ITypes of Statistics

1.Descriptive
2. Inferential

## 1. Descriptive Statistics

**Descriptive statistics** summarizes and organizes data to describe its main features. It provides a clear picture of the data you have, without making any conclusions or predictions about a larger group. You're just describing what you see. Think of it as a way to simplify a large amount of information into something more understandable.

- **Key Measures:**
    - **Measures of Central Tendency:** These describe the center of the data, like the **mean** (average), **median** (middle value), and **mode** (most frequent value).
    - **Measures of Variability (or Spread):** These describe how spread out the data is, including the **range** (difference between the highest and lowest values), **variance**, and **standard deviation**.
    - **Frequency Distribution:** This shows how often each value or range of values appears in the data, often presented in tables or graphs like histograms.
- **Example:** Imagine you have the test scores for all 30 students in a class. Using descriptive statistics, you could calculate the average score (the mean), find the middle score (the median), and see how much the scores vary (the standard deviation). This tells you about the performance of *this specific group* of 30 students.

## 2. Inferential Statistics

**Inferential statistics** uses data from a small sample to make predictions, inferences, or generalizations about a larger population. The goal is to move beyond the immediate data and draw conclusions that apply to a wider group. This type of statistics relies on probability theory to account for the uncertainty that comes with using a sample.

- **Key Techniques:**
    - **Hypothesis Testing:** This is a formal process for using sample data to test a claim or hypothesis about a population. For example, a t-test or ANOVA test.
    - **Estimation:** This involves using sample data to estimate a population parameter, usually with a **confidence interval**, which provides a range of values where the true population parameter is likely to be found.
    - **Regression Analysis:** This technique helps in understanding the relationship between variables to make predictions.
- **Example:** Taking the previous example, you might take a random sample of 30 students' test scores from a high school and use inferential statistics to estimate the average test score for *all students* in the entire school. You'd calculate a confidence interval to express how certain you are about this estimate.

## The Key Difference

The fundamental difference lies in their purpose:

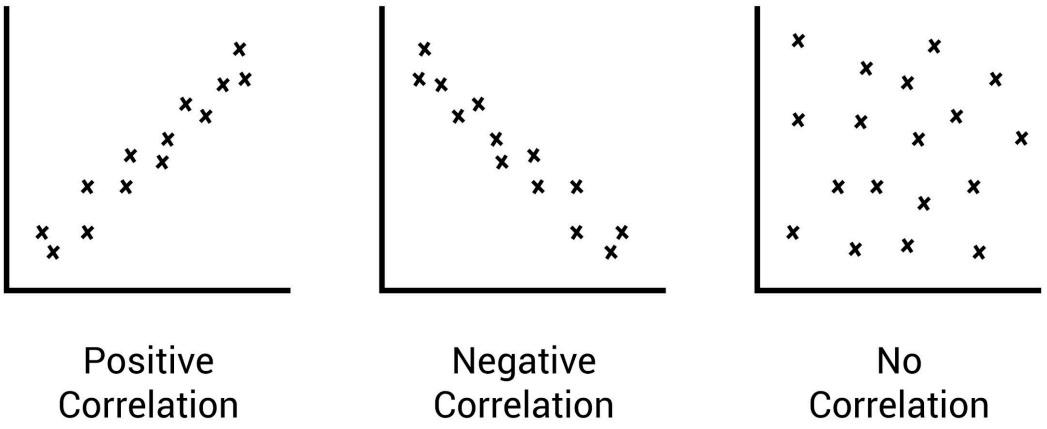
- **Descriptive statistics describes** a known dataset. It's about summarizing and presenting facts.
- **Inferential statistics infers** or predicts from a sample to a larger population. It's about making educated guesses and drawing conclusions.

## 3.Correlation

Correlation indicates that two variables change together. When one variable increases or decreases, the other variable tends to do the same, either in the same or opposite direction. It's a relationship, not necessarily a cause-and-effect link.

- **Positive Correlation:** As one variable increases, the other also increases. For example, the more hours you study, the higher your test scores tend to be.
- **Negative Correlation:** As one variable increases, the other decreases. For example, the more you exercise, the less you may weigh.
- **No Correlation:** There is no clear relationship between the two variables.

A correlation coefficient (often represented as 'r') is a number between -1 and +1 that quantifies the strength and direction of the relationship. An 'r' value of +1 is a perfect positive correlation, -1 is a perfect negative correlation, and 0 means no linear correlation.




## Causation

Causation means that a change in one variable is the direct cause of a change in another variable. To establish causation, you must demonstrate that:

1. **Temporal Precedence:** The cause must occur *before* the effect.
2. **Covariation:** The two variables must be correlated.
3. **No Confounding Variables:** There is no other plausible explanation for the effect.

The most reliable way to prove causation is through a controlled experiment, where you can manipulate one variable and observe the effect on another while holding all other factors constant.

Key terms in statistics

1,population

2, sample

Sampling distribution

- Probability distribution
- Area under the curve
- Guassian distribution/normal Distribution
- Sampling distribution of mean
- Central limit theory

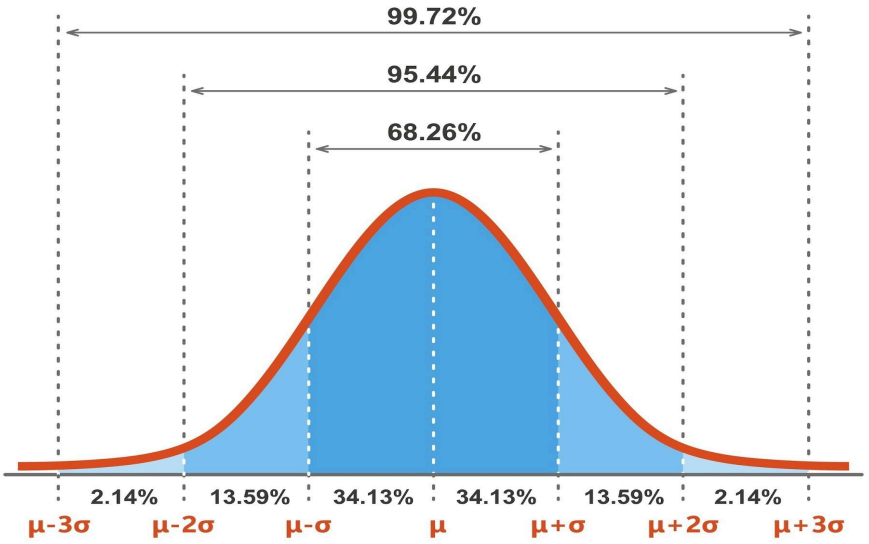
## 4,probability Distribution

A **probability distribution** describes all the possible values a random variable can take and how often each value is expected to occur. It's a foundational concept that can represent a population. For a continuous variable, this is shown as a curve where the height of the curve indicates the likelihood of a value. The total area under the curve is always equal to 1, or 100%, because it represents the probability of all possible outcomes.

## 5.Gaussian Distribution / Normal Distribution

The **Gaussian distribution**, more commonly known as the **normal distribution**, is a specific type of probability distribution. It's a symmetrical, bell-shaped curve where the majority of values cluster around a central mean, with values tapering off equally in both directions. Many natural phenomena, like height, blood pressure, and test scores, follow a normal distribution. It's defined by its mean (μ) and standard deviation (σ).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$




## 5. Area Under the Curve

For a probability distribution, the **area under the curve** represents the probability of an outcome occurring within a certain range. For a normal distribution, the total area is 1, and specific percentages of the data fall within certain standard deviations from the mean. For example, about 68% of the data lies within one standard deviation, and about 95% lies within two standard deviations. Calculating the area under a specific part of the curve allows you to find the probability of a value falling in that range.

## 6 Sampling Distribution

A **sampling distribution** is a specific type of probability distribution. Instead of showing the distribution of individual data points, it shows the distribution of a **statistic** (like the mean) calculated from many random samples of the same size taken from a population. For example, if you repeatedly take samples of size *n* from a population and calculate the mean for each sample, the distribution of all those sample means is the sampling distribution of the mean.

## 7.Sampling Distribution of the Mean

This is the most common type of sampling distribution. It is the probability distribution of the **sample means** that you get from drawing an infinite number of samples of a specific size from a population. This distribution has a few important properties:

- Its mean ($\mu_{\bar{x}}$) is equal to the population mean ($\mu$).
- Its standard deviation, known as the **standard error of the mean**, is equal to the population standard deviation ($\sigma$) divided by the square root of the sample size ( n):

## 8. Central Limit Theorem (CLT)

The **Central Limit Theorem** is one of the most important concepts in statistics. It states that if you take a large enough sample size (typically n≥30), the sampling distribution of the mean will be **approximately normal**, regardless of the shape of the original population distribution. This allows us to use the properties of the normal distribution to make inferences and predictions about the population even when we don't know its original distribution

**9. Hypothesis testing**

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It's a way to determine if there's enough evidence to reject a particular assumption or hypothesis about a population parameter. This process helps researchers and analysts determine if an observed pattern or difference in data is statistically significant or if it could have occurred by chance.

## The Core Concepts of Hypothesis Testing

- **Null Hypothesis (H0)**: This is the initial assumption, a statement of no effect, no difference, or no relationship. It represents the status quo and is what you are trying to disprove. For example, H0 might state that the average height of men is equal to the average height of women.
- **Alternative Hypothesis (Ha or H1):** This is the claim you are trying to prove. It's the opposite of the null hypothesis. It suggests that there is a significant effect, difference, or relationship. Using the previous example, Ha would be that the average height of men is not equal to the average height of women. The alternative hypothesis can be one-tailed (directional, e.g., men are taller than women) or two-tailed (non-directional, e.g., men and women have different heights).
- **Significance Level (α)**: This is a predetermined threshold for rejecting the null hypothesis. It represents the maximum probability of making a Type I error. Common values are 0.05 (5%) and 0.01 (1%). A smaller α means you require stronger evidence to reject H0.
- **Test Statistic**: A value calculated from sample data that is used to evaluate the null hypothesis. The choice of test statistic depends on the type of data and the hypothesis being tested (e.g., z-score, t-score, F-statistic).
- **P-value:** The probability of observing a test statistic as extreme as, or more extreme than, the one calculated from your sample, assuming the null hypothesis is true. A small p-value (typically less than α) suggests that your observed data is unlikely under the null hypothesis, providing evidence to reject H0.

## The Hypothesis Testing Procedure

The process of hypothesis testing is a structured, multi-step process:

1. State the Hypotheses: Clearly define both the null (H0) and alternative (Ha) hypotheses.
2. Set the Significance Level (α): Decide on the probability of a Type I error you are willing to accept before conducting the test.
3. Choose the Appropriate Test Statistic: Select a statistical test (e.g., t-test, z-test, ANOVA) based on the data type, sample size, and the nature of the hypothesis.
4. Collect and Analyze Data: Gather your sample data and use it to calculate the test statistic and the p-value.

5. Make a Decision: Compare the p-value to the significance level (α).
    - If p-value ≤α, you reject the null hypothesis. This means your results are statistically significant, and you have enough evidence to support the alternative hypothesis.
    - If p-value $ > \alpha$, you fail to reject the null hypothesis. This means you don't have enough evidence to conclude that a significant difference or relationship exists.
6. Draw a Conclusion: State your findings in the context of the original problem. For instance, "Based on the data, we have sufficient evidence to conclude that the new medicine is more effective."

### Types of Errors

In hypothesis testing, there's always a risk of making an incorrect decision. There are two types of errors you can make:

- Type I Error (False Positive): Rejecting the null hypothesis when it is actually true. The probability of this error is equal to the significance level, α.
- Type II Error (False Negative): Failing to reject the null hypothesis when it is actually false. The probability of this error is denoted by β.

## Major statistical tests

A Z-test is a statistical hypothesis test used to determine if there's a significant difference between a sample mean and a population mean, or between the means of two different samples. It's a key tool in inferential statistics, allowing you to make conclusions about a larger population based on data from a smaller sample.

### When to Use a Z-Test

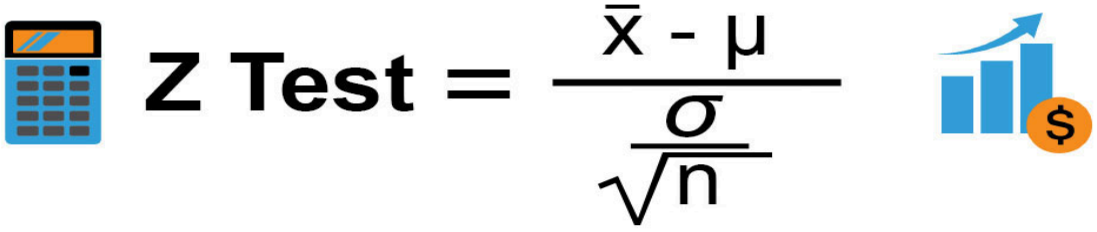
You should use a Z-test when your data meets two critical conditions:

1. **Known Population Standard Deviation (σ)**: The Z-test assumes you know the true standard deviation of the population. This is a crucial distinction from the t-test, which is used when the population standard deviation is unknown.
2. **Large Sample Size (n≥30)**: The Z-test relies on the Central Limit Theorem, which states that for a large enough sample size (typically 30 or more), the distribution of sample means will be approximately normal, regardless of the population's original distribution. This allows you to use the standard normal (Z) distribution for your calculations.

### The Z-Test Formula and Process

The formula for a one-sample Z-test is:

# Z Test Statistics Formula




Where:

- $\bar{x}$ is the sample mean.
- μ is the population mean (as stated in the null hypothesis).
- σ is the known population standard deviation.
- n is the sample size.

**The Z-test procedure follows these steps:**

1. **State the Hypotheses**: Define the **null hypothesis** (H0) and the **alternative hypothesis** (Ha). H0 usually states there is no difference, while Ha suggests there is a difference.
2. **Set the Significance Level (α)**: Choose your desired significance level, which is the probability of a Type I error (rejecting a true null hypothesis). Common values are 0.05 or 0.01.
3. **Calculate the Z-Statistic**: Use the formula above to calculate the Z-score for your sample data. The Z-statistic tells you how many standard deviations your sample mean is from the hypothesized population mean.
4. **Find the P-value or Critical Value**: Use a Z-table or statistical software to find the p-value associated with your calculated Z-statistic. Alternatively, you can find the critical Z-value for your chosen α level.
5. **Make a Decision**:
    - **P-value method**: If the p-value is less than or equal to your significance level (α), you **reject the null hypothesis**.
    - **Critical value method**: If your calculated Z-statistic falls into the critical region (the rejection zone), you **reject the null hypothesis**.
6. **Conclude**: State your conclusion in the context of the problem. For example, "The data provides statistically significant evidence that the new training method improves performance."

# One tail and two tail test

One-tailed and two-tailed tests are two types of **alternative hypotheses** in statistical testing that determine the directionality of the test. The choice between them affects where you look for evidence to reject the null hypothesis and how the significance level (α) is distributed.

**One-Tailed Test (Directional)**
A one-tailed test is used when you have a specific, directional hypothesis about the population parameter. You're only interested in seeing if there's a difference in one specific direction (e.g., greater than or less than).

- **Hypotheses**: The alternative hypothesis (Ha) uses a directional sign, either ">" (greater than) or "<" (less than).
    - **Right-tailed test**: Ha:μ>100 (e.g., a new drug **increases** a patient's heart rate).
    - **Left-tailed test**: Ha:μ<100 (e.g., a new diet plan **reduces** average weight).
- **Rejection Region**: The entire significance level (α) is placed in **one tail** of the distribution. For a right-tailed test, the rejection region is in the far right; for a left-tailed test, it's in the far left. This makes it "easier" to reject the null hypothesis if the result falls in the predicted direction.
- **Use Case**: You use a one-tailed test when you have strong prior knowledge or a theory that makes a specific directional prediction. For example, if you're testing a new fertilizer and you only care if it increases crop yield, not if it decreases it.

**Two-Tailed Test (Non-Directional)**
A two-tailed test is used when you want to know if there's a difference, but you don't have a specific prediction about the direction of that difference. You're testing for the possibility of a significant effect in **either** direction.

- **Hypotheses**: The alternative hypothesis (Ha) uses a "not equal to" sign (=).
    - Ha:μ=100 (e.g., a new training program **changes** the average score, but you don't know if it will increase or decrease it).
- **Rejection Region**: The significance level (α) is split between **both tails** of the distribution. For a common α=0.05, each tail would have a rejection region of 0.025. This makes the test more conservative, as your test statistic needs to be more extreme to fall into one of the smaller rejection regions.
- **Use Case**: This is the default or most common choice when there is no prior knowledge or reason to assume a specific direction. For example, if you are comparing two groups (e.g., men and women) on a particular

measure, you would use a two-tailed test to see if there's any difference between their means.

**Key Differences at a Glance**

| Feature | One-Tailed Test | Two-Tailed Test |
|---|---|---|
| **Hypothesis (Ha)** | Directional: > or < | Non-directional: ≠ |
| **Rejection Region** | Concentrated in one tail | Split between two tails |
| **P-value** | The probability of getting a result in one direction. | The probability of getting a result in either direction. |
| **Sensitivity** | Higher power to detect an effect in the predicted direction. | More conservative; detects a difference in either direction. |
| **When to Use** | When you have a clear, a priori directional prediction. | When you are testing for any difference, regardless of direction. |

## T-test

A **t-test** is a statistical hypothesis test used to compare the means of two groups. It's particularly useful for determining if the difference between those means is statistically significant or if it could have occurred by chance. The t-test is the go-to test when you have a **small sample size** (typically less than 30) and the **population standard deviation is unknown**.

**T-Test vs. Z-Test**

The key difference between a t-test and a z-test lies in what you know about the population.

- **Z-Test**: Used when the **population standard deviation (σ) is known** and/or you have a **large sample size** (n≥30).
- **T-Test**: Used when the **population standard deviation is unknown** and you must estimate it using the sample standard deviation (s). The t-test uses the **t-distribution**, which has "fatter tails" than the normal distribution to account for the increased uncertainty that comes with having a smaller sample.

As the sample size increases, the t-distribution becomes more like the normal distribution, and the results of a t-test and a z-test will converge.

**Types of T-Tests**
There are three main types of t-tests, each used for a specific scenario:

1. **One-Sample T-Test**: Compares the mean of a **single sample** to a known value or a hypothesized population mean.
    - **Example**: You want to know if the average height of students in your class is significantly different from the national average height of 5'8".
2. **Independent Samples T-Test** (or Two-Sample T-Test): Compares the means of **two independent, unrelated groups** to see if there's a significant difference between them.
    - **Example**: You want to know if a new teaching method (Group A) results in higher test scores than the old method (Group B).
3. **Paired Samples T-Test**: Compares the means from the **same group at different times** or under different conditions. The samples are dependent or "paired" in some way.
    - **Example**: You want to measure if a new workout program has a significant effect on participants' weight by comparing their weight **before** the program to their weight **after** the program.
    -

**T-Test Assumptions**

For a t-test to be valid, your data should meet these assumptions:

- **Continuous or Ordinal Data**: The data you're analyzing should be on a continuous scale (e.g., height, weight) or an ordinal scale (e.g., ratings).
- **Random Sampling**: The data should be collected from a random sample of the population.
- **Approximate Normal Distribution**: The data should be approximately normally distributed. This is especially important for very small sample sizes.
- **Homogeneity of Variance** (for Independent Samples T-Test): The variances of the two groups being compared should be roughly equal.