Лабораторная работа № 10

Статистика. Проверка статистических гипотез

Указания к выполнению лабораторной работы

В лабораторной работе требуется написать собственную программу на языке программирования Python с использованием стандартных функций из библиотек **pandas, Matplotlib** и др.

В качестве отчета по работе преподавателю предъявляются решения в электронном виде (файлы .py или .ipynb). При необходимости нужно ответить на дополнительные вопросы.

Задание на лабораторную работу

Задание 1. *Загрузка данных и предварительная обработка*

1. Загрузить данные из файла *avocado.csv*. Вывести начало таблицы, определить, какие поля и какого типа присутствуют.

Подсказка: датасет сформирован на основе готового набора с ресурса https://www.kaggle.com/datasets/neuromusic/avocado-prices?resource=download

2. В качестве исследуемого параметра возьмём среднюю цену на авокадо (AveragePrice). Требуется провести предварительный анализ:

- проверить данные на наличие значений NaN, пустых полей;
- с помощью стандартного инструмента библиотеки pandas describe() найти минимальное, максимальное значения, оценить их адекватность.
- Записи со значением NaN и аномальные выбросы удалить.

3. Для средней цены составить интервальный вариационный ряд. Количество интервалов определить самостоятельно.

4. Для полученного вариационного ряда найти среднее значение, медиану и моду.

Задание 2. *Визуализация данных*

1. Построить гистограмму и кумуляту. Какой из способов является более наглядным?

2. По гистограмме попробуйте выдвинуть гипотезу о законе распределения параметра.

Задание 3. *Проверка гипотезы о законе распределения*

1. Выдвинуть гипотезу H_0 относительно закона распределения: «Выборка извлечена из генеральной совокупности со стандартным нормальным распределением».

2. Построить эмпирическую и теоретическую оценки плотности распределения.

3. Зафиксировать уровень значимости критерия $\alpha = 0.05$.

4. Применить критерий хи-квадрат, используя встроенные функции в Python, и получить p-value.

5. Сравнить p-value с выбранным уровнем значимости α, сделать выводы.

Задание 4. *Проверка статистической значимости*

1. Сформируйте два датасета, в которые входят количественные оценки качества двух методов:

первый: содержит оценки качества для двух статистически неотличимых методов (то есть установите для обоих методов одинаковое среднее значение и сгенерируйте N оценок качества с заданным стандартным отклонением (отклонения можно установить разными для методов)),

второй: то же самое, но средние значения для методов должны отличаться.

2. Посчитайте статистическую значимость для уровней $\alpha = 0.05, 0.01, 0.001$ для обоих датасетов по двухстороннему t-критерию Стьюдента. Выведите на графике распределения Стьюдента критические значения.

3. Посчитайте доверительные интервалы для обоих датасетов на уровне доверия 0.95.