Basic Descriptive Statistics (Notes)

1. Data Basics

Understanding data is the cornerstone of statistics. Data is the raw material that we analyze to extract insights, make decisions, and build models. To truly master this concept, we need to explore it in depth.

1.1 Types of Data

Data can be classified into two main categories: Categorical and Numerical . Each category has subtypes, and understanding these distinctions is critical for choosing the right statistical methods and visualizations.

1.1.1 Categorical Data

Categorical data represents categories or labels . These categories do not involve numerical values but describe **qualitative attributes.**

- Nominal Data :
    - Categories have no inherent order.
    - Examples:
        - Colors: Red, Blue, Green
        - Marital Status: Single, Married, Divorced
        - Cities: New York, London, Tokyo
    - Key Characteristics:
        - Cannot perform mathematical operations (e.g., "Red + Blue" makes no sense).
        - Often summarized using frequency counts or percentages.
- Ordinal Data :
    - Categories have a meaningful order or ranking.
    - Examples:
        - Education Levels: High School, Bachelor's, Master's, Ph.D.
        - Movie Ratings: Poor, Average, Good, Excellent
        - Survey Responses: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
    - Key Characteristics:
        - Order matters, but the differences between categories are not necessarily equal.
            - For example, the difference between "High School" and "Bachelor's" is not the same as the difference between "Master's" and "Ph.D."
        - Can summarize using medians or modes, but not means.

1.1.2 Numerical Data

Numerical data represents quantitative measurements and involves numbers. It can be further divided into discrete and continuous data.

- Discrete Data :
    - Represents countable values.
    - Examples:
        - Number of students in a classroom: 25, 30, 40
        - Number of cars in a parking lot: 10, 15, 20
        - Number of goals scored in a soccer match: 0, 1, 2, 3
    - Key Characteristics:
        - Values are whole numbers (integers).
        - Finite or countably infinite set of possible values.
- Continuous Data :
    - Represents measurable values that can take any value within a range.
    - Examples:
        - Height: 1.75 m, 1.80 m, 1.85 m
        - Weight: 68.5 kg, 70.2 kg, 72.1 kg
        - Temperature: 23.5°C, 24.1°C, 25.0°C
    - Key Characteristics:
        - Values can include decimals or fractions.
        - Infinite number of possible values within a range.

1.2 Variables

A variable is a characteristic or attribute that can take different values. Variables are used to represent data and are central to statistical analysis.

Types of Variables

1. Independent Variable :
    - The variable that is manipulated or controlled in an experiment.
    - Example:
        - In a study examining the effect of hours studied on exam scores:
            - Independent Variable: Hours studied.
            - Dependent Variable: Exam scores.
2. Dependent Variable :
    - The variable that is being measured or observed.
    - Example:
        - In the same study:
            - Dependent Variable: Exam scores.
3. Confounding Variable :

- A variable that affects both the independent and dependent variables, potentially causing misleading results.
- Example:
    - In the same study:
        - Confounding Variable: Prior knowledge of the subject (students who already know the material may score higher regardless of hours studied).

1.3 Populations vs. Samples

Understanding the distinction between populations and samples is crucial for statistical inference.

- Population :
    - The entire group of individuals or items that you're interested in studying.
    - Example:
        - All students in a university.
        - All adults in a country.
- Sample :
    - A subset of the population that is selected for analysis.
    - Example:
        - 100 students randomly chosen from the university.
        - 1,000 adults surveyed from the country.

Why Use Samples?

- Studying an entire population is often impractical due to time, cost, or logistical constraints.
- Statistical techniques allow us to generalize findings from a sample to the population.

Examples to Understand Data Basics

Example 1: Classifying Data Types

Classify the following data as categorical (nominal/ordinal) or numerical (discrete/continuous):

1. Favorite ice cream flavor (chocolate, vanilla, strawberry).
2. Number of books read in a year.
3. Temperature in degrees Celsius.
4. Education level (high school, bachelor's, master's, Ph.D.).

Solution :

1. Categorical (Nominal) : Ice cream flavors have no inherent order.
2. Numerical (Discrete) : The number of books is countable.
3. Numerical (Continuous) : Temperature can take any value within a range.
4. Categorical (Ordinal) : Education levels have a meaningful order.

Example 2: Identifying Variables

In a study examining the effect of exercise on weight loss:

- What is the independent variable?
- What is the dependent variable?

Solution :

- Independent Variable : Amount of exercise (manipulated by the researcher).
- Dependent Variable : Weight loss (measured outcome).

Example 3: Population vs. Sample

A researcher wants to study the average income of all adults in a country but only surveys 1,000 adults.

- What is the population?
- What is the sample?

Solution :

- Population : All adults in the country.
- Sample : The 1,000 adults surveyed.

Exercises for Practice

Exercise 1: Classify the Data

Classify the following data as categorical (nominal/ordinal) or numerical (discrete/continuous):

1. Marital status (single, married, divorced).
2. Number of cars in a parking lot.
3. Time taken to complete a race.
4. Ratings of a movie (poor, average, good, excellent).

Answers :

1. Categorical (Nominal) : Marital status has no inherent order.
2. Numerical (Discrete) : The number of cars is countable.
3. Numerical (Continuous) : Time can take any value within a range.
4. Categorical (Ordinal) : Movie ratings have a meaningful order.

Exercise 2: Identify Variables

In a study examining the relationship between hours of sleep and test performance:

- What is the independent variable?
- What is the dependent variable?

Answers :

- Independent Variable : Hours of sleep (manipulated by the researcher).
- Dependent Variable : Test performance (measured outcome).

Exercise 3: Population vs. Sample

A researcher wants to study the average height of all adults in a country but only surveys 1,000 adults.

- What is the population?
- What is the sample?

Answers :

- Population : All adults in the country.
- Sample : The 1,000 adults surveyed.

Detailed Explanation of Key Concepts

Why Is Classifying Data Important?

Understanding the type of data helps you choose the right statistical methods and visualizations. For example:

- Categorical data is often summarized using frequency tables and bar charts.
- Numerical data can be analyzed using measures like mean, median, and standard deviation.

Why Distinguish Between Population and Sample?

In most cases, it's impractical to study an entire population (e.g., all adults in a country). Instead, we analyze a sample and use statistical techniques to generalize findings to the population.

2. Descriptive Statistics: A Beginner-Friendly Guide

Descriptive statistics is the foundation of understanding data. It helps us summarize and describe the main features of a dataset in a way that's easy to interpret. We'll cover:

1. Measures of Central Tendency : Mean, Median, Mode
2. Measures of Dispersion : Range, Variance, Standard Deviation, Interquartile Range (IQR)
3. Data Visualization : Histograms, Box Plots, Scatter Plots

Let's explore each topic in extreme detail, step by step.

2.1 Measures of Central Tendency

These measures tell us where the "center" or "typical value" of the data lies. They are essential because they give us a single number that represents the entire dataset. Think of them as a way to summarize the data without looking at every single number.

2.1.1 Mean (Average)

The mean is the most commonly used measure of central tendency. It's like finding the "balance point" of the data.

What Is the Mean?

- The mean is the sum of all values divided by the total number of values.
- Formula:

$$\text{Mean} = \frac{\text{Sum of All Values}}{\text{Number of Values}}$$

Mathematically:

$$\text{Mean} = \frac{\sum x_i}{n}$$

where:

- x_i = individual data points
- n = total number of data points

Step-by-Step Calculation :

- Add up all the numbers in the dataset.
- Divide the sum by the total number of numbers.

Why Use the Mean?

- The mean considers every data point, so it gives a good overall picture of the dataset.
- However, the mean is sensitive to outliers (extreme values). For example, if one person earns $1 million in a group of people earning $50,000, the mean income will be skewed upward.

Example :

Let's calculate the mean for the dataset: $[5, 10, 15, 20]$.

1. Add up all the numbers:

$$5 + 10 + 15 + 20 = 50$$

2. Divide by the total number of values ($n = 4$):

$$\text{Mean} = \frac{50}{4} = 12.5$$

Real-World Example :

Imagine you're a teacher calculating the average score of your students on a test. If the scores are $[70, 80, 90, 100]$, the mean score is:

$$\text{Mean} = \frac{70 + 80 + 90 + 100}{4} = \frac{340}{4} = 85$$

This tells you that, on average, students scored 85%.

Exercise :

Calculate the mean of the dataset: $[8, 12, 15, 20, 25]$.

- Solution:
  1. Add up all the numbers:

$$8 + 12 + 15 + 20 + 25 = 80$$

  2. Divide by the total number of values ($n = 5$):

$$\text{Mean} = \frac{80}{5} = 16$$

2.1.2 Median

The median is the middle value when the data is sorted in ascending order. It's like finding the "middle child" in a family.
What Is the Median?

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

Step-by-Step Calculation :

- Sort the data in ascending order.
- Find the middle value(s).

Why Use the Median?

- The median is robust to outliers. Even if there are extreme values, the median remains stable.
- Example: In income data, the median gives a better sense of what a "typical" person earns compared to the mean.

Example :

Let's calculate the median for the dataset: $[5,10,15,20]$.

1. Sort the data (already sorted): $[5,10,15,20]$.
2. Since there are 4 values (even number), take the average of the two middle values:

$$\text{Median} = \frac{10 + 15}{2} = 12.5$$

Real-World Example :

Imagine you're analyzing house prices in a neighborhood. The prices are $[200,000, 250,000, 300,000, 1,000,000]$.

The mean price is:

$$\text{Mean} = \frac{200,000 + 250,000 + 300,000 + 1,000,000}{4} = 437,500$$

But the median price is:

$$\text{Median} = \frac{250,000 + 300,000}{2} = 275,000$$

The median gives a better idea of what a "typical" house costs, ignoring the outlier mansion.

Exercise :

Find the median of the dataset: $[8, 12, 15, 20, 25]$.

- Solution:
  1. Sort the data (already sorted): $[8, 12, 15, 20, 25]$.
  2. Since there are 5 values (odd number), the middle value is:

$$\text{Median} = 15$$

2.1.3 Mode

The mode is the value that appears most frequently in the dataset. It's like finding the "most popular" item.

What Is the Mode? The mode is useful for identifying the most common value in a dataset.

A dataset can have:

- One mode (unimodal),
- More than one mode (bimodal, multimodal), or
- No mode if no value repeats.

Why Use the Mode?

- The mode is helpful for categorical data (e.g., favorite colors, shoe sizes).
- It's less commonly used for numerical data unless there's a clear peak.

Example :

Let's find the mode for the dataset: $[5,10,10,15,20]$.
Count the frequency of each value:

- 5: 1 time
- 10: 2 times
- 15: 1 time
- 20: 1 time

The mode is 10 (appears twice).

Real-World Example :

- Imagine you're a store manager analyzing shoe sizes sold in a week. The sizes are $[7,8,8,9,10,10,10]$. The mode is: $\text{Mode}=10$

This tells you that size 10 is the most popular shoe size.

Exercise : Identify the mode of the dataset: $[8,12,15,15,20,25]$.
Solution:
Count the frequency of each value:

- 8: 1 time
- 12: 1 time
- 15: 2 times
- 20: 1 time
- 25: 1 time

The mode is 15 (appears twice).

2.2 Measures of Dispersion

These measures tell us how spread out the data is. They are crucial because they give us a sense of variability in the dataset.

2.2.1 Range

The range is the simplest measure of dispersion. It's like measuring the "width" of the data.

What Is the Range?

- The range is the difference between the maximum and minimum values.
- Formula:
  - $\text{Range} = \text{Max} - \text{Min}$

Why Use the Range?

- The range gives a quick sense of the spread of the data.
- However, it's sensitive to outliers because it only considers the extremes.

Example :
Let's calculate the range for the dataset: $[5,10,15,20]$.

- Maximum: 20
- Minimum: 5
- Range:
  - $\text{Range} = 20 - 5 = 15$

Real-World Example :

Imagine you're analyzing daily temperatures over a week. The temperatures are $[60,65,70,75,80]$. The range is:

- $\text{Range} = 80 - 60 = 20$

This tells you the temperature fluctuated by 20 degrees during the week.

Exercise :

Calculate the range of the dataset: $[8,12,15,20,25]$.

Solution:

- Maximum: 25
- Minimum: 8
- Range: 25 - 8 = 17

2.2.2 Variance

The variance measures how far each number in the dataset is from the mean. It's like measuring the "average distance" of each data point from the center.

What Is Variance?

- Variance quantifies the spread of the data around the mean.
- Formula:

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n}$$

where:

- x_i = individual data points
- $\bar{x}$ = mean
- n = total number of data points

Step-by-Step Calculation :

1. Calculate the mean ($\bar{x}$).
2. Subtract the mean from each data point ($x_i - \bar{x}$).
3. Square each result ($(x_i - \bar{x})^2$).
4. Add up all the squared differences.
5. Divide by the total number of data points (n).

Why Use Variance?

- Variance tells us how much the data points vary from the mean.
- Larger variance indicates greater variability.

Example :
Let's calculate the variance for the dataset: [5,10,15,20].

1. Calculate the mean:

$$\text{Mean} = \frac{5 + 10 + 15 + 20}{4} = 12.5$$

2. Subtract the mean from each data point:

$$5 - 12.5 = -7.5, \quad 10 - 12.5 = -2.5, \quad 15 - 12.5 = 2.5, \quad 20 - 12.5 = 7.5$$

3. Square each result:

$$(-7.5)^2 = 56.25, \quad (-2.5)^2 = 6.25, \quad (2.5)^2 = 6.25, \quad (7.5)^2 = 56.25$$

4. Add up the squared differences:

$$56.25 + 6.25 + 6.25 + 56.25 = 125$$

5. Divide by the total number of data points ($n = 4$):

$$\text{Variance} = \frac{125}{4} = 31.25$$

Exercise : Compute the variance of the dataset: [8,12,15,20,25].

Solution:

1. Calculate the mean:

$$\text{Mean} = \frac{8 + 12 + 15 + 20 + 25}{5} = \frac{80}{5} = 16$$

2. Subtract the mean from each data point:

$$8 - 16 = -8, \quad 12 - 16 = -4, \quad 15 - 16 = -1, \quad 20 - 16 = 4, \quad 25 - 16 = 9$$

3. Square each result:

$$(-8)^2 = 64, \quad (-4)^2 = 16, \quad (-1)^2 = 1, \quad (4)^2 = 16, \quad (9)^2 = 81$$

4. Add up the squared differences:

$$64 + 16 + 1 + 16 + 81 = 178$$

5. Divide by the total number of data points ($n = 5$):

$$\text{Variance} = \frac{178}{5} = 35.6$$

Real-World Example : Imagine you're analyzing the daily sales of a store over five days: [8,12,15,20,25]. The variance of 35.6 tells you that the sales figures vary significantly around the mean of 16. This could indicate inconsistent customer traffic or fluctuating demand.

2.2.3 Standard Deviation

The standard deviation is simply the square root of the variance. It measures the spread of the data in the same units as the original data, making it easier to interpret than variance.

What Is Standard Deviation? Standard deviation quantifies how much individual data points deviate from the mean.

Formula:

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

Why Use Standard Deviation? Unlike variance, which is expressed in squared units, standard deviation is expressed in the original units of the data. It's widely used in fields like finance, quality control, and machine learning to measure variability.

Example :
Let's calculate the standard deviation for the dataset: $[5,10,15,20]$.

- From earlier, we calculated the variance as 31.25.
- Take the square root of the variance:

Example :

Let's calculate the standard deviation for the dataset: $[5,10,15,20]$.

1. From earlier, we calculated the variance as 31.25.
2. Take the square root of the variance:

$$\text{Standard Deviation} = \sqrt{31.25} \approx 5.59$$

Real-World Example :
Imagine you're analyzing test scores: $[70, 80, 90, 100]$. The variance was 125, so the standard deviation is:

$$\text{Standard Deviation} = \sqrt{125} \approx 11.18$$

This tells you that most students scored within about 11.18 points of the mean score of 85.
Exercise :
Find the standard deviation of the dataset: $[8, 12, 15, 20, 25]$.

- Solution:
  1. From earlier, we calculated the variance as 35.6.
  2. Take the square root of the variance:

$$\text{Standard Deviation} = \sqrt{35.6} \approx 5.97$$

2.2.4 Interquartile Range (IQR)

The interquartile range (IQR) is the range between the first quartile (Q1) and the third quartile (Q3). It represents the middle 50% of the data and is robust to outliers.

What Is the IQR? The IQR is the difference between Q3 (the 75th percentile) and Q1 (the 25th percentile).

Formula: $IQR = Q3 - Q1$

Step-by-Step Calculation :

- Sort the data in ascending order.
- Find Q1 (the median of the lower half) and Q3 (the median of the upper half).
- Subtract Q1 from Q3.

Why Use the IQR?

- The IQR is robust to outliers because it only considers the middle 50% of the data.
- It's often used alongside box plots to identify outliers.

Example :

Let's calculate the IQR for the dataset: $[5, 10, 15, 20]$.

1. Sort the data (already sorted): $[5, 10, 15, 20]$.
2. Find Q1 and Q3:
   - Q1: Median of the lower half ($[5, 10]$) = $\frac{5+10}{2} = 7.5$
   - Q3: Median of the upper half ($[15, 20]$) = $\frac{15+20}{2} = 17.5$
3. Calculate the IQR:

$$\text{IQR} = Q3 - Q1 = 17.5 - 7.5 = 10$$

Real-World Example :

Imagine you're analyzing test scores: $[50, 60, 70, 80, 90]$. The IQR is:

$$\text{IQR} = Q3 - Q1 = 80 - 60 = 20$$

This tells you that the middle 50% of students scored between 60 and 80.

Real-World Example :

Imagine you're analyzing test scores: $[50, 60, 70, 80, 90]$. The IQR is:

$$\text{IQR} = Q3 - Q1 = 80 - 60 = 20$$

This tells you that the middle 50% of students scored between 60 and 80.

Exercise :

Calculate the IQR for the dataset: $[8, 12, 15, 20, 25]$.

- Solution:
  1. Sort the data (already sorted): $[8, 12, 15, 20, 25]$.
  2. Find Q1 and Q3:
     - Q1: Median of the lower half ($[8, 12]$) = $\frac{8+12}{2} = 10$
     - Q3: Median of the upper half ($[20, 25]$) = $\frac{20+25}{2} = 22.5$
  3. Calculate the IQR:

$$\text{IQR} = Q3 - Q1 = 22.5 - 10 = 12.5$$

2.3 Data Visualization

Visualizations help us understand patterns, trends, and outliers in the data. They make complex datasets easier to interpret.

2.3.1 Histograms

A histogram shows the distribution of numerical data by dividing it into bins (intervals).

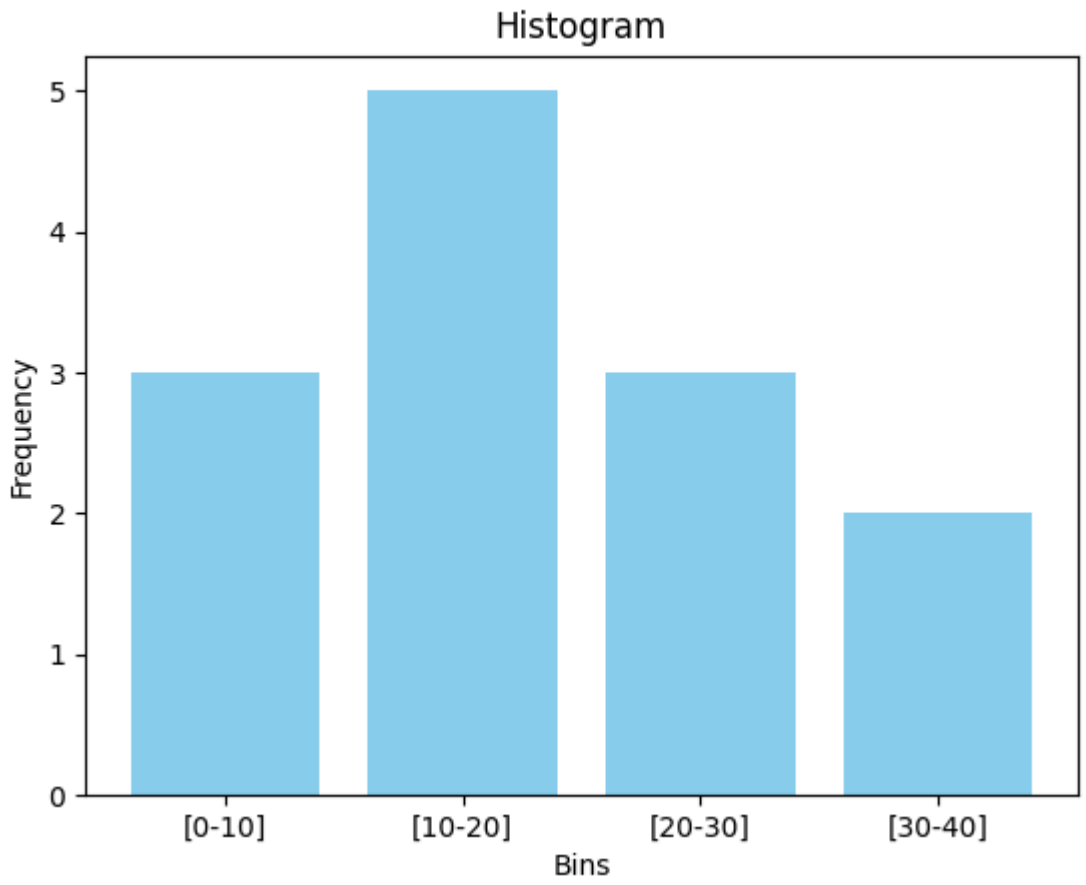
What Is a Histogram? A histogram groups data into bins and shows the frequency (count) of values in each bin. It helps visualize the shape of the data (e.g., symmetric, skewed).

Why Use Histograms? Histograms reveal the overall pattern of the data and identify outliers.

Example :

Let's create a histogram for the dataset: [5,10,15,20, 29, 21, 15, 14, 9, 6, 35, 31, 11].

- Choose bins: [0−10],[10−20] [20-30], [30 -40].



2.3.2 Box Plots

A box plot summarizes the distribution of data using quartiles and identifies outliers.

What Is a Box Plot?

A box plot shows:

- Minimum, Q1, Median, Q3, Maximum
- Outliers: Points outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$

‘

Why Use Box Plots?

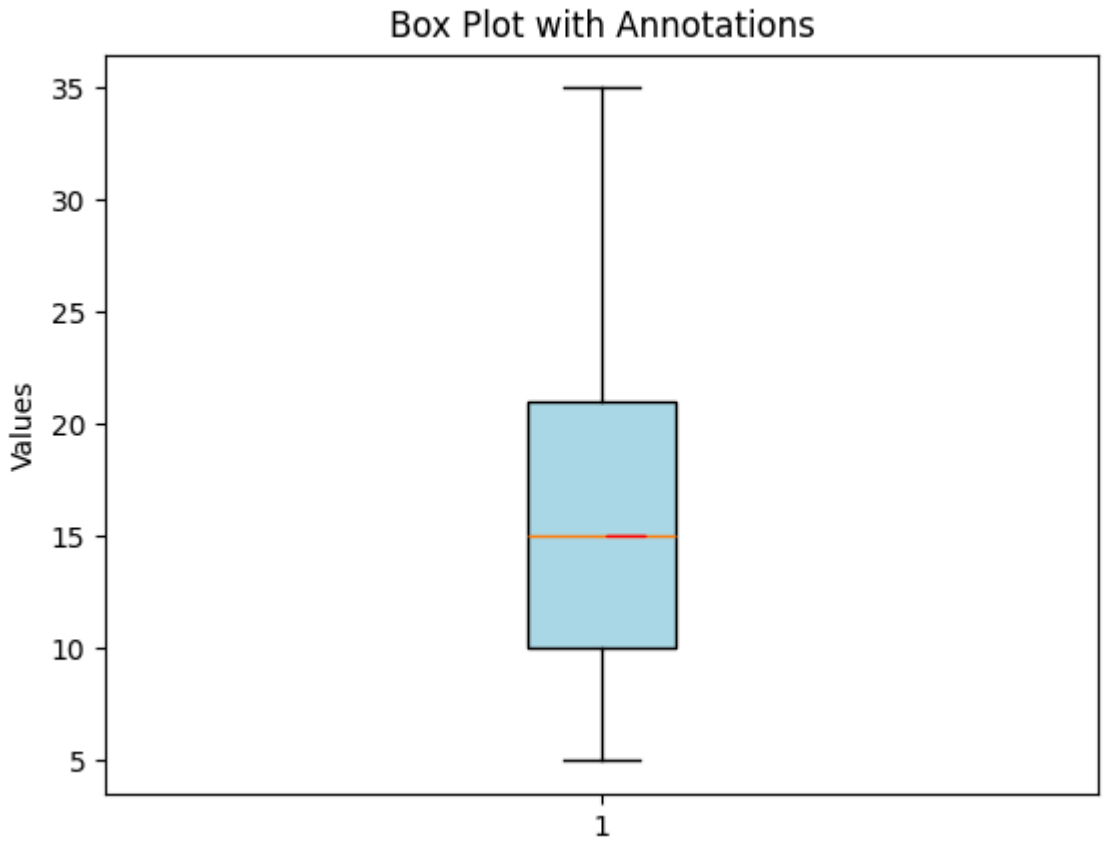
- Box plots are excellent for comparing distributions across groups and detecting outliers.

Example :

Let's create a box plot for the dataset: [5, 10, 15, 20, 29, 21, 15, 14, 9, 6, 35, 31, 11].

- Q1: 10.0
- Median: 15.0
- Q3: 21.0
- IQR: 11.0
- Lower Outlier Bound: -6.5
- Upper Outlier Bound: 37.5

Formulas (Lower Bound: $Q1-1.5\times IQR$ and Upper Bound: $Q3+1.5\times IQR$]



2.3.3 Scatter Plots

A scatter plot shows the relationship between two numerical variables.

A scatter plot is a powerful tool for visualizing the relationship between two numerical variables. It helps us identify trends, correlations, and patterns in the data.

Why Use Scatter Plots?

- Scatter plots are particularly useful when you want to explore whether two variables are related.
- They can reveal:
    - Positive Correlation : As one variable increases, the other tends to increase.
    - Negative Correlation : As one variable increases, the other tends to decrease.
    - No Correlation : No clear relationship between the variables.

Step-by-Step Process :

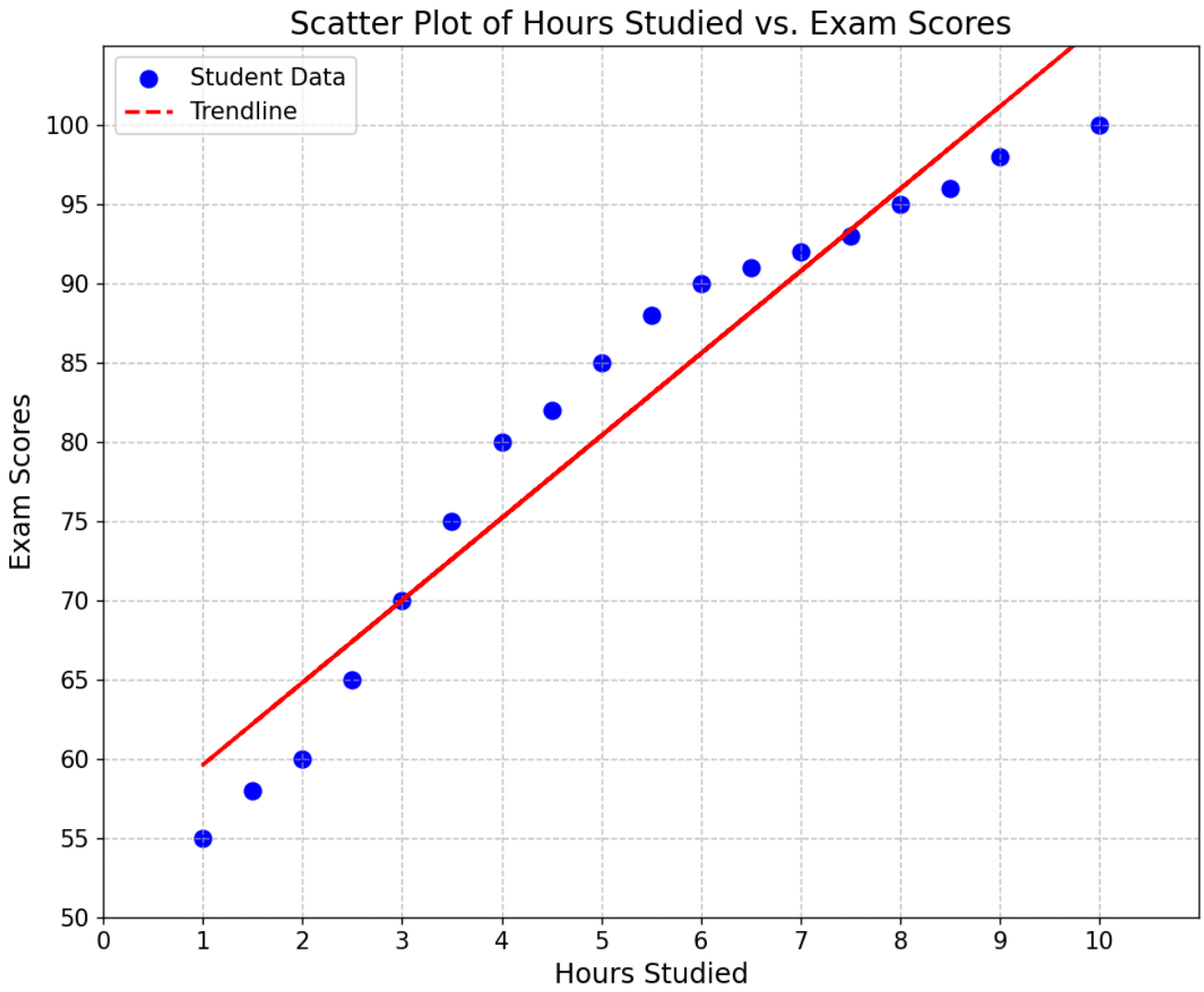
- Identify the two variables you want to compare.
- Plot each data point on a graph, where:
    - The x-axis represents one variable.
    - The y-axis represents the other variable.
- Look for patterns or trends in the plotted points.

Example :
Let's create a scatter plot for hours studied vs. exam scores:

Dataset: [(2, 60), (4, 80), (6, 90), (8, 95), (1, 55), (3, 70), (5, 85), (7, 92), (9, 98), (2.5, 65), (3.5, 75), (5.5, 88), (7.5, 93), (10, 100), (1.5, 58), (4.5, 82), (6.5, 91), (8.5, 96)]

Interpretation:

- As hours studied increase, exam scores tend to increase as well.
- This suggests a positive correlation between study time and performance.



Real-World Example :

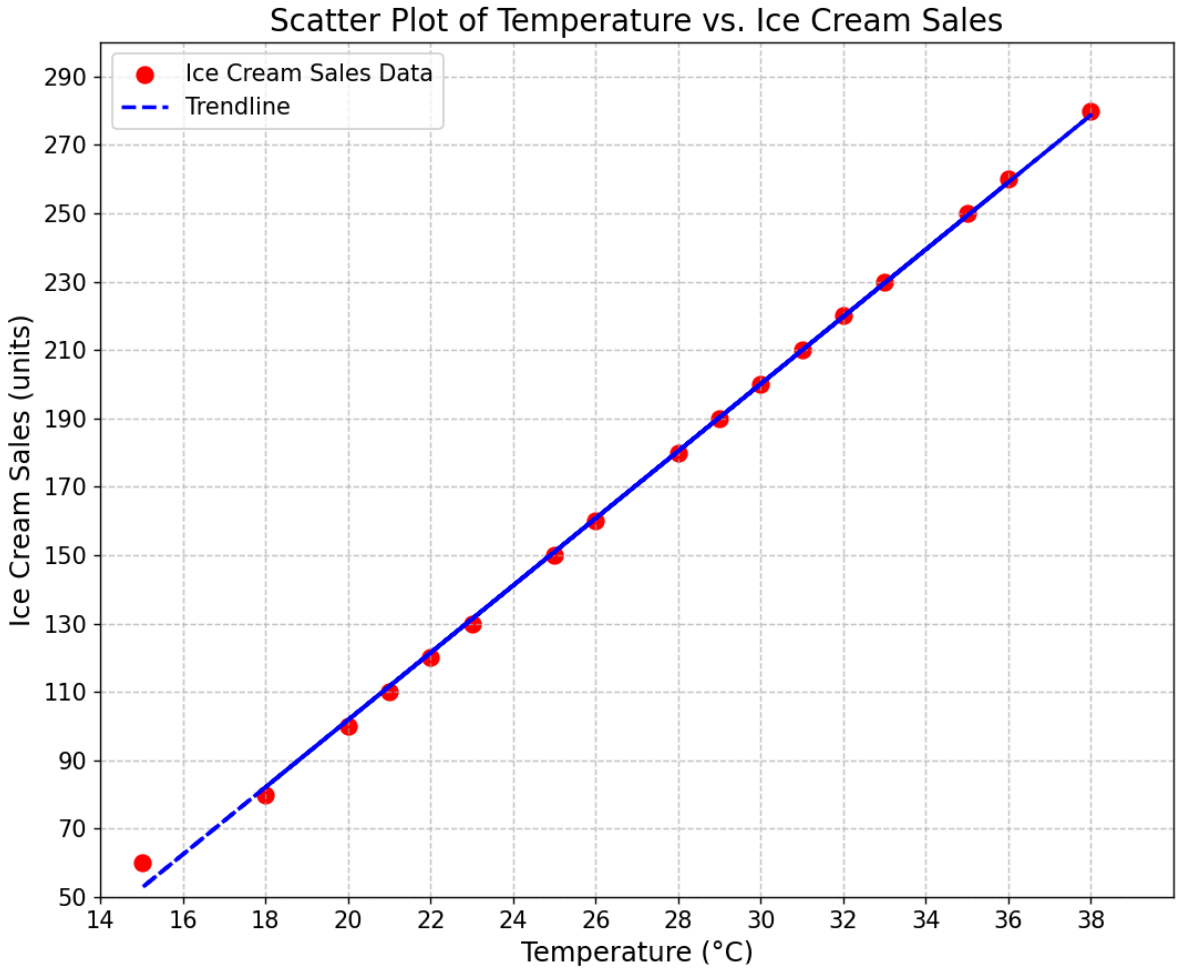
Suppose you're analyzing the relationship between advertising spend and sales revenue for a company. A scatter plot might show that higher advertising budgets lead to increased sales, indicating a positive correlation.

Exercise :

- Create a scatter plot for the following dataset, which shows the relationship between temperature (°C) and ice cream sales (units):

[(20, 100), (25, 150), (30, 200), (35, 250), (18, 80), (22, 120), (28, 180), (32, 220), (38, 280), (21, 110), (26, 160), (31, 210), (36, 260), (15, 60), (23, 130), (29, 190), (33, 230)

]



Solution:

- Plot temperature on the x-axis and ice cream sales on the y-axis.
- Interpretation:
    - As temperature increases, ice cream sales also increase, suggesting a positive correlation.

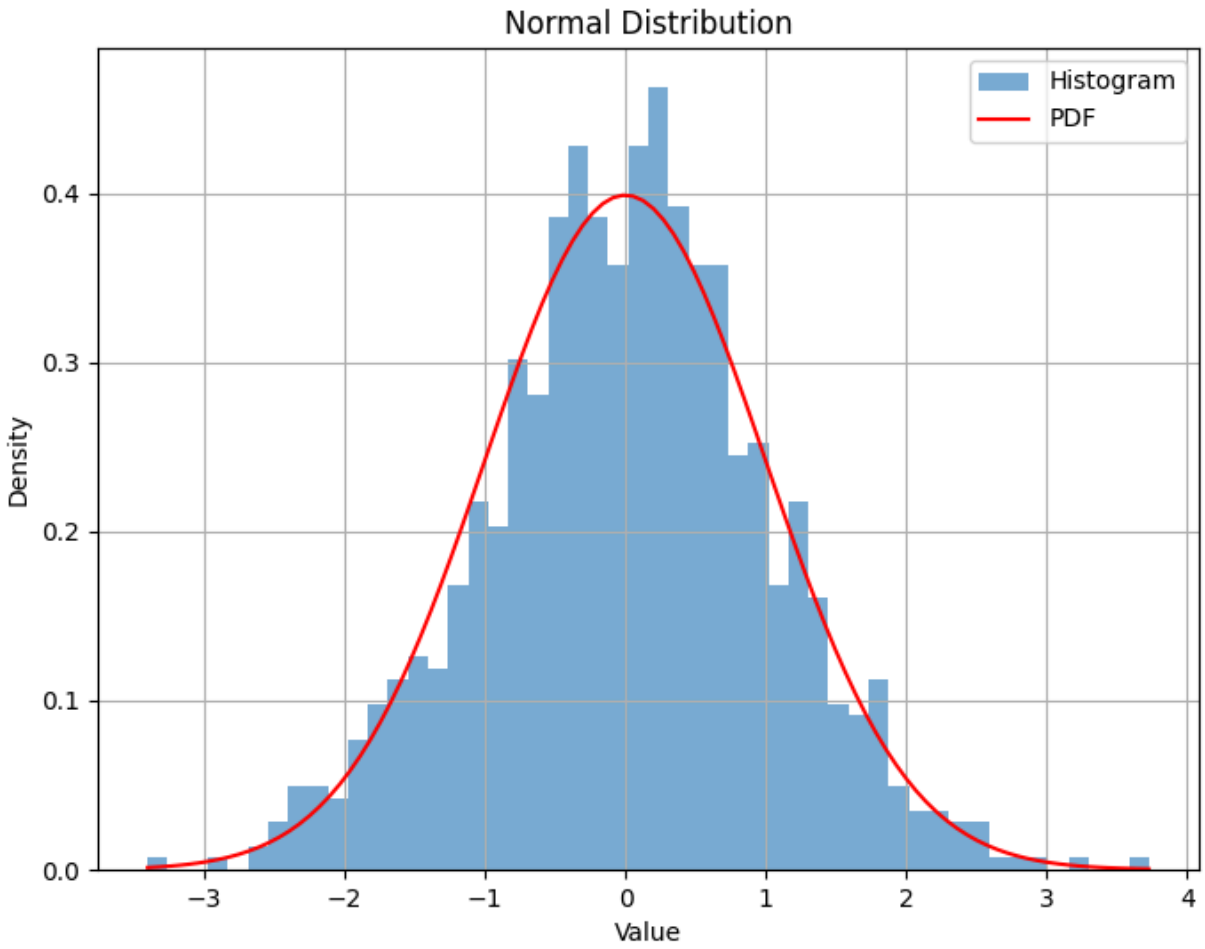
Skewness and Kurtosis

These are measures of the shape of the data distribution. While we've discussed measures like mean, median, and standard deviation, skewness and kurtosis provide deeper insights into how the data is distributed.

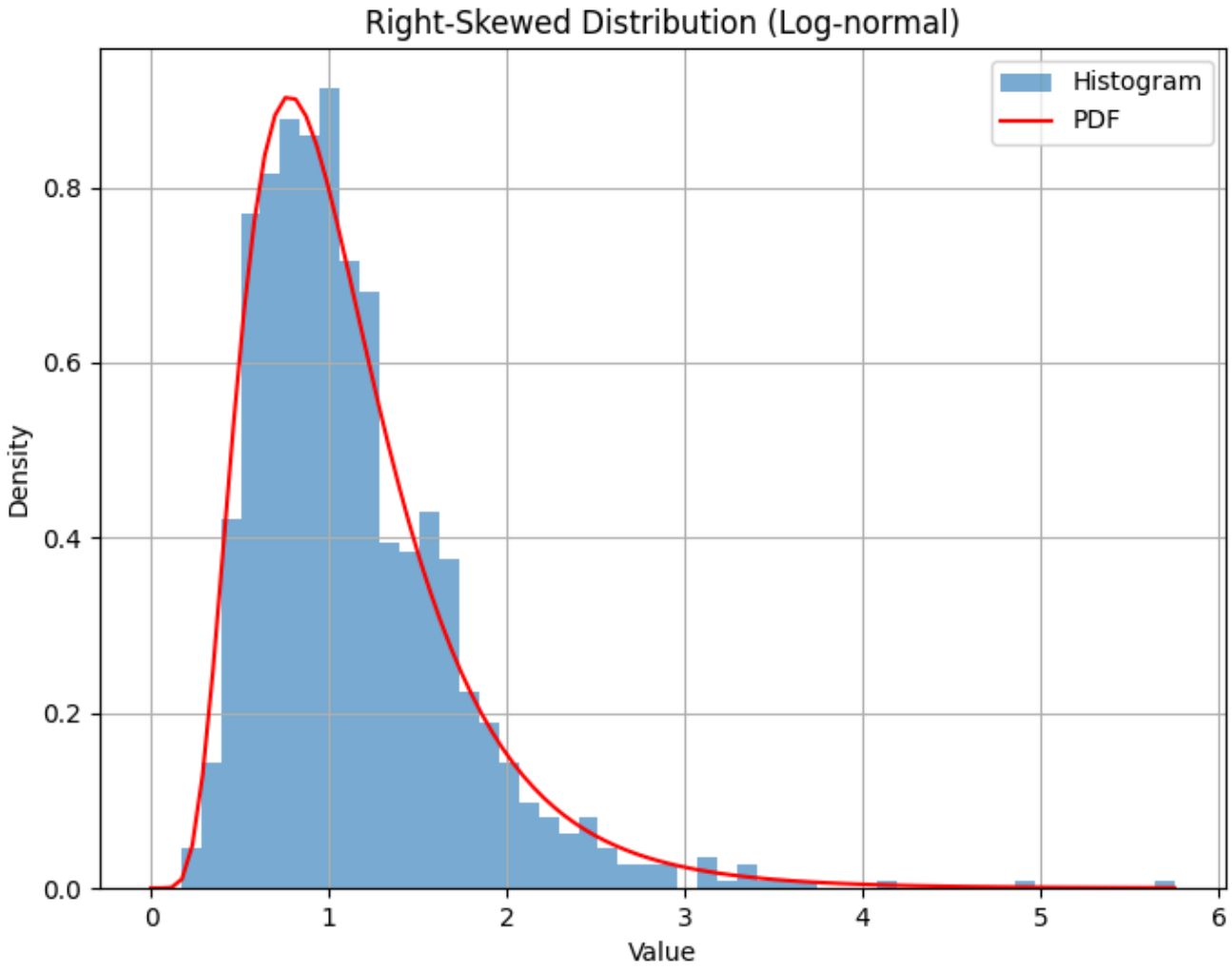
1.1 Skewness

What Is Skewness? Skewness measures the asymmetry of the data distribution.

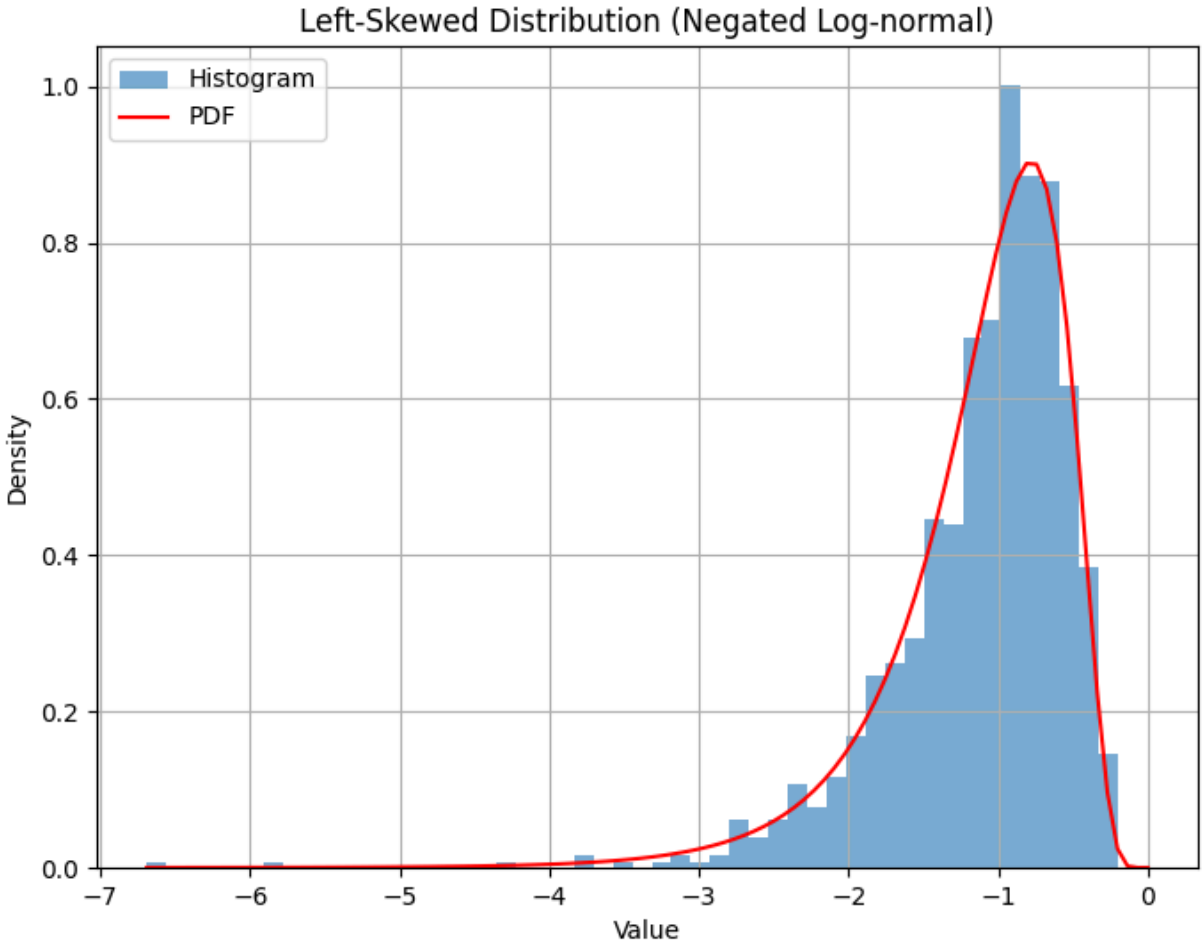
A symmetric distribution (like the normal distribution) has zero skewness.



Positive skewness: The tail on the right side is longer (data is skewed to the right).



Negative skewness: The tail on the left side is longer (data is skewed to the left).



Why Use Skewness?

- Skewness helps identify whether the data is evenly distributed or if it has a long tail on one side.
- It's particularly useful when deciding whether to use the mean or median as a measure of central tendency.
- Example :
    - Dataset: [5,10,15,20,100]
        - Mean: 30, Median: 15

- The mean is much larger than the median, indicating positive skewness.

Real-World Example : Income data often has positive skewness because a small number of people earn extremely high incomes, pulling the mean upward.

1.2 Kurtosis

- What Is Kurtosis?
  - Kurtosis measures the "tailedness" or "peakedness" of the data distribution.
  - High kurtosis: The distribution has heavy tails and a sharp peak (leptokurtic).
  - Low kurtosis: The distribution has light tails and a flatter peak (platykurtic).
  - Normal distributions have a kurtosis value of 3 (often adjusted to 0 for simplicity).
- Why Use Kurtosis?
  - Kurtosis helps identify whether the data has extreme outliers or is tightly clustered around the mean.
- Example :
  - Dataset: [5,10,15,20,100]
    - The presence of the extreme value (100) suggests high kurtosis.
- Real-World Example :
  - Stock market returns often exhibit high kurtosis due to occasional extreme price movements.

---

2. Weighted Mean

The weighted mean is a variation of the mean that accounts for different weights assigned to data points. This is useful when some values are more important than others.

Formula:

$$\text{Weighted Mean} = \frac{\sum(w_i \cdot x_i)}{\sum w_i}$$

where:

- x_i = individual data points
- w_i = weights assigned to each data point

Why Use the Weighted Mean?

- The weighted mean is used when certain data points contribute more to the overall average (e.g., grades with different credit hours).

Example :

- Grades: $[80, 90, 70]$ with weights: $[3, 2, 1]$
- Weighted Mean:

$$\text{Weighted Mean} = \frac{(80 \cdot 3) + (90 \cdot 2) + (70 \cdot 1)}{3 + 2 + 1} = \frac{240 + 180 + 70}{6} = \frac{490}{6} \approx 81.67$$

Real-World Example :

- Calculating GPA (Grade Point Average) involves assigning weights to grades based on credit hours.

3. Percentiles and Quartiles

We briefly touched on quartiles (Q1, Q2, Q3) earlier, but let's expand on percentiles , which divide the data into 100 equal parts.

- What Are Percentiles?
  - The p-th percentile is the value below which p% of the data falls.
  - For example, the 50th percentile is the median.
- Why Use Percentiles?

    - Percentiles are useful for understanding the relative position of a data point within the dataset.
    - Common percentiles include:
        - 25th percentile (Q1)
        - 50th percentile (Median)
        - 75th percentile (Q3)
- Example :
    - Dataset: $[10,20,30,40,50]$
    - 20th percentile: The value below which 20% of the data falls.
        - Position: $0.2 \cdot (n+1) = 0.2 \cdot 6 = 1.2$
        - Interpolate between the 1st and 2nd values: $10 + 0.2 \cdot (20-10) = 12$
- Real-World Example :
    - Standardized test scores (e.g., SAT, GRE) are often reported in percentiles to show how a student performed relative to others.

4. Outliers

Outliers are extreme values that deviate significantly from the rest of the data. Identifying and handling outliers is an important part of descriptive statistics.

- What Are Outliers?
    - Outliers can be identified using:
        - Z-scores: Values more than 3 standard deviations from the mean.
        - IQR Method: Values outside $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$.
- Why Handle Outliers?
    - Outliers can skew measures like the mean and variance, leading to misleading conclusions.
    - They may indicate errors in data collection or rare events worth investigating.
- Example :
    - Dataset: $[5, 10, 15, 20, 100]$
        - Using the IQR method:
            - Q1: 10, Q3: 20, IQR: 10
            - Lower Bound: $10 - 15 = -5$
            - Upper Bound: $20 + 15 = 35$
            - 100 is an outlier.
- Real-World Example :
    - In financial data, outliers might represent fraudulent transactions or market anomalies.

5. Coefficient of Variation (CV)

The coefficient of variation is a relative measure of dispersion that compares the standard deviation to the mean.

- What Is the Coefficient of Variation?
    - Formula

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

- Expressed as a percentage.

Why Use CV?

- CV allows you to compare the variability of datasets with different units or scales.
- A lower CV indicates less relative variability.

Example :

- Dataset 1: $[10, 20, 30]$
    - Mean: 20, Standard Deviation: 10
    - CV: $\frac{10}{20} = 0.5$ (50%)
- Dataset 2: $[100, 200, 300]$
    - Mean: 200, Standard Deviation: 100
    - CV: $\frac{100}{200} = 0.5$ (50%)
- Both datasets have the same relative variability.

Real-World Example :

- Comparing the variability of stock prices vs. bond yields.

2.4 Why Descriptive Statistics Matters

Now that we've covered measures of central tendency, dispersion, and visualization, let's take a moment to reflect on why descriptive statistics is so important.

- **Summarizing Large Datasets :** Real-world datasets often contain thousands or even millions of data points. Descriptive statistics allows us to summarize this information into a few key numbers or visualizations, making it easier to understand.
- **Identifying Patterns and Trends :** By calculating measures like the mean, median, and standard deviation, we can quickly identify the "center" and "spread" of the data. Visualizations like histograms and scatter plots help us spot trends and relationships.
- **Making Informed Decisions :** Descriptive statistics provides the foundation for decision-making. For example, a business might use the mean and standard deviation of sales data to set targets or identify underperforming products.
- **Preparing for Advanced Analysis :** Before diving into advanced statistical techniques like hypothesis testing or machine learning, it's crucial to first understand the basic characteristics of your data using descriptive statistics.

Created by Shailesh | *https://beginnersblog.org* | *https://t.me/openailearning*