Statistics for Data Science

Mayank Pratap Singh

September 19, 2024

1 Introduction to Statistics

Statistics is a vital tool in *Data Analysis*, allowing us to **summarize data**, make sense of large datasets, and arrive at informed decisions.

2 Motivation for Statistics

The need for statistics arises due to the impracticality of working with the entire population. We work with **samples**, which are subsets of a **population**, to infer trends and patterns.

2.1 Population and Sample

- **Population:** The complete dataset, typically large, e.g., population of a city (10,000 individuals).
- **Sample:** A subset of the population (e.g., a sample of 500 people) that is used to estimate the characteristics of the population.
- **Rule of Thumb:** For most statistical analyses, we prefer **n ¿ 30** in sample size.

3 Statistical Modeling

Statistical models help identify **patterns** and **distributions** in data. These can be continuous or discrete, and different distributions include:

- **Binomial Distribution**
- **Normal Distribution**
- **Poisson Distribution**

Each distribution is described by either a **Probability Density Function (PDF)** for continuous random variables or a **Probability Mass Function (PMF)** for discrete random variables.

4 Types of Statistics

4.1 Descriptive Statistics

Descriptive statistics allow us to **summarize and describe** the data. Key concepts include:

- Measures of Central Tendency: **Mean, Median, Mode**
- Measures of Dispersion: **Standard Deviation, Variance, Range**
- Visualization: **Histograms, Box Plots, Scatter Plots**
- Data Distribution

4.2 Inferential Statistics

Inferential statistics allow us to make inferences about the population based on sample data. Key techniques include:

- **Hypothesis Testing**
- **Confidence Intervals**
- **Regression Analysis**
- **P-value and Critical Value**
- **Statistical Tests:** Z-test, t-test, ANOVA, Chi-Square Test

5 Real World Applications of Statistics

Statistics are used in various fields, including:

- **Financial Predictions:** Stock market trends
- **Healthcare:** Medical research and clinical trial data (e.g., drug recovery times)
- **Sports:** Performance evaluation and improvement
- **Business & Marketing:** Customer retention strategies

6 Key Python Libraries for Statistics

- **SciPy:** For scientific and mathematical computations
- **Scikit-learn:** Includes statistical functions and data normalization
- **StatsModels:** Excellent for hypothesis testing

7 Types of Data

7.1 Qualitative Data

- **Nominal Data:** Categories without order (e.g., gender, colors)
- **Ordinal Data:** Categories with meaningful order (e.g., education levels, satisfaction ratings)

7.2 Quantitative Data

- **Interval Data:** Ordered data with meaningful intervals but no true zero (e.g., temperature)
- **Ratio Data:** Ordered data with a true zero (e.g., income)

7.3 Discrete vs. Continuous Data

- **Discrete Data:** Countable, separate values (e.g., number of cars)
- **Continuous Data:** Measurable values within a range (e.g., height, weight)

8 Measures of Central Tendency

For discrete data, **mean** and **mode** are meaningful measures. For continuous data, **mean** and **median** are more relevant.

- **Mean:** Average of all data points.
- **Median:** Middle value after sorting.
- **Mode:** Most frequent value.

9 Key Takeaways

- Statistics simplify complex data.
- Models help us understand data patterns.
- Statistical methods lead to informed decision-making.