# Types of Data

## Qualitative or Categorical Data

Qualitative data, also known as categorical data, describes data that fits into specific categories. It is non-numerical and involves categorical variables that define features such as a person’s gender or hometown. Categorical measures are defined using natural language rather than numbers.

Sometimes categorical data can hold numerical values (quantitative values), but these values do not have mathematical meaning. Examples include birthdates, favorite sports, and school postcodes. Although birthdates and school postcodes contain numerical values, they do not have a numerical significance in mathematical operations.

### Nominal Data

- A type of qualitative information that labels variables without assigning numerical values.
- Also known as the nominal scale.
- Cannot be ordered or measured.
- Examples: Letters, symbols, words, gender, etc.
- Examined using the **grouping method**, where data is grouped into categories and analyzed by frequency or percentage.
- **Visualization**: Represented using **pie charts**.

### Ordinal Data

- A type of data that follows a natural order.
- Unlike nominal data, ordinal data has a meaningful order, but the differences between values are not measurable.
- Found in surveys, finance, economics, and questionnaires.
- **Visualization**: Represented using **bar charts**.
- Often expressed using tables, where each row represents a distinct category.

---

## Quantitative or Numerical Data

Quantitative data, also known as numerical data, represents values in numerical form (e.g., how much, how often, how many). It provides information about quantities of specific things.

Examples include height, length, size, and weight.

Numerical data is classified into **discrete data** and **continuous data**.

### Discrete Data

- Can take only **distinct values**.
- Contains a finite number of possible values.
- Values **cannot be subdivided meaningfully**.
- **Example**: Number of students in a class.

### Continuous Data

- Can be **measured** rather than counted.
- Has an **infinite number of possible values** within a given range.
- **Example**: Temperature range.

---

# Inferential Statistics

Inferential statistics involves drawing conclusions or making inferences about a population based on data collected from a sample of that population.

## How It Works

1. **Sampling**: Data is collected from a subset of the population, called a sample.
2. **Analysis**: Various statistical techniques such as means, standard deviations, correlations, or regression coefficients are applied.
3. **Inference**: Generalizations are made about the population based on the analyzed sample data, assuming it is representative.

Inferential statistics includes techniques like hypothesis testing, confidence intervals, and regression analysis to determine statistical significance and generalizability.

## Types of Inferential Statistics

### 1. Hypothesis Testing

Hypothesis testing is a fundamental technique in inferential statistics used to test assumptions about a population parameter using sample data.

### Z-Test

- Used when the population variance is known and the sample size is large (n > 30).
- Based on the standard normal distribution (Z-distribution).
- **Example**: Determining if the mean height of a population differs from 65 inches using a large sample with a known standard deviation.

### T-Test

- Used when the population standard deviation is unknown or the sample size is small (n < 30).
- Based on the Student's t-distribution.
- **Types**:
  - **Independent samples t-test**: Compares means of two independent groups.
  - **Paired samples t-test**: Compares means of two related groups.
- **Example**: Comparing exam scores between two student groups.

### F-Test

- Compares variances of two or more populations.
- Used in Analysis of Variance (ANOVA) to test differences among multiple group means.
- **Example**: Analyzing the effectiveness of three teaching methods on student performance.

## 2. Confidence Intervals

- Provide a range of values where a population parameter is likely to lie with a specific confidence level (e.g., 95%).
- **Example**: Estimating the proportion of voters supporting a candidate within a confidence interval.

## 3. Regression Analysis

- Examines relationships between independent variables and a dependent variable.
- Used for prediction and hypothesis testing about variable relationships.
- **Example**: Predicting exam scores based on hours of study using regression models.

---

# Hypothesis Testing

## What Is Hypothesis Testing?

Hypothesis testing, sometimes called **significance testing**, is a statistical method used to test an assumption about a **population parameter**. The methodology depends on the nature of the data and the purpose of the analysis.

Hypothesis testing assesses the plausibility of a hypothesis using **sample data**, which may come from a larger population or a data-generating process. For simplicity, the term "population" will be used to refer to both cases in the following descriptions.

# Key Takeaways

- Hypothesis testing evaluates the plausibility of a hypothesis using **sample data**.
- It provides **evidence** concerning the validity of a hypothesis.
- **Statistical analysts** test hypotheses by analyzing a **random sample** from the population.
- The **four steps** of hypothesis testing include:
  1. Stating the hypotheses.
  2. Formulating an analysis plan.
  3. Analyzing the sample data.
  4. Evaluating the results.

# How Hypothesis Testing Works ?

In hypothesis testing, an analyst examines a **statistical sample** to determine whether there is sufficient evidence to reject the **null hypothesis**.

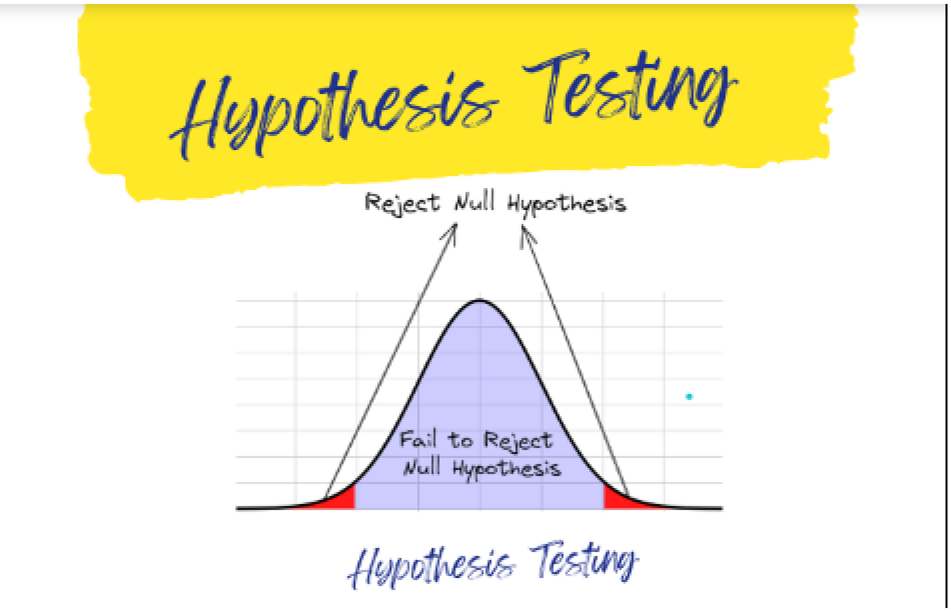
All analysts use a **random population sample** to test two different hypotheses:

1. **Null Hypothesis ($H_0$):** Assumes no effect or no difference (e.g., the population mean return is equal to zero).
2. **Alternative Hypothesis ($H_1$ or $H_a$):** Represents the opposite of the null hypothesis and suggests an effect or a difference.

Since these hypotheses are **mutually exclusive**, only one can be true, but one of them **must be true**.

## The 4-Step Hypothesis Testing Process

1. **State the hypotheses:** Clearly define the null and alternative hypotheses.
2. **Formulate an analysis plan:** Outline how the data will be evaluated.
3. **Analyze the sample data:** Perform statistical calculations based on the chosen method.
4. **Evaluate the results:** Either reject the null hypothesis or state that it remains plausible given the data.



# Confidence Intervals

## What Is a Confidence Interval?

A **confidence interval (CI)** is a range of values that estimates a population parameter with a specified level of confidence. It is calculated as:

**Confidence Interval = Point Estimate ± Margin of Error**

This means that if you were to repeat the same study multiple times, the true population parameter would fall within the confidence interval a certain percentage of the time.

## Understanding Confidence Levels

The **confidence level** represents the probability that the confidence interval contains the true population parameter. It is typically expressed as:

**Confidence Level = 1 − α**

where **α** is the significance level of the statistical test. Common confidence levels include **90%**, **95%**, and **99%**.

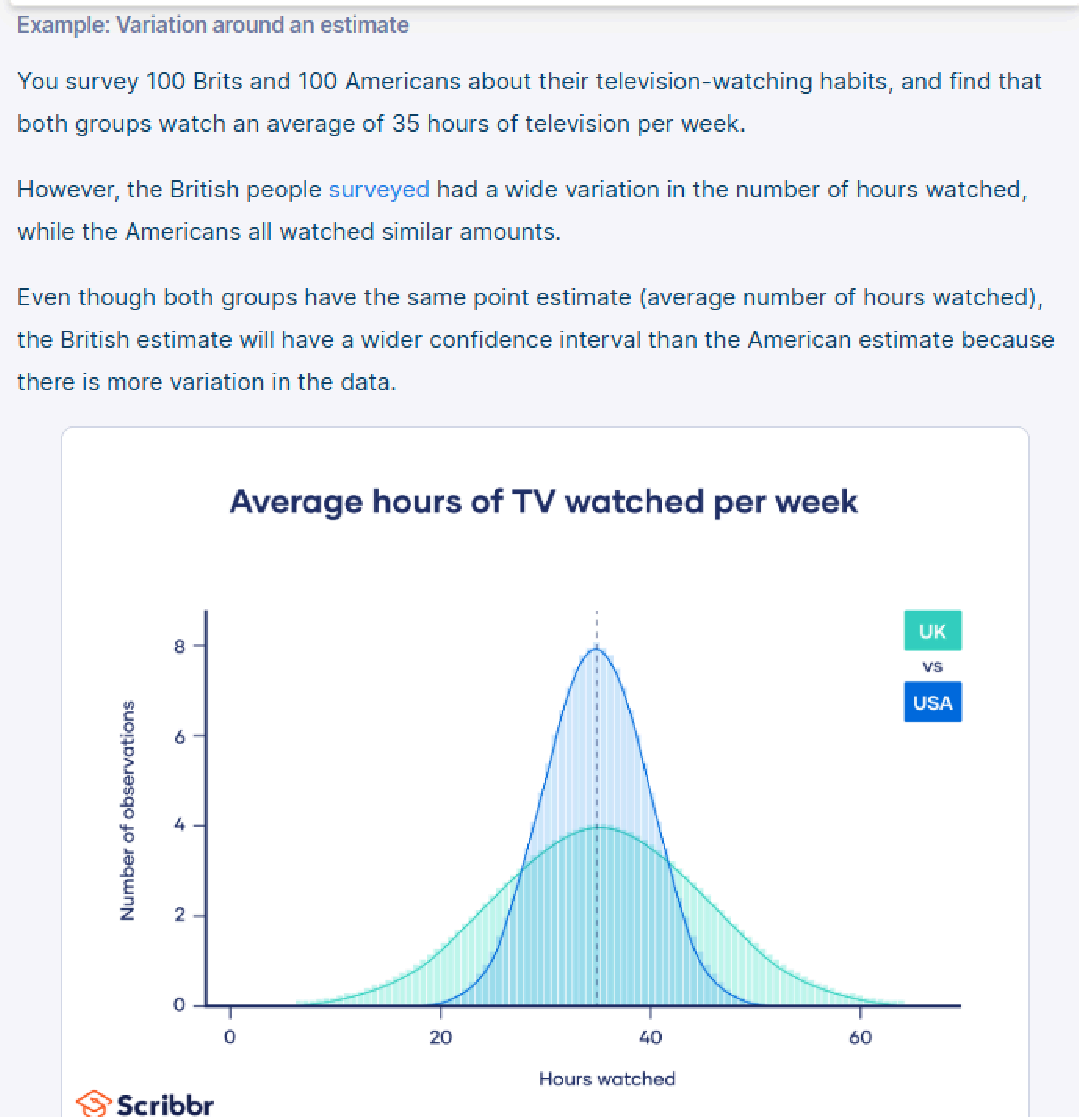
For example, a **95% confidence interval** means that if we were to take 100 samples, approximately 95 of them would contain the true population parameter within their confidence intervals.

## When to Use Confidence Intervals

Confidence intervals are useful in various statistical analyses, including:

- **Estimating population parameters** such as means or proportions.
- **Comparing differences** between groups.
- **Measuring variability** among data points.

By using confidence intervals, analysts can communicate the uncertainty around an estimate rather than relying solely on a single value.



Example: Variation around an estimate

You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.