# **What is statistics and types?**

**Statistics** is a branch of mathematics that involves the collection, analysis, interpretation, presentation, and organization of data

**1. Descriptive Statistics**

Descriptive statistics refers to methods for summarizing and describing the main features of a data set. It provides a way to present large amounts of data in a more understandable and manageable form.

### 2. ****Inferential Statistics****

Inferential statistics involves using data from a sample to make inferences or predictions about a population. It allows for generalizations beyond the observed data, and is often used in hypothesis testing.

# What is Data?

**Data** refers to raw facts, figures, or information that can be collected, analysed and measured.

### **# **Population****

A **population** refers to the entire set of individuals or items that are of interest in a particular study.

### ****# Sample****

A **sample** is a subset of the population, selected for the purpose of analysis.

# **Sampling techniques**

 **Simple Random Sampling**

* **Description**: Every member of the population has an equal chance of being selected. This is one of the most basic and widely used techniques.

**Example**: Selecting 100 students randomly from a university's entire student body.

 **Systematic Sampling**

* **Description**: A sample is selected by choosing every *k*-th individual from the population after selecting a random starting point.
* **Example**: If you're sampling every 10th person from a list of 1000 employees.

 **Stratified Sampling**

* **Description**: The population is divided into homogeneous subgroups (strata) based on certain characteristics, and a random sample is taken from each subgroup.
* **Example**: A survey of a country’s population might divide people by age groups (18-24, 25-34, etc.) and take random samples from each age group.

 **Convenience Sampling**

* **Description**: When survey is conducted with people who has expertise in the particular domain.

**Example**: Survey on drug test

 **Variable**

**variable** is any characteristic, number, or quantity that can be measured or quantified.

### ****1. Qualitative (Categorical) Variables****

Qualitative variables, also called **categorical variables**, represent categories or groups that the data can be divided into. These variables describe qualities or characteristics rather than numerical value

### ****2. Quantitative (Numerical) Variables****

Quantitative variables, also called **numerical variables**, represent quantities and can be measured in numerical terms.

#### Subtypes of Quantitative Variables:

* **Discrete Variables**:
  + These variables represent counts of items or occurrences and take on distinct, separate values (usually integers). There are no values between two consecutive values.
  + **Example**:
    - **Number of children in a family**
    - **Number of cars owned**
    - **Number of books in a shelf**
* **Continuous Variables**:
  + These variables can take on an infinite number of values within a given range and can be measured with great precision. They often represent measurements of quantities that can be broken down into smaller and smaller units.
  + **Example**:
    - **Height** (e.g., 5.5 feet, 5.56 feet, 5.555 feet, etc.)
    - **Weight** (e.g., 150.5 pounds, 150.55 pounds)
    - **Temperature** (e.g., 25.2°C, 25.25°C, etc.)

**Variable measurement scales:**

### ****1. Nominal Scale****

The **nominal scale** is the most basic level of measurement. It is used for labeling variables without any quantitative value. The categories or labels are **mutually exclusive** and **exhaustive**, but there is no inherent order or ranking among them.

#### Examples:

* **Gender** (Male, Female, Non-binary)
* **Nationality** (American, British, Japanese, etc.)
* **Marital Status** (Single, Married, Divorced)

### ****2. Ordinal Scale****

The **ordinal scale** categorizes data into distinct categories with a meaningful order or ranking. However, the **distances** between the categories are **not necessarily equal**, meaning we know the order but not the exact differences between each level.

#### Examples:

* **Education Level** (High School, Bachelor's, Master's, Doctorate)
* **Survey Responses** (Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree)
* **Socioeconomic Status** (Low, Medium, High)

### ****3. Interval Scale****

The **interval scale** measures variables where the distances between values are **equal** and consistent. While the interval scale does not have a true zero point, the differences between measurements are meaningful, allowing for a wide range of statistical operations.

#### Examples:

* **Temperature** (in Celsius or Fahrenheit) — the difference between 10°C and 20°C is the same as between 30°C and 40°C, but 0°C doesn't mean "no temperature."
* **IQ Scores** — the difference between scores is consistent, but an IQ of 0 doesn't mean the absence of intelligence.

### ****4. Ratio Scale****

The **ratio scale** is the highest level of measurement. It has all the properties of the interval scale, but it also has a **true zero point**, which signifies the complete absence of the measured attribute. This allows for meaningful **ratios** between values.

#### Examples:

* **Height** — Zero means no height at all.
* **Weight** — Zero means no weight at all.
* **Age** — Zero means no age (newborn).
* **Income** — Zero means no income at all.

**# Frequency:** in statistics refers to the number of times a particular value or category occurs in a data set.

### ****Types of Frequency****

1. **Absolute Frequency**:
   * The count of how many times a specific value or category appears in the dataset.
   * **Example**: In a survey of 20 people, 12 people like coffee, 5 like tea, and 3 like juice. The absolute frequency for coffee is 12, tea is 5, and juice is 3.
2. **Cumulative Frequency**:
   * The cumulative sum of the absolute frequencies up to and including a certain value. This helps in understanding how data accumulates across categories or values.
   * **Example**: If the number of students in each grade is as follows:
     + Grade A: 10 students
     + Grade B: 15 students
     + Grade C: 25 students
     + Cumulative frequency for Grade A = 10
     + Cumulative frequency for Grade B = 10 + 15 = 25
     + Cumulative frequency for Grade C = 25 + 25 = 50

**Measures of Central Tendency** are statistical tools used to summarize a set of data by identifying a central point within the dataset. These measures give a single value that represents the center or typical value of the data distribution. The three primary measures of central tendency are:

### ****1. Mean (Arithmetic Mean)****

The **mean** is the average of all the values in a dataset. It is the most commonly used measure of central tendency.

#### Formula:

Mean=∑XN\text{Mean} = \frac{\sum X}{N}Mean=N∑X​

#### Example:

Suppose you have the following dataset representing the ages of 5 people: 25, 30, 35, 40, 45.

Mean=25+30+35+40+455=1755=35\text{Mean} = \frac{25 + 30 + 35 + 40 + 45}{5} = \frac{175}{5} = 35Mean=525+30+35+40+45​=5175​=35

So, the **mean** age is 35 years.

### ****2. Median****

The **median** is the middle value in a dataset when the data points are arranged in ascending or descending order. If the dataset has an odd number of values, the median is the middle number. If the dataset has an even number of values, the median is the average of the two middle numbers.

#### Steps to calculate the Median:

1. **Sort the data** in ascending or descending order.
2. If the number of data points is **odd**, the median is the middle value.
3. If the number of data points is **even**, the median is the average of the two middle values.

#### Example:

For the dataset: 25, 30, 35, 40, 45

* The number of data points is odd (5 values).
* The sorted data is already in order: 25, 30, 35, 40, 45.
* The middle value is **35**, so the **median** is 35.

### ****3. Mode****

The **mode** is the value that occurs most frequently in a dataset. It is the only measure of central tendency that can be used with nominal (categorical) data.

#### Example:

For the dataset: 25, 30, 35, 30, 45

* The value **30** appears twice, which is more frequent than any other value.
* So, the **mode** is **30**.

**Measures of Dispersion:** (also called **measures of variability**) are statistical tools that describe the spread or variability of a dataset.

#### Example:

For the dataset: 3, 7, 9, 12, 15

* **Maximum value** = 15
* **Minimum value** = 3

### 2. ****Variance****

Variance measures the **average squared deviation** of each data point from the mean. It tells us how spread out the values in a dataset are around the mean.

#### Formula for Population Variance:

σ2=∑(Xi−μ)2N\sigma^2 = \frac{\sum (X\_i - \mu)^2}{N}σ2=N∑(Xi​−μ)2​

Where:

* XiX\_iXi​ is each data point.
* μ\muμ is the population mean.
* NNN is the number of data points.

#### Formula for Sample Variance:

s2=∑(Xi−Xˉ)2n−1s^2 = \frac{\sum (X\_i - \bar{X})^2}{n - 1}s2=n−1∑(Xi​−Xˉ)2​

Where:

* Xˉ\bar{X}Xˉ is the sample mean.
* nnn is the sample size.

#### Example:

For the dataset: 2, 4, 6, 8, 10

* **Mean** (μ\muμ) = 2+4+6+8+105=6\frac{2 + 4 + 6 + 8 + 10}{5} = 652+4+6+8+10​=6
* Calculate the squared deviations from the mean: (2−6)2=16,(4−6)2=4,(6−6)2=0,(8−6)2=4,(10−6)2=16(2-6)^2 = 16, \quad (4-6)^2 = 4, \quad (6-6)^2 = 0, \quad (8-6)^2 = 4, \quad (10-6)^2 = 16(2−6)2=16,(4−6)2=4,(6−6)2=0,(8−6)2=4,(10−6)2=16
* Sum of squared deviations = 16+4+0+4+16=4016 + 4 + 0 + 4 + 16 = 4016+4+0+4+16=40
* For **population variance**: σ2=405=8\sigma^2 = \frac{40}{5} = 8σ2=540​=8
* For **sample variance**: s2=404=10s^2 = \frac{40}{4} = 10s2=440​=10

### 3. ****Standard Deviation****

The **standard deviation** is the square root of the variance. It provides a measure of spread in the same unit as the original data, making it easier to interpret. It is one of the most commonly used measures of dispersion.

#### Formula for Population Standard Deviation:

σ=∑(Xi−μ)2N\sigma = \sqrt{\frac{\sum (X\_i - \mu)^2}{N}}σ=N∑(Xi​−μ)2​​

#### Formula for Sample Standard Deviation:

s=∑(Xi−Xˉ)2n−1s = \sqrt{\frac{\sum (X\_i - \bar{X})^2}{n - 1}}s=n−1∑(Xi​−Xˉ)2​​

#### Example:

For the dataset: 2, 4, 6, 8, 10 (with sample variance = 10):

* The **standard deviation** = 10≈3.16\sqrt{10} \approx 3.1610​≈3.16

### ****1. Percentage****

**Percentage** is a way of expressing a number as a fraction of 100. It is used to describe how much something is relative to a whole. Percentages are widely used in many fields like finance, education, and health to express proportions or parts of a total.

#### Example:If you scored 45 out of 50 on a test, you can calculate your percentage score as:

Percentage=(4550)×100=90%\text{Percentage} = \left( \frac{45}{50} \right) \times 100 = 90\%Percentage=(5045​)×100=90%

### ****2. Percentile****

**Percentile** is a measure used in statistics to describe the position of a particular value in a dataset. It indicates the percentage of data points that fall below a given value.

#### Formula for Percentile:

Percentiles are calculated based on the rank of a value in a sorted dataset. The general process for finding the percentile rank of a value is:

P=(Number of values below the given valueTotal number of values)×100P = \left( \frac{ \text{Number of values below the given value}}{\text{Total number of values}} \right) \times 100P=(Total number of valuesNumber of values below the given value​)×100

#### Example:

Consider the following dataset of scores: 10,20,30,40,50,60,70,80,90,10010, 20, 30, 40, 50, 60, 70, 80, 90, 10010,20,30,40,50,60,70,80,90,100

* **Step 1**: Arrange the data in ascending order (already done).
* **Step 2**: Find the position of the value you are interested in. For instance, if you are interested in the score **60**, it is the 6th value in this dataset.
* **Step 3**: Calculate the percentile rank:

P=(510)×100=50%P = \left( \frac{5}{10} \right) \times 100 = 50\%P=(105​)×100=50%

* **Interpretation**: The score of 60 is in the **50th percentile**, meaning that 50% of the values in the dataset are below 60.

### ****Five Number Summary****

The **five-number summary** is a set of descriptive statistics that provides a concise overview of a dataset's distribution. It is especially useful for understanding the spread and central tendency of the data, as well as identifying potential outliers.

The five-number summary consists of the following five values:

1. **Minimum**: The smallest value in the dataset.
2. **First Quartile (Q1)**: The median of the lower half of the dataset (25th percentile).
3. **Median (Q2)**: The middle value of the dataset (50th percentile).
4. **Third Quartile (Q3)**: The median of the upper half of the dataset (75th percentile).
5. **Maximum**: The largest value in the dataset.

### ****Outliers****

Outliers are typically identified using the **Interquartile Range (IQR)**:

* **Outlier Rule**: A data point is considered an outlier if it is more than 1.5 times the IQR away from either Q1 or Q3. Lower Bound=Q1−1.5×IQR\text{Lower Bound} = Q1 - 1.5 \times \text{IQR}Lower Bound=Q1−1.5×IQR Upper Bound=Q3+1.5×IQR\text{Upper Bound} = Q3 + 1.5 \times \text{IQR}Upper Bound=Q3+1.5×IQR

### ****Data Distribution****

**Data distribution** refers to how the values in a dataset are spread or distributed across different ranges. It provides insights into the general pattern, spread, and central tendency of the data. Understanding the distribution of a dataset is crucial for selecting the right statistical methods and making inferences about the data.

### ****Types of Data Distributions****

Here are some common types of data distributions:

#### **1. Normal Distribution**

* **Description**: A **normal distribution** is one of the most well-known distributions in statistics. It is symmetric, with the data points clustering around the mean.
* **Characteristics**:
  + Symmetrical: The left and right sides are mirror images.
  + Bell-shaped curve: Most of the data points are near the mean, and fewer data points occur as you move away from the mean.
  + The mean, median, and mode are all equal and located at the center of the distribution.
* **Example**: Heights of people in a population or SAT scores.
* **Key Properties**:
  + **68%** of the data falls within 1 standard deviation of the mean.
  + **95%** falls within 2 standard deviations.
  + **99.7%** falls within 3 standard deviations.

#### **2. Skewed Distribution**

* **Description**: A **skewed distribution** occurs when the data is not symmetrically distributed, meaning it has a "tail" on one side.
* **Types**:
  + **Positively Skewed (Right Skewed)**: The tail of the distribution is on the right side, meaning that the right tail has extreme values (outliers) that are larger than the rest of the data. The mean is greater than the median.
  + **Negatively Skewed (Left Skewed)**: The tail of the distribution is on the left side, meaning that the left tail has extreme values (outliers) that are smaller than the rest of the data. The mean is less than the median.
* **Example**: Income distribution (positive skew) or age at retirement (negative skew).

#### **3. Uniform Distribution**

* **Description**: A **uniform distribution** occurs when all values in a dataset are equally likely. This type of distribution is often called a **rectangular distribution** because it forms a rectangle when plotted.
* **Characteristics**:
  + All values are equally likely to occur.
  + The mean and median are both in the center of the range of values.
* **Example**: Rolling a fair die (each face has an equal probability of appearing).

#### **5. Exponential Distribution**

* **Description**: An **exponential distribution** is used to model the time between events in a Poisson process, where events happen at a constant rate, but are independent of each other.
* **Characteristics**:
  + It has a single peak at zero and then decreases exponentially as you move away from zero.
  + Used for modeling waiting times, such as the time between phone calls at a call center.
* **Example**: The time between arrivals of customers at a store or the lifetime of a light bulb.
* **Z-Score**
* A **Z-score** (also called a **standard score**, **z-value**, or **standardized score**) is a statistical measurement that describes a value's relationship to the mean of a group of values. It tells you how many **standard deviations** a data point is from the mean of the dataset. The Z-score is commonly used to understand if a data point is typical (close to the mean) or unusual (far from the mean).
* **Formula for Z-Score**
* The Z-score for a data point is calculated using the following formula:
* Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ​
* Where:

### ****Common Types of Normalization****

There are several methods of normalization, with the choice depending on the nature of the data and the requirements of the model or analysis. Below are the most common techniques:

#### **1. Min-Max Normalization (Rescaling)**

* **Description**: Min-Max normalization scales the data to a fixed range, usually 0 to 1, by transforming each feature to a specific range.
* **Formula**:

Xnorm=X−min(X)max(X)−min(X)X\_{\text{norm}} = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}Xnorm​=max(X)−min(X)X−min(X)​

#### **1. Binomial Distribution**

* **Description**: The binomial distribution models the number of successes in a fixed number of independent trials, each with two possible outcomes (success or failure). It is used when the trials are independent, and the probability of success remains constant across trials.
* **Example**: If you flip a fair coin 5 times, the probability of getting exactly 3 heads (successes) is given by the binomial distribution

### ****Bernoulli Distribution****

The **Bernoulli distribution** is one of the simplest and most fundamental discrete probability distributions in statistics. It models a random experiment or trial that has only two possible outcomes: a **success** (usually denoted as 1) or a **failure** (denoted as 0). This makes it a special case of the **binomial distribution**, where the number of trials is 1.

### ****Confidence Level and Confidence Interval****

In statistics, **confidence levels** and **confidence intervals** are crucial concepts used to estimate the range of values for an unknown population parameter, based on sample data. These concepts help provide a measure of how reliable an estimate is.

### ****1. Confidence Level****

The **confidence level** is the probability that a confidence interval contains the true population parameter. It represents how certain we are that the interval we have calculated from the sample data contains the actual parameter.

* **Common confidence levels**:
  + **90% Confidence Level**: This means that there is a 90% probability that the confidence interval contains the true population parameter.
  + **95% Confidence Level**: This is the most commonly used level, meaning that 95% of all confidence intervals, generated from random samples of the same size and from the same population, would contain the true parameter.
  + **99% Confidence Level**: This means there is a 99% chance that the confidence interval contains the true population parameter.

A higher confidence level means a wider interval, because we are more confident that the true parameter lies within the interval, but we are also less precise. Conversely, a lower confidence level results in a narrower interval but less confidence that it contains the true parameter.

### ****2. Confidence Interval****

A **confidence interval (CI)** is a range of values, derived from sample data, that is used to estimate an unknown population parameter. It provides an estimated range that is likely to contain the true value of the parameter with a certain level of confidence.

The general formula for a **confidence interval for a population mean** when the population standard deviation is known is:

CI=xˉ±Z(σn)CI = \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right)CI=xˉ±Z(n​σ​)

Where:

* xˉ\bar{x}xˉ = Sample mean
* ZZZ = Z-value corresponding to the desired confidence level (e.g., for 95% confidence, Z≈1.96Z \approx 1.96Z≈1.96)
* σ\sigmaσ = Population standard deviation
* nnn = Sample size

If the population standard deviation is unknown, we use the **t-distribution** instead of the Z-distribution:

CI=xˉ±t(sn)CI = \bar{x} \pm t \left( \frac{s}{\sqrt{n}} \right)CI=xˉ±t(n​s​)

Where:

* ttt = t-value based on the desired confidence level and degrees of freedom (df=n−1df = n - 1df=n−1)
* sss = Sample standard deviation
* nnn = Sample size

### ****Hypothesis Testing****

**Hypothesis testing** is a fundamental concept in statistics used to make inferences or draw conclusions about a population based on sample data. It involves testing an assumption (called a **hypothesis**) about a population parameter, such as a population mean or proportion, and determining whether the sample data provides sufficient evidence to reject that assumption.

### ****Steps in Hypothesis Testing****

1. **State the Null and Alternative Hypotheses**
   * **Null Hypothesis (H₀)**: The default assumption that there is no effect, no difference, or no relationship. It is the hypothesis we seek to test.
   * **Alternative Hypothesis (H₁ or Ha)**: The hypothesis that contradicts the null hypothesis. It represents the effect, difference, or relationship we suspect exists.

**Example**:

* + **Null Hypothesis (H₀)**: The average height of students is 170 cm (μ=170\mu = 170μ=170).
  + **Alternative Hypothesis (H₁)**: The average height of students is not 170 cm (μ≠170\mu \neq 170μ=170).

1. **Choose the Significance Level (α)**
   * The **significance level** (α\alphaα) is the probability of rejecting the null hypothesis when it is true. Common choices for α\alphaα are 0.05, 0.01, and 0.10.
   * **α=0.05\alpha = 0.05α=0.05** means there is a 5% chance of making a Type I error (rejecting a true null hypothesis).
2. **Select the Appropriate Test** Depending on the type of data, sample size, and hypothesis, you choose a statistical test. Common tests include:
   * **t-test**: Used when the sample size is small and the population standard deviation is unknown.
   * **z-test**: Used when the sample size is large, or the population standard deviation is known.
   * **Chi-square test**: Used for categorical data to test the relationship between variables.
   * **ANOVA**: Used to compare means among three or more groups.
3. **Calculate the Test Statistic** The test statistic is a standardized value that measures the degree of deviation of the sample statistic from the population parameter under the null hypothesis.
   * For a **t-test**, the test statistic is calculated as:

t=xˉ−μ0snt = \frac{\bar{x} - \mu\_0}{\frac{s}{\sqrt{n}}}t=n​s​xˉ−μ0​​

Where:

* + - xˉ\bar{x}xˉ = sample mean
    - μ0\mu\_0μ0​ = hypothesized population mean
    - sss = sample standard deviation
    - nnn = sample size
  + For a **z-test**:

z=xˉ−μ0σnz = \frac{\bar{x} - \mu\_0}{\frac{\sigma}{\sqrt{n}}}z=n​σ​xˉ−μ0​​

Where:

* + - σ\sigmaσ = population standard deviation

1. **Determine the p-value** The **p-value** is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.
   * A **low p-value** (typically less than α\alphaα) indicates strong evidence against the null hypothesis, leading to rejection of H0H₀H0​.
   * A **high p-value** indicates weak evidence against H0H₀H0​, so you fail to reject the null hypothesis.
2. **Make a Decision**
   * **Reject the null hypothesis (H₀)** if the p-value is less than the significance level (p<αp < \alphap<α).
   * **Fail to reject the null hypothesis** if the p-value is greater than the significance level (p≥αp \geq \alphap≥α).
3. **Draw a Conclusion** Based on the decision, conclude whether or not there is sufficient evidence to support the alternative hypothesis.

### ****Types of Errors in Hypothesis Testing****

* **Type I Error (False Positive)**: Rejecting a true null hypothesis. This occurs when you conclude that there is an effect or difference when there actually is not.
  + **Example**: Concluding that a drug works when it doesn't.
* **Type II Error (False Negative)**: Failing to reject a false null hypothesis. This occurs when you fail to detect a true effect or difference.
  + **Example**: Concluding that a drug doesn't work when it actually does.