\text{распределение}}(df_{\text{between}}, df_{\text{within}}) \geq F_{\text{наблюдаемая}}\right)$.

Выводы:

- Для **Выборки 1** ANOVA не отвергла H_0 (высокое **P-значение**), указывая на отсутствие общих значимых различий.

- Для **Выборки 2** ANOVA отвергла H_0 (низкое **P-значение**), подтверждая наличие как минимум одного значимого различия между группами. Ручной расчет **F-статистики** совпал с библиотечной функцией.

Анализ методов

- **Вычислительная эффективность:** ANOVA значительно более эффективна для сравнения трех и более групп, требуя всего одного расчета **F-статистики** вместо множества парных тестов.
- **Полнота информации:**
  - Парные тесты (с корректировкой на множественные сравнения) дают детализированную информацию о том, **какие именно** пары групп различаются.
  - ANOVA быстро определяет **наличие** общего различия, но не указывает конкретные пары. Для этого требуются дополнительные **пост-хок тесты**.

Заключение

Проведенный анализ показал, что ANOVA является более эффективным методом для выявления общего различия между тремя и более группами, тогда как парные тесты дают более детальную информацию о конкретных различиях между парами. Выбор метода зависит от цели исследования и необходимости учета проблемы множественных сравнений. Полученные результаты подтвердили теоретические аспекты дисперсионного анализа и его практическую применимость.