Descriptive Statistics for Data Science

Data Science Student

September 19, 2024

Contents

1 Descriptive Statistics **1**
1.1 Population vs. Sample 1
1.2 Types of Data 1

2 Measure of Central Tendency **2**
2.1 Mean 2
2.2 Median 2
2.3 Mode 2

3 Measure of Dispersion **2**
3.1 Range 2
3.2 Interquartile Range (IQR) 3
3.3 Variance and Standard Deviation 3

4 Normal Distribution and the Empirical Rule **3**

5 Key Takeaways **3**

1 Descriptive Statistics

Descriptive statistics provide methods for summarizing the data, allowing us to get a quick overview of the dataset.

1.1 Population vs. Sample

- **Population:** The complete dataset.
- **Sample:** A subset of the population, used to infer trends.

1.2 Types of Data

- **Numerical Data:** Data that represents measurable quantities (e.g., height, income).
- **Categorical Data:** Data that represents categories (e.g., gender, colors).

2 Measure of Central Tendency

Measures of central tendency give us an idea of where the center of the data lies.

2.1 Mean

The mean, or average, is given by:

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: Given the dataset {4, 8, 6, 5, 3, 8, 9, 8, 2}, the mean is:

$$\text{Mean} = \frac{4+8+6+5+3+8+9+8+2}{9} = 5.89$$

2.2 Median

The median is the middle value of an ordered dataset.

- **Odd dataset size:** The middle value.
- **Even dataset size:** The average of the two middle values.

Example: After sorting the dataset {4, 8, 6, 5, 3, 8, 9, 8, 2}:

$$\text{Ordered Set} = \{2, 3, 4, 5, 6, 8, 8, 8, 9\}$$

The median is 6.

2.3 Mode

The mode is the most frequently occurring value in a dataset.

Example: In the dataset {4, 8, 6, 5, 3, 8, 9, 8, 2}, the mode is 8 (appears 3 times).

3 Measure of Dispersion

Dispersion gives us an idea of how spread out the data is.

3.1 Range

The range is the difference between the maximum and minimum values in a dataset:

$$\text{Range} = \max - \min$$

Example: In the dataset {56, 64, 75, 80, 83, 90, 92, 95, 98, 100}, the range is:

$$100 - 56 = 44$$

While range is a quick measure of variability, it is sensitive to outliers.

3.2 Interquartile Range (IQR)

The interquartile range (IQR) is the range between the 25th percentile (Q1) and the 75th percentile (Q3), giving a robust measure of variability that is less influenced by outliers:

$$\text{IQR} = Q_3 - Q_1$$

Example: In the dataset {56, 64, 75, 80, 83, 90, 92, 95, 98, 100}, we have:

$$Q_1 = 75,\ Q_3 = 95 \quad \Rightarrow \text{IQR} = 95 - 75 = 20$$

3.3 Variance and Standard Deviation

Variance measures the overall spread of data, while standard deviation is the square root of the variance, providing the average deviation from the mean.

Variance:

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

For the sample variance, we use $n-1$ instead of n for an unbiased estimate:

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

Standard Deviation:

$$s = \sqrt{s^2}$$

Standard deviation is widely used in data science as it provides a direct measure of variability in the same units as the data itself.

4 Normal Distribution and the Empirical Rule

The normal distribution is a common probability distribution. According to the empirical rule:

- 68% of the data lies within one standard deviation of the mean.
- 95% of the data lies within two standard deviations of the mean.
- 99.7% of the data lies within three standard deviations of the mean.

5 Key Takeaways

- Descriptive statistics help simplify complex data and provide insights into data trends.
- Measures of central tendency (mean, median, mode) are essential for summarizing data.
- Measures of dispersion (range, IQR, variance, standard deviation) give insights into data spread and variability.
- The normal distribution is foundational in many statistical methods, with the empirical rule offering a simple understanding of data spread.

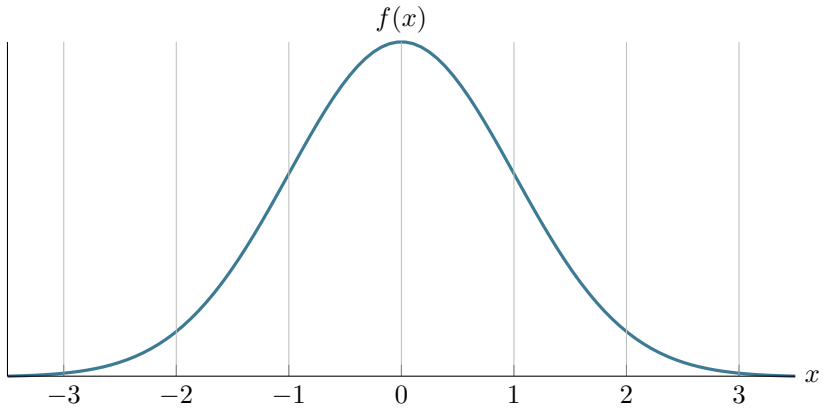

Figure 1: Normal Distribution: 68% of the data within one standard deviation.