

Лабораторная работа №1 по математической статистике

Тема: Анализ и моделирование случайных величин на основе эмпирических данных

Цель работы

- Изучить методы анализа и визуализации эмпирических данных.
- Освоить построение гистограмм и различных теоретических распределений.
- Научиться оценивать параметры распределений (математическое ожидание и дисперсию) на основе полученных данных.
- Развить навыки программирования на языке Python с использованием Jupyter Notebook.
- Реализовать методы статистического анализа без использования специализированных библиотек.

Задачи

1. Импорт данных:

- Загрузить три предоставленных файла данных в Jupyter Notebook.
- Проверить корректность загрузки и структуру данных.

2. Исследовательский анализ данных:

- Рассчитать математическое ожидание и дисперсию для каждого набора данных.
- Построить гистограммы с нормировкой 100% для визуализации распределения данных.

3. Построение теоретических распределений:

- На основе рассчитанных параметров создать функции распределения:
 - Геометрическое распределение
 - Биномиальное распределение
 - Распределение Пуассона
 - Равномерное распределение
 - Нормальное распределение
 - Экспоненциальное распределение
- Реализовать алгоритмы построения этих распределений вручную, без использования соответствующих функций из библиотек (разрешено использовать библиотеку numpy)

4. Визуальное сравнение:

- Наложить теоретические распределения на соответствующие гистограммы.

- Проанализировать соответствие эмпирических данных и теоретических моделей.

5. Документация и выводы:

- Подробно описать каждый этап работы.
- Сделать выводы о том, какое теоретическое распределение наиболее подходит для каждого набора данных.

Описание данных

В предоставленных трех файлах содержатся числовые данные, сгенерированные с помощью генератора случайных величин по одному из следующих законов распределения:

- Геометрическое распределение
- Биномиальное распределение
- Распределение Пуассона
- Равномерное распределение
- Нормальное распределение
- Экспоненциальное распределение

Каждый файл содержит последовательность чисел, представляющих измерения или наблюдения, полученные из соответствующего распределения.

Требования к выполнению работы

• Оборудование и ПО:

- Устанавливаем Python и Jupyter Notebook.

• Язык программирования:

- Python

• Инструменты:

- Для построения графиков использовать matplotlib.
- Теоретические распределения можно генерировать с помощью стандартной библиотеки Python, но необходимо реализовать методы самостоятельно (без использования функций библиотеки для расчета распределений).

• Структура проекта:

- Один Jupyter Notebook с комментариями и описаниями выполняемых шагов.
- Чистый и хорошо структурированный код, который легко читается и понимается.

• Обязательные элементы:

- Построение гистограмм 100% для каждого набора данных.
- Расчет математического ожидания и дисперсии для каждого набора данных.
- Построение теоретических распределений на основе полученных

- параметров.
- Сопоставление эмпирических и теоретических распределений.
- Описание выполненной работы и выводы.

Пошаговая инструкция

1. Подготовка окружения:

- Убедитесь, что Python и Jupyter Notebook установлены на вашем компьютере

2. Импорт данных:

- Все три файла содержат числовые данные, сгенерированные по разным распределениям.
- Реализуйте загрузку всех файлов в цикле, чтобы автоматизировать процесс обработки.
- Добавьте обработку потенциальных ошибок, таких как отсутствие файла или некорректный формат данных.

3. Анализ данных:

- Рассчитайте математическое ожидание и дисперсию для каждого набора данных.
- Реализуйте функции для расчета этих статистик.

4. Обнаружение и удаления выбросов

В процессе обработки и визуализации данных часто возникают ситуации, когда некоторые значения (выбросы) существенно отличаются от остальных и негативно влияют на качество графиков. Такие выбросы могут растягивать оси графиков, скрывать общие тенденции и делать визуализацию менее информативной. Существует несколько математических методов для обнаружения и удаления или обработки таких значений. Сравните приведенные методы между собой и выберите один, чтобы улучшить данные

- Метод Z-оценки
- Метод IQR (Interquartile Range)
- Тримминг (Обрезка по процентилям)
- Винзоризация (Winsorization)
- Использование медианы и медианного абсолютного отклонения (MAD)

5. Построение гистограмм:

- Используйте matplotlib для построения гистограмм с нормировкой 100%.
- Автоматизируйте построение гистограмм для всех наборов данных.

6. Реализация теоретических распределений:

- Реализуйте функции для каждого распределения, используя математические формулы.
- Автоматизируйте процесс построения распределений для всех наборов данных.
- Обработайте случаи, когда параметры распределения недоступны.
- Не забывайте, что некоторые распределения могут быть недоступны для

некоторых наборов данных.

- Всего максимум 6 теоретических распределений для каждого набора данных.

7. Сопоставление и анализ:

- Сравните построенные теоретические распределения с эмпирическими гистограммами.
- Определите, какое распределение визуально наиболее точно описывает каждый набор данных.

8. Документация и выводы:

- Оформите все шаги работы в виде понятно написанного отчета внутри Jupyter Notebook.
- Сделайте выводы о соответствии данных теоретическим моделям и обоснование выбора наиболее подходящего распределения.

Рекомендации

• Чистота кода:

- Используйте функции для повторяющихся операций.
- Следуйте стилю кодирования PEP8.
- Добавляйте комментарии для пояснения логики кода.

• Визуализация:

- Подбирайте понятные и контрастные цвета для графиков.
- Добавляйте легенды, подписи и заголовки для улучшения читаемости графиков.
- На графиках не должно быть необъяснимых пробелов и пустых столбцов

• Творческий подход:

- Экспериментируйте с различными типами графиков и их настройками.
- Добавляйте дополнительные анализы, если это обосновано данными.

Критерии оценки лабораторной работы

Лабораторная работа оценивается по следующим критериям:

1. Корректность выполнения задания (40%)

- Правильный импорт и обработка данных.
- Верное построение гистограмм с нормировкой 100%.
- Точный расчет математического ожидания и дисперсии.
- Корректная реализация теоретических распределений.
- Без построения достаточного числа гистограмм для каждого набора данных лабораторная не будет принята.

2. Качество визуализации (20%)

- Четкость и информативность графиков.
- Графики не содержат выбросы, осуществлена очистка от выбросов
- Соответствие требований к оформлению.

- Наложение теоретических распределений на гистограммы.

3. Структура и чистота кода (20%)

- Использование функций и модульность кода.
- Соблюдение стиля кодирования (PEP8).
- Отсутствие избыточного или неиспользуемого кода, в том числе и закомментированного.

4. Документация и оформление (10%)

- Наличие подробных комментариев и описаний.
- Логичная структура Jupyter Notebook.
- Четкое представление выводов и интерпретаций.

5. Творческий подход и самостоятельность (10%)

- Инновационные решения в визуализации.
- Самостоятельная реализация алгоритмов без использования специализированных библиотек.
- Внесение собственных дополнений и улучшений в работу.

Оценка может быть снята в случае:

- Отсутствия требуемой гистограммы (немедленный ноль).
- Использования готовых функций для построения распределений без самостоятельной реализации методов.
- Некачественного или непонятного кода, затрудняющего проверку работы.
- Недостаточного описания выполненных шагов и выводов.
- Отсутствия обработки ошибок, приводящей к сбоям при запуске кода.

Полезные ресурсы

- [Официальная документация Python \(https://www.python.org/doc/\)](https://www.python.org/doc/)
- [Руководство по Jupyter Notebook \(https://jupyter.org/documentation\)](https://jupyter.org/documentation)
- [Уроки по Matplotlib \(https://matplotlib.org/stable/users/index.html\)](https://matplotlib.org/stable/users/index.html)
- [Курс "Matplotlib с Python" на Udemy \(https://www.udemy.com/course/matplotlib-with-python/\)](https://www.udemy.com/course/matplotlib-with-python/)
- [PEP 8 – Стиль кодирования для Python \(https://pep8.org/\)](https://pep8.org/)

Удачи в выполнении лабораторной работы!