# **Statistics**

Statistics is the science of collecting and organizing the and analysing data

## Data:

Data refers to facts or pieces of information that can be measured. The most interesting meaning of data is its ability to be measured.

**Types of Statistics**

Statistics can be broadly classified into Descriptive Statistics and Inferential Statistics:

**1. Descriptive Statistics**

Descriptive statistics summarize and organize data to describe the main features of a dataset. It focuses on presenting data meaningfully to identify patterns and trends.

**Measures of Central Tendency**: These describe the center of a dataset.

Mean (average)

Median (middle value)

Mode (most frequent value)

**Measures of Dispersion**: These describe the spread or variability in the dataset.

Range

Variance

Standard Deviation

**Data Visualization Tools**: Graphical representations to visualize data.

Histograms

Pie charts

Bar graphs

Box plots

**2. Inferential Statistics**

Inferential statistics involve making predictions, inferences, or decisions about a population based on a sample of data. This type of statistics uses probability theory to estimate population parameters and test hypotheses.

**Estimation:**

**Point estimation** (e.g., estimating the mean of a population)

**Interval estimation** (e.g., confidence intervals)

**Hypothesis Testing**: Testing assumptions about a population (e.g., null hypothesis vs. alternative hypothesis).

**T-tests**

**Chi-square tests**

**ANOVA (Analysis of Variance)**

**Regression Analysis**: Examines relationships between variables.

**Simple regression**

**Multiple regression**

**Sampling Theory: Methods**

|  |
| --- |
| Key Differences Between Descriptive and Inferential Statistics |
| | **Aspect** | **Descriptive Statistics** | **Inferential Statistics** | | --- | --- | --- | | **Purpose** | **Summarize and describe the data** | **Make inferences about a population** | | **Data Usage** | **Deals with the entire dataset** | **Analyzes a sample to generalize about a population** | | **Methods** | **Measures of central tendency and dispersion** | **Hypothesis testing, confidence intervals** | | **Tools** | **Charts, graphs, and tables** | **Probability distributions and statistical tests** | |

# Sampling techniques

**What is Sampling?**

**Sampling** is the process of selecting a subset (called a **sample**) from a larger group (called the **population**) in such a way that the sample represents the characteristics of the population. In many cases, it’s not feasible to study an entire population (e.g., all people in a country, all students in a school, etc.), so researchers use sampling to draw conclusions from a smaller, manageable group.

**Notation:**

* **Population (N)**: The entire group from which you want to draw conclusions.
* **Sample (n)**: The smaller subset of the population you will actually study.
* **Sampling Frame**: A list of all individuals or units in the population from which the sample is drawn.
* **Sampling Unit**: The individual element (person, item, etc.) from the population.

**Types of Sampling Techniques**

Now, let’s break down **four types of sampling techniques** with their notations and good examples:

**1. Simple Random Sampling (SRS)**

In **Simple Random Sampling**, every member of the population has an equal chance of being selected for the sample. It’s like randomly picking people out of a group without any bias or systematic order.

**Notation:**

* Suppose you have a population of **N** elements (e.g., 100 students).
* You want to select **n** elements (e.g., 10 students).
* The probability of selecting any one individual is nN\frac{n}{N}Nn​.

**Example:**

Imagine a class of **100 students** (population size **N = 100**) and you want to randomly select **10 students** for a survey (sample size **n = 10**).

* You assign a unique number to each student (1 through 100).
* Using a random number generator, you randomly select 10 numbers (e.g., 3, 17, 25, etc.).
* The students whose numbers match the randomly generated ones will be selected.

Since every student had the same chance of being selected, this is a simple random sample.

**Key Point: Every member has an equal chance of selection.**

**2. Stratified Sampling**

In **Stratified Sampling**, the population is divided into distinct subgroups, called **strata**, that share similar characteristics. A random sample is then selected from each stratum, ensuring that each subgroup is well-represented in the final sample.

**Notation:**

* **Population (N)** is divided into **k** strata.
* **n₁, n₂, ..., nk** are the sample sizes from each stratum.
* **N₁, N₂, ..., Nk** are the sizes of each stratum.

**Example:**

Suppose you have a population of **100 students** with **40 men** and **60 women**, and you want to select a sample of **10 students**.

* You divide the population into two strata: **men** (N₁ = 40) and **women** (N₂ = 60).
* You decide to sample **4 men** and **6 women** to ensure representation from both genders.

Here, you select a random sample of **4 men** and **6 women**. This ensures that both men and women are represented proportionally in the sample.

**Key Point: Divide the population into strata and sample from each group.**

**3. Systematic Sampling**

In **Systematic Sampling**, you select every *k-th* individual from a list of the population, after choosing a random starting point. This method is useful when the population is already ordered in some way.

**Notation:**

* **N** is the total population size.
* **n** is the sample size.
* **k** is the sampling interval, calculated as k=Nnk = \frac{N}{n}k=nN​.

**Example:**

Suppose you have a list of **100 employees** and you want to select a sample of **10 employees**.

Calculate the sampling interval k=10010=10k = \frac{100}{10} = 10k=10100​=10.

* Randomly choose a starting point between 1 and 10 (e.g., 3).
* From the 3rd employee, you select every 10th employee: 3rd, 13th, 23rd, 33rd, and so on.

This way, you are systematically selecting every *10th* individual after a random start.

**Key Point: Every *k-th* individual is selected after a random starting point.**

**4. Convenience Sampling**

In **Convenience Sampling**, you select individuals who are easiest to reach or most convenient for the researcher. This is often used because it’s quick and easy, but it can lead to bias because the sample might not represent the entire population well.

**Notation:**

* **N** is the total population size.
* **n** is the sample size.

**Example:**

Imagine you are a researcher doing a survey on student preferences. Instead of randomly selecting students, you simply ask the first 10 students you see in the cafeteria to participate.

While it’s convenient and easy, this sample may not represent all students accurately, especially if the cafeteria only has students from one particular class or group.

**Key Point: The sample is chosen based on what’s easiest or most convenient, which can introduce bias.**

**Summary of Sampling Types**

|  |  |  |
| --- | --- | --- |
| Sampling Type | Definition | Example |
| Simple Random Sampling | Every individual has an equal chance of selection. | Randomly picking 10 students from 100. |
| Stratified Sampling | Divide the population into strata and sample from each. | Selecting 4 men and 6 women from 100 students. |
| Systematic Sampling | Select every *k-th* individual after a random start. | Selecting every 10th employee from a list. |
| Convenience Sampling | Choose the easiest or most convenient individuals. | Asking 10 students in the cafeteria. |

These sampling techniques help in different scenarios and are essential for conducting accurate and unbiased research. **Simple Random Sampling** is ideal for general population surveys, **Stratified Sampling** is great when you want to ensure all subgroups are represented, **Systematic Sampling** is efficient when the population is ordered, and **Convenience Sampling** is quick but may lead to biased results.

# Types of variables

**What are Variables in Statistics?**

A **variable** is something that you are measuring or observing. It can change or vary, which is why we call it a "variable." For example, in a survey about students, the **age** or **gender** of the students are variables.

Variables in statistics can be divided into two main types: **Quantitative** (numbers) and **Qualitative** (categories).

**1. Quantitative Variables (Numbers)**

These are **numbers** that you can measure or count. You can perform math with them (like adding or subtracting). Quantitative variables are about **how much** or **how many**.

**Examples:**

* **Age**: 15 years, 18 years, 20 years (you can count how many years)
* **Height**: 5 feet, 5.5 feet, 6 feet (you can measure the height)
* **Number of pets**: 2 pets, 3 pets, 5 pets (you can count the number of pets)

**Types of Quantitative Variables:**

1. **Discrete**: These are whole numbers (countable).
   * Example: Number of books in a library (you can’t have 2.5 books).
2. **Continuous**: These can be any number, including decimals or fractions.
   * Example: Weight (you can have 60.5 kg, 70.3 kg).

**2. Qualitative Variables (Categories)**

These are **categories** or **labels** that describe something. You **cannot** do math with them because they are not numbers. Qualitative variables are about **what type** or **what kind**.

**Examples:**

* **Color**: Red, Blue, Green (these are categories, not numbers)
* **Gender**: Male, Female, Other (labels)
* **Type of Fruit**: Apple, Banana, Orange (categories)

**Types of Qualitative Variables:**

1. **Nominal**: These are categories with **no order**. There’s no “higher” or “lower” category.
   * Example: **Eye color** (Blue, Green, Brown) — no order between colors.
2. **Ordinal**: These are categories with a **meaningful order** or ranking.
   * Example: **Education level** (High School, Bachelor’s, Master’s) — there’s an order to the levels.

**Summary of Quantitative vs. Qualitative Variables**

| **Type of Variable** | **What It Is** | **Examples** |
| --- | --- | --- |
| **Quantitative** (Numbers) | Values you can count or measure. | Age, Height, Salary, Number of pets |
| **Qualitative** (Categories) | Describes qualities or categories, not numbers. | Color, Gender, Type of fruit |

**Real-World Examples to Make It Clear:**

**Example 1: A Survey About Students**

Imagine you are conducting a survey about students.

* **Quantitative Variables**:
  + **Age** (15 years, 18 years, 20 years)
  + **Number of books** (5 books, 10 books)
* **Qualitative Variables**:
  + **Favorite subject** (Math, Science, English) — **Nominal**, because no subject is “better” or “worse” than the others.
  + **Grade level** (Freshman, Sophomore, Junior, Senior) — **Ordinal**, because the grade levels have a clear order.

**Example 2: A Study on Animals**

Suppose you want to study animals in a zoo.

* **Quantitative Variables**:
  + **Number of legs** (4, 2, 8)
  + **Weight** (5 kg, 10 kg, 15 kg)
* **Qualitative Variables**:
  + **Animal type** (Lion, Tiger, Elephant) — **Nominal**, because there’s no ranking of animals.
  + **Animal size** (Small, Medium, Large) — **Ordinal**, because there is an order from small to large.

**In Short:**

* **Quantitative variables** are **numbers** you can measure or count, like **age** or **height**.
* **Qualitative variables** are **categories** that describe things, like **color** or **gender**. Some can have an order (like **education level**), and some just label things without any order (like **eye color**).

Data Visualization Tools:

Graphical representations to visualize data.

Histograms

Pie charts

Bar graphs

Box plots

Intermediate

# **Measure of Central Tendency**

**What is the Measure of Central Tendency?**

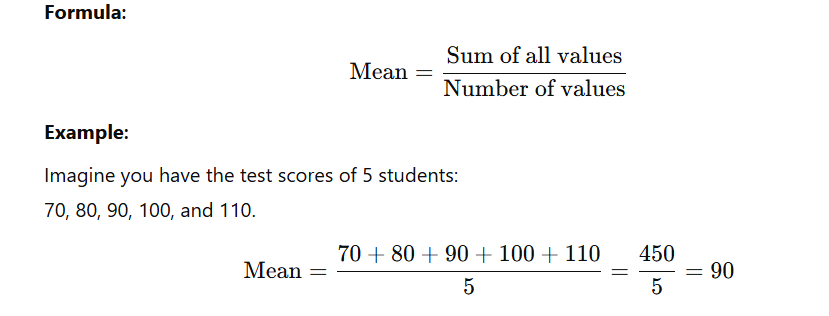
* The **Measure of Central Tendency** is a way to describe the **center point** or typical value of a dataset. It gives us a single value that summarizes an entire dataset and represents the "average" or "middle" of the data.

The three most common measures of central tendency are:

1. **Mean** (Arithmetic Average)
2. **Median** (Middle Value)
3. **Mode** (Most Frequent Value)

**1. Mean (Arithmetic Average)**

The **mean** is the most common and familiar measure of central tendency. It is calculated by **adding up all the values in a dataset and dividing by the total number of values**.

****

**Key Points:**

* The mean is influenced by extreme values (outliers). For example, if one test score was 500, the mean would increase significantly, even though it doesn’t represent most of the students' scores.

**2. Median (Middle Value)**

The median is the middle value when all the numbers in a dataset are arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle numbers.

**How to Find the Median:**

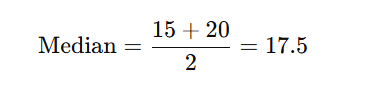
1. Arrange the data in order.
   * Find the middle value (or average the two middle values if there are an even number of values).

**Example 1 (Odd Number of Values):**

Dataset: 10, 15, 20, 25, 30  
The middle value is **20**, so the median is **20**.

**Example 2 (Even Number of Values):**

* Dataset: 10, 15, 20, 25  
  The middle values are **15** and **20**.



So, the median is **17.5**.

**Key Points:**

* The median is **not affected by outliers**. For example, if one of the values in the dataset is 500, the median remains unchanged, unlike the mean.

**3. Mode (Most Frequent Value)**

The mode is the value that **appears** most frequently in a dataset. A dataset can have:

* **No mode** (if no value repeats),
* **One** mode (unimodal), or

**Multiple modes** (bimodal or multimodal).

**Example 1:**

Dataset: 10, 15, 20, 20, 25  
The number **20** appears twice, so the mode is **20**.

**Example 2:**

Dataset: 10, 15, 20, 20, 25, 25  
Both **20** and **25** appear twice, so this dataset is **bimodal** (two modes).

**Key Points:**

* The mode is often used for categorical data (like colors, brands, etc.).
* For example, in a survey of favorite colors: Blue (5 votes), Red (7 votes), Green (3 votes) — the mode is **Red**.

**Comparison of Mean, Median, and Mode**

| **Measure** | **How to Calculate** | **Best When** | **Not Good When** |
| --- | --- | --- | --- |
| **Mean** | Add all values and divide by the count | Data is evenly distributed | There are outliers (extreme values) |
| **Median** | Middle value (or average of 2 middle values) | Data has outliers or skewed distribution | Large datasets where sorting is time-consuming |
| **Mode** | Most frequent value | Data is categorical or has repeated values | Every value is unique |

**Which Measure to Use?**

1. **Use Mean**:
   * When the data is evenly distributed without extreme outliers.
   * Example: Heights of students in a class (if no one is abnormally short or tall).
2. **Use Median**:
   * When there are **outliers** or the data is **skewed**.
   * Example: Income data, where a few people earn significantly more than the majority.
3. **Use Mode**:
   * When working with **categorical data** or finding the most popular value.
   * Example: Finding the most preferred car brand or the most common test score.

**Conclusion**

The mean is useful but can be misleading with outliers.

The median is better for skewed data or data with outliers.

The mode is best for categorical data or when looking for the most common value.

# Measure of Dispersion

**What is Measure of Dispersion?**

The **Measure of Dispersion** tells us how spread out or scattered the data values are in a dataset. It complements measures of central tendency (mean, median, mode) by showing the **variability** in the data.

* It answers the question: **"How far are the data points from the center?"**
* For example, two datasets can have the same mean but very different levels of spread.

**Why is it Important?**

* It helps to understand how consistent or inconsistent the data is.
* A low dispersion means the data points are close to the center (mean/median), while a high dispersion means they are more spread out.
* Dispersion is crucial in fields like data science, finance, and research to assess variability in datasets.

**Types of Measures of Dispersion**

There are two main categories:

1. **Absolute Measures**: Directly measure variability in terms of the original units of data (e.g., range, variance, standard deviation).
2. **Relative Measures**: Measure variability relative to the size of the data (e.g., coefficient of variation).

**1. Range**

The **range** is the simplest measure of dispersion. It is the difference between the largest and smallest value in a dataset.



**Example:**

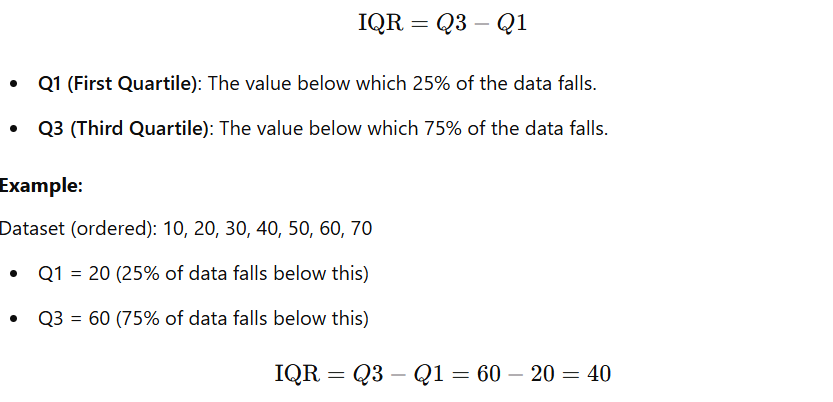
Dataset: 10, 20, 30, 40, 50

Range= 50 - 10 = 40

* **Advantages**: Easy to calculate.
* **Disadvantages**: Only considers two values (max and min), so it is sensitive to outliers.

**2. Interquartile Range (IQR)**

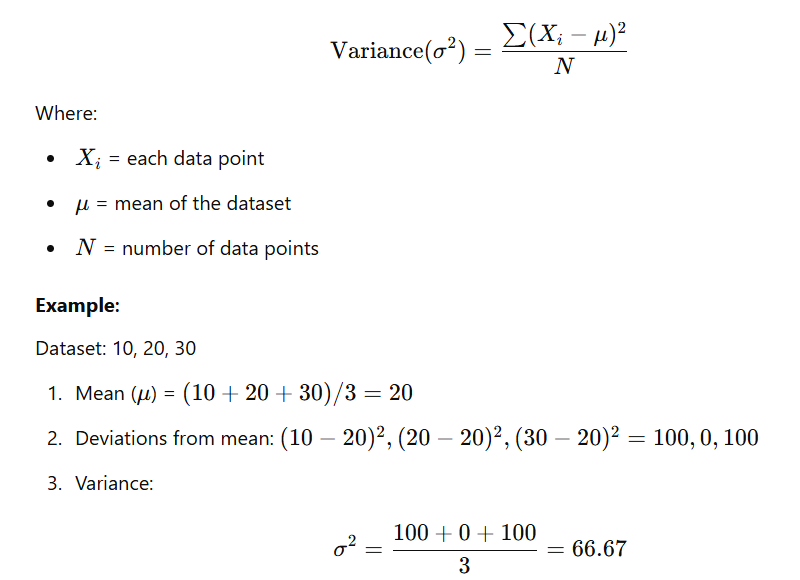
The **IQR** measures the spread of the middle 50% of data. It is the difference between the **third quartile (Q3)** and the **first quartile (Q1)**.



* **Advantages**: Not affected by outliers.
* **Disadvantages**: Ignores variability outside the middle 50%.

**3. Variance**

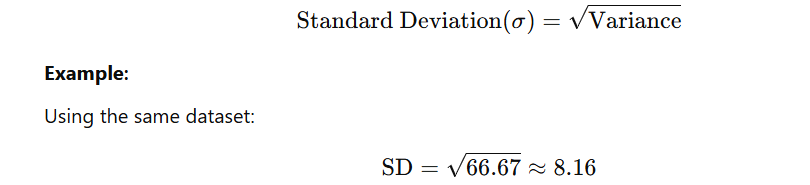
The **variance** measures the average squared deviation of each data point from the mean. It gives a sense of how spread out the data is.



* **Advantages**: Uses all data points.
* **Disadvantages**: Measured in squared units, making it harder to interpret directly.

**4. Standard Deviation**

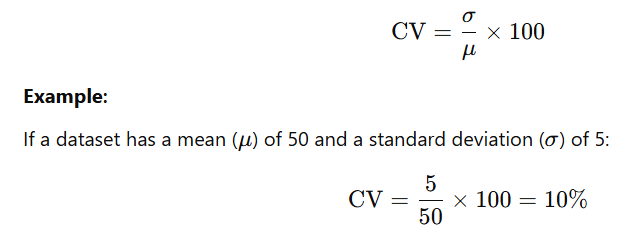
The **standard deviation (SD)** is the square root of the variance. It measures the average distance of each data point from the mean in the same units as the original data.



* **Advantages**: Easy to interpret because it's in the same unit as the data.
* **Disadvantages**: Can be affected by outliers.

**5. Coefficient of Variation (CV)**

The **Coefficient of Variation** is a relative measure of dispersion. It expresses the standard deviation as a percentage of the mean.



* **Advantages**: Useful for comparing variability across datasets with different units or scales.
* **Disadvantages**: Cannot be used if the mean is 0.

**Comparison Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Definition | Best When | Limitations |
| Range | Difference between max and min values | Quick, simple measure of spread | Affected by outliers |
| Interquartile Range | Spread of middle 50% of data | Skewed datasets or when ignoring outliers | Ignores variability outside Q1 and Q3 |
| Variance | Average squared distance from the mean | Measuring spread for analysis or modeling | Hard to interpret directly (squared units) |
| Standard Deviation | Average distance from the mean (same unit as data) | When interpreting spread in original units | Sensitive to outliers |
| Coefficient of Variation | Variability relative to the mean (in %) | Comparing spread across different datasets | Not defined for zero or very small mean |

**Example Use Case**

Suppose you are analyzing test scores of two classes:

* Class A: 70, 80, 90, 100, 110
* Class B: 50, 60, 90, 120, 150

**Central Tendency:**

* Both have the **same mean**: μ=90\mu = 90μ=90.  
  But clearly, Class B has more variability.

**Dispersion:**

* **Range**:
  + Class A: 110−70=40110 - 70 = 40110−70=40
  + Class B: 150−50=100150 - 50 = 100150−50=100
* **Standard Deviation**:
  + Class A has a smaller SD, meaning the scores are closer to the mean.
  + Class B has a larger SD, meaning the scores are more spread out.

**Conclusion**

The **measure of dispersion** is critical in data science to understand variability in data. It helps complement measures of central tendency by providing insights into how spread out the data is, which is essential for decision-making and data modeling.

# Distributions in Statistics

In data science and statistics, **distributions** describe how values in a dataset are spread or organized. Different types of distributions help us model various real-world scenarios. Below are six key distributions explained for beginners, along with examples.

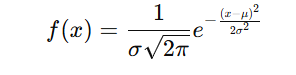
**1. Normal Distribution (Gaussian Distribution)**

The **Normal Distribution** is one of the most commonly used distributions in statistics. It is a symmetrical, bell-shaped curve where most of the data points are concentrated around the mean, and the probabilities decrease as you move further away.

**Key Features:**

* Symmetrical around the mean (μ\muμ).
* The mean, median, and mode are equal.
* Spread is determined by the standard deviation (σ\sigmaσ).
* 68% of the data lies within 1 standard deviation (σ) of the mean, 95% within 2σ, and 99.7% within 3σ. (This is called the 68-95-99.7 rule.)

**Equation:**

****

**Example:**

* Heights of people in a population often follow a normal distribution (e.g., most people have an average height, with fewer being very tall or very short).
* Dataset: Heights of 100 people, with a mean of 170 cm and a standard deviation of 10 cm.

**2. Standard Normal Distribution**

The **Standard Normal Distribution** is a special case of the normal distribution where:

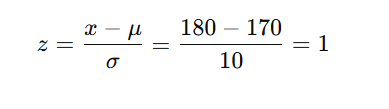
* The mean (μ\muμ) = 0
* The standard deviation (σ\sigmaσ) = 1

It is essentially a **normalized version of the normal distribution**, used to simplify calculations and comparisons.

**Example:**

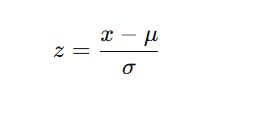
If you have a dataset of heights with a mean of 170 cm and a standard deviation of 10 cm:

* Height of 180 cm corresponds to a **z-score** of:



**3. Z-Score**

The **z-score** (or standard score) tells us how many standard deviations a particular value (x) is away from the mean.

****

Where:

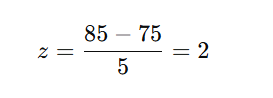
* x = individual data point,
* μ = mean of the dataset,
* σ= standard deviation.

**Uses of Z-Score:**

* To compare scores from different datasets.
* To identify outliers (z-scores beyond ±3pm are considered outliers).

**Example:**

If a student scores 85 in a test where the class mean is 75 and the standard deviation is 5:



The student scored 2 standard deviations above the mean.

**4. Log-Normal Distribution**

A **Log-Normal Distribution** is a distribution where the logarithm of the data follows a normal distribution. It is commonly used when data cannot be negative and grows multiplicatively (e.g., stock prices, income).

**Key Features:**

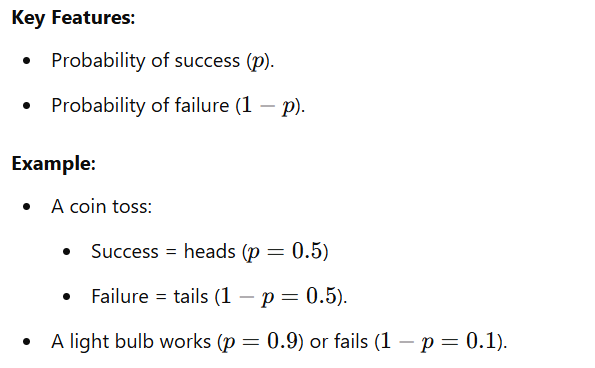
* Skewed to the right (not symmetrical).
* Values are strictly positive.

**Example:**

* Stock prices over time often follow a log-normal distribution.
* If stock prices today are $100, they may follow a log-normal distribution in the future due to compounding growth.

**5. Bernoulli Distribution**

The **Bernoulli Distribution** is the simplest probability distribution. It models a single trial that results in one of two outcomes: **success** (1) or **failure** (0).

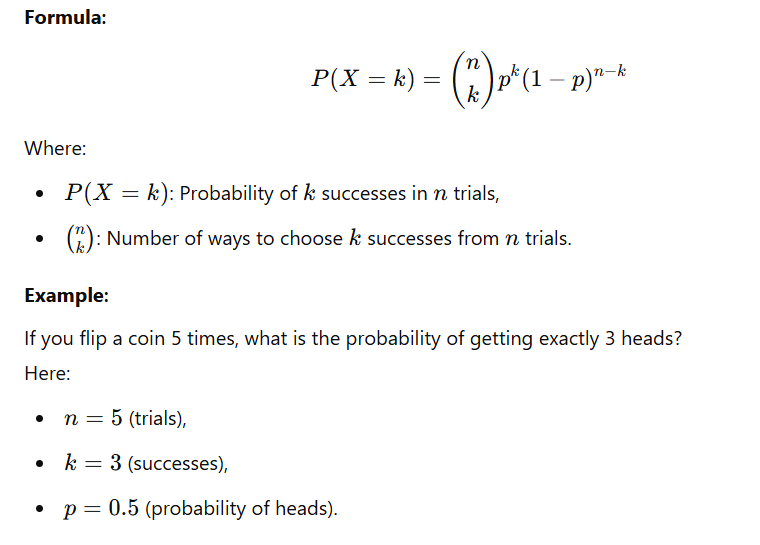


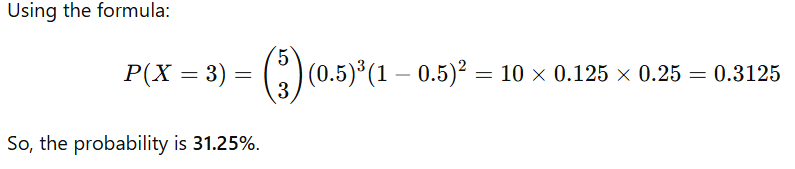
**6. Binomial Distribution**

The **Binomial Distribution** extends the Bernoulli Distribution to **multiple independent trials**. It models the number of successes in nnn trials, where each trial has a probability of success ppp.

**Key Features:**

* The outcomes are binary (success/failure).
* The probability of success remains the same in all trials.
* Total trials = n





**Summary Table**

|  |  |  |
| --- | --- | --- |
| Distribution | Key Features | Example |
| Normal Distribution | Bell-shaped curve, symmetrical, mean = median = mode | Heights of people, IQ scores |
| Standard Normal | Normal distribution with μ=0\mu = 0μ=0, σ=1\sigma = 1σ=1 | Z-scores, standardized test scores |
| Z-Score | Measures how far a data point is from the mean in terms of standard deviations | Outlier detection, comparing test scores |
| Log-Normal Distribution | Skewed, logarithm of data is normally distributed | Stock prices, income distribution |
| Bernoulli Distribution | Single trial, two outcomes (success/failure) | Tossing a coin, passing a test (pass/fail) |
| Binomial Distribution | Number of successes in nnn independent trials | Flipping coins, number of defective items in a batch |

# Hypothesis Testing

**1. What is Hypothesis Testing?**

**Hypothesis Testing** is a statistical method used to:

* Test assumptions or claims about a population parameter (e.g., mean, proportion).
* Make decisions using sample data (evidence) and probability.

**2. Steps in Hypothesis Testing**

1. **State the Hypotheses**: Define the null hypothesis and alternative hypothesis.
2. **Collect Data**: Perform an experiment or gather data.
3. **Perform a Test**: Use statistical tests like Z-test, t-test, etc., to analyze the data.
4. **Draw Conclusions**: Based on the p-value and significance level, reject or accept the null hypothesis.

**3. Null Hypothesis (H₀)**

The **Null Hypothesis** is the default assumption that there is **no effect or difference**.

* It is what you **assume to be true** unless proven otherwise.
* We test this hypothesis and look for evidence to reject it.

**Example:**

A company claims that the average battery life of its smartphone is **10 hours**.

* H0H\_0H0​: The average battery life = 10 hours.

**4. Alternative Hypothesis (H₁ or Hₐ)**

The **Alternative Hypothesis** contradicts the null hypothesis. It states there **is an effect or difference**.

* If we reject the null hypothesis, we **support the alternative hypothesis**.

**Example:**

Using the same battery life example:

* HaH\_aHa​: The average battery life ≠ 10 hours (two-tailed test).  
  OR
* HaH\_aHa​: The average battery life < 10 hours (one-tailed test).

**5. Experiments and Data Collection**

Once hypotheses are stated, data is collected through experiments, surveys, or observations. The data will help determine whether to reject H0H\_0H0​.

**Example:**

Test 50 smartphones and record their battery lives. Use this sample data to perform a statistical test.

**6. P-Value (Probability Value)**

The **p-value** is the probability of obtaining results **as extreme or more extreme** than the observed results, assuming H0H\_0H0​ is true.

* **Low p-value (< significance level)**: Reject H0H\_0H0​. The data supports HaH\_aHa​.
* **High p-value (> significance level)**: Fail to reject H0H\_0H0​. Not enough evidence to support HaH\_aHa​.

**Example:**

* A p-value of **0.03** means there is a 3% chance that the observed data would occur under the null hypothesis.
* If the significance level (α\alphaα) = 0.05: Since 0.03<0.050.03 < 0.050.03<0.05, **reject H0H\_0H0​**.

**7. Significance Level (α)**

The **significance level** (α\alphaα) is the threshold for rejecting H0H\_0H0​. It represents the **risk of Type I error** (false positive).

* Common values: α=0.05\alpha = 0.05α=0.05 (5%) or α=0.01\alpha = 0.01α=0.01 (1%).
* It means you are willing to accept a 5% or 1% chance of incorrectly rejecting H0H\_0H0​.

**8. Reject or Fail to Reject the Null Hypothesis**

* If p-value<αp \text{-value} < \alphap-value<α: **Reject H0H\_0H0​** (evidence supports HaH\_aHa​).
* If p-value>αp \text{-value} > \alphap-value>α: **Fail to reject H0H\_0H0​** (not enough evidence to support HaH\_aHa​).

**9. Confidence Interval**

A **Confidence Interval (CI)** is a range of values where the true population parameter is likely to lie.

* A **95% confidence interval** means: "We are 95% confident that the true parameter lies within this range."

**Example:**

If a 95% CI for the average battery life is [9.5, 10.5] hours:

* The true mean battery life is likely between **9.5 and 10.5 hours**.

**10. Errors in Hypothesis Testing**

There are two types of errors:

1. **Type I Error** (False Positive):
   * Rejecting H0H\_0H0​ when H0H\_0H0​ is true.
   * Controlled by the significance level (α\alphaα).

**Example**: You conclude the battery life ≠ 10 hours, but it actually is 10 hours.

1. **Type II Error** (False Negative):
   * Failing to reject H0H\_0H0​ when HaH\_aHa​ is true.
   * Probability of Type II error = β\betaβ.

**Example**: You fail to prove the battery life is different when it actually is.

**11. One-Tailed and Two-Tailed Tests**

**One-Tailed Test**

* Tests for a **specific direction** (e.g., greater than or less than).
* Example: HaH\_aHa​: Battery life < 10 hours.

**Two-Tailed Test**

* Tests for **any difference** (not direction-specific).
* Example: HaH\_aHa​: Battery life ≠ 10 hours.

**12. Point Estimate**

A **Point Estimate** is a single value used to estimate a population parameter.

**Example:**

* Sample mean (xˉ\bar{x}xˉ) = 9.8 hours estimates the average battery life.

**Summary of Key Steps**

1. **State Hypotheses**:
   * H0H\_0H0​: Average battery life = 10 hours.
   * HaH\_aHa​: Average battery life ≠ 10 hours.
2. **Collect Data**: Conduct an experiment on 50 smartphones.
3. **Choose Significance Level (α\alphaα)**: Use α=0.05\alpha = 0.05α=0.05.
4. **Perform Test**: Calculate the p-value using statistical software or a formula.
5. **Draw Conclusions**:
   * If p<0.05p < 0.05p<0.05, reject H0H\_0H0​.
   * If p>0.05p > 0.05p>0.05, fail to reject H0H\_0H0​.
6. **Calculate Confidence Interval**: Estimate the range where the true mean lies.
7. **Understand Errors**: Type I (false positive) and Type II (false negative).