MIT World Peace University

Data Science for Cybersecurity and Forensics
Third Year B. Tech, Semester 6

Statistical Approaches in Data Science

Assignment 3

Prepared By

Krishnaraj Thadesar
Cyber Security and Forensics
Batch A1, PA 10

April 16, 2024

Contents

1 Aim

Learning some statistical approaches that are used in data science.

2 Objectives

1. Write a Python program to implement central tendency for housing data.
2. Using python compute variance in the weather.
3. Compute variance in the weather to find best time to visit New Delhi(or any city).
4. Using histogram find the best time to visit Delhi (or any)s on any dataset.

3 Theory

3.1 Types of Statistics

Statistics can be broadly categorized into two main types: descriptive statistics and inferential statistics.

3.2 Descriptive Statistics

Descriptive statistics involve methods for summarizing and describing the features of a dataset. It provides insights into the central tendency, variability, and distribution of the data.

3.2.1 Measures of Central Tendency

Measures of central tendency are statistics that describe the center or average of a dataset. Common measures of central tendency include the mean, median, and mode.

- **Mean**: The arithmetic average of a set of values, calculated by summing all the values and dividing by the number of observations.

  Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median**: The middle value in a dataset when the values are arranged in ascending order. It divides the dataset into two equal halves.

  Formula:

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2}+x_{n/2+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

- **Mode**: The value that appears most frequently in a dataset.

  Formula:

$$\text{Mode} = \text{value with highest frequency}$$

3.2.2 Measures of Dispersion

Measures of dispersion quantify the spread or variability of the data points in a dataset. They provide information about how the data is distributed around the central tendency.

- **Range**: The difference between the maximum and minimum values in a dataset.

  Formula:
  $$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

- **Variance**: The average of the squared differences from the mean. It measures the average distance of each data point from the mean.

  Formula:
  $$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

- **Standard Deviation**: The square root of the variance. It provides a measure of the dispersion of data points around the mean.

  Formula:
  $$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

3.3 Inferential Statistics

Inferential statistics involve methods for making predictions or inferences about a population based on a sample of data. It uses probability theory to draw conclusions about the population parameters.

3.3.1 Hypothesis Testing

Hypothesis testing is a statistical method used to determine whether there is enough evidence to reject a null hypothesis in favor of an alternative hypothesis. It involves setting up a null hypothesis and an alternative hypothesis, collecting data, and using statistical tests to make a decision.

3.3.2 Regression Analysis

Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It helps in understanding how the value of the dependent variable changes when one or more independent variables are varied.

3.3.3 Correlation Analysis

Correlation analysis is a statistical method used to measure the strength and direction of the relationship between two variables. It helps in understanding how changes in one variable are associated with changes in another variable.

4 Platform

Operating System: Windows 11
IDEs or Text Editors Used: Visual Studio Code
Compilers or Interpreters: Python 3.10.1

5 Requirements

```
python==3.10.1
matplotlib==3.8.3
numpy==1.26.4
pandas==2.2.2
seaborn==0.13.2
```

6 Code

7 FAQs

7.1 Question 1

1. **What do you understand by Statistics for Data science?**
   Statistics for data science involves the application of statistical methods and techniques to analyze, interpret, and derive insights from data. It encompasses a wide range of methods, including descriptive statistics, inferential statistics, and predictive modeling, to explore patterns, relationships, and trends within datasets. In data science, statistics plays a crucial role in data preprocessing, exploratory data analysis, hypothesis testing, and model evaluation, enabling data scientists to make informed decisions and derive actionable insights from data.

7.2 Question 2

1. **Do we need preprocessing to perform statistics for Data science? Justify, your answer**
   Yes, preprocessing is essential for performing statistics in data science. Preprocessing involves cleaning, transforming, and preparing raw data to make it suitable for statistical analysis. Without preprocessing, raw data may contain missing values, outliers, inconsistencies, or other irregularities that can affect the accuracy and reliability of statistical results. Preprocessing techniques such as handling missing values, outlier detection, data normalization, and feature engineering help ensure that the data meets the assumptions and requirements of statistical methods. By preprocessing the data, data scientists can improve the quality of statistical analysis, enhance the performance of models, and derive more accurate and meaningful insights from the data.

7.3 Question 3

1. **Describe the different Statistical approaches in Data science using Python?**
   In data science, various statistical approaches are used to analyze and model data using Python. Some common statistical approaches include:

   - Descriptive Statistics: Summarizing and describing the features of a dataset using measures of central tendency, dispersion, and visualization techniques.
   - Inferential Statistics: Making inferences and predictions about populations based on sample data using hypothesis testing, confidence intervals, and regression analysis.
   - Predictive Modeling: Building predictive models to forecast future outcomes or classify data into different categories using techniques such as linear regression, logistic regression, decision trees, and ensemble methods.
   - Time Series Analysis: Analyzing and forecasting time-series data using methods like autoregressive integrated moving average (ARIMA), seasonal decomposition, and exponential smoothing.

   Python provides a wide range of libraries and tools for implementing these statistical approaches, including NumPy, pandas, SciPy, scikit-learn, and Statsmodels, making it a popular choice for statistical analysis in data science.

8 Conclusion

In this assignment, we learned about various statistical approaches used in data science, including measures of central tendency, dispersion, hypothesis testing, regression analysis, and correlation analysis. We implemented these statistical concepts using Python and explored how they can be applied to analyze and interpret data. By understanding and applying statistical methods, data scientists can gain valuable insights from data, make informed decisions, and build predictive models to solve real-world problems.