Week 1: Introduction to Data Science and Descriptive Statistics

1. Introduction to Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from noisy, structured, and unstructured data [1]. It combines various fields, including statistics, computer science, and business knowledge, to understand and analyze actual phenomena with data. The ultimate goal of data science is to make better decisions and predictions.

2. Understanding Data

Data refers to raw, unorganized facts, figures, objects, symbols, and events that are collected from various sources. In its raw form, data holds little meaning. However, once it is processed, organized, and analyzed, it can provide valuable insights and information. Data is the foundation of any data science project.

Types of Data

Data can be broadly categorized into two main types: **Qualitative** and **Quantitative** data.

2.1. Qualitative Data

Qualitative data, also known as categorical data, describes qualities or characteristics that cannot be measured numerically. It is often descriptive and can be observed but not calculated. Examples include colors, names, and types of animals.

- **Nominal Data**: Data that can be labeled or classified into mutually exclusive categories without any order or ranking. Examples: Gender (Male, Female), Marital Status (Single, Married, Divorced).

- **Ordinal Data**: Data that can be categorized and ranked in a meaningful order, but the differences between categories are not precisely measurable. Examples: Education Level (High School, Bachelor's, Master's, PhD), Customer Satisfaction (Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied).

2.2. Quantitative Data

Quantitative data, also known as numerical data, consists of numerical values that can be measured or counted. This type of data can be subjected to mathematical operations and statistical analysis. Examples include height, weight, age, and temperature.

- **Discrete Data**: Data that can only take specific, distinct values and cannot be broken down into smaller units. It is often the result of counting. Examples: Number of children in a family, number of cars in a parking lot.
- **Continuous Data**: Data that can take any value within a given range. It is typically the result of measuring. Examples: Height of a person, temperature of a room, time taken to complete a task.

3. Descriptive Statistics

Descriptive statistics are used to summarize and describe the main features of a collection of information quantitatively. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

3.1. Mean (Average)

The **mean**, or arithmetic average, is the most commonly used measure of central tendency. It is calculated by summing all the values in a dataset and then dividing by the number of values.

Formula:

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

Where: * $\bar{x}$ is the mean * $\sum x_i$ is the sum of all values * n is the number of values in the dataset

Example: For the dataset `[10, 20, 30, 40, 50]`, the mean is $(10 + 20 + 30 + 40 + 50)/5 = 150/5 = 30$.

3.2. Median

The **median** is the middle value in a dataset when the values are arranged in ascending or descending order. It is particularly useful when the dataset contains outliers, as it is less affected by extreme values compared to the mean.

- If the number of values is odd, the median is the middle value.
- If the number of values is even, the median is the average of the two middle values.

Example 1 (Odd number of values): For the dataset `[10, 20, 30, 40, 50]`, the sorted dataset is `[10, 20, 30, 40, 50]`. The median is `30`.

Example 2 (Even number of values): For the dataset `[10, 20, 30, 40]`, the sorted dataset is `[10, 20, 30, 40]`. The median is $(20 + 30)/2 = 25$.

3.3. Standard Deviation

The **standard deviation** measures the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Formula (Population Standard Deviation):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Where: * σ is the population standard deviation * x_i is each individual value * μ is the population mean * N is the number of values in the population

Formula (Sample Standard Deviation):

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Where: * s is the sample standard deviation * x_i is each individual value * $\bar{x}$ is the sample mean * n is the number of values in the sample

Example: For the dataset `[10, 20, 30, 40, 50]`, with a mean of `30`:

1. Calculate the difference from the mean for each value: `[-20, -10, 0, 10, 20]`
2. Square each difference: `[400, 100, 0, 100, 400]`
3. Sum the squared differences: `400 + 100 + 0 + 100 + 400 = 1000`
4. Divide by `n` (for population) or `n-1` (for sample). Let's assume population, so `1000 / 5 = 200` (variance).
5. Take the square root: $\sqrt{200} \approx 14.14$.

This means, on average, the values in the dataset deviate by approximately 14.14 units from the mean of 30.

References

[1] IBM. (n.d.). *What is data science?* Retrieved from https://www.ibm.com/topics/data-science