SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT III –Standard Error of the Mean

EMPATHY:

Standard Error of the Mean

- The concept of standard error of the mean is a foundation for statistical and data analysis across mathematics, science, and real-life applications.
- Understanding SEM equips students to interpret sample data, estimate population averages, and tackle common exam questions with clarity.

DEFINE:

What Is Standard Error of the Mean?

- The standard error of the mean (SEM) is a statistical measure that describes how much the sample mean is likely to deviate from the actual population mean.
- It’s commonly used in statistics, data analysis, and experiment results to assess the reliability of sample averages.

Key Formula for Standard Error of the Mean

Here's the standard formula for the standard error of the mean:

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}}$$

Where:

SD = standard deviation of the sample
n = sample size (number of data points)

IDEATE:

Step-by-Step Illustration

Let’s solve a standard error of the mean problem stepwise:

1. List your sample data(xi): 5, 10, 15, 20, 25

2. Calculate the mean($\bar{x}$): (5 + 10 + 15 + 20 + 25) / 5 = 15

3. Compute each data point’s difference from mean, then square:
$(5-15)^2 = 100$, $(10-15)^2 = 25$, $(15-15)^2 = 0$, $(20-15)^2 = 25$, $(25-15)^2 = 100$

4. Add squared differences: 100 + 25 + 0 + 25 + 100 = 250

5. Find sample variance: 250 / (5−1) = 62.5

6. Take square root for SD: √62.5 ≈ 7.91

7. Find SEM: SEM = 7.91 / √5 ≈ 7.91 / 2.24 = 3.53

8. Final answer: The standard error of the mean is 3.53 for this sample.

Standard Error vs Standard Deviation

PROTOTYPING:

Feature	Standard Deviation (SD)	Standard Error of the Mean (SEM)
Meaning	Spread of all data points around the mean	How far the sample mean is from the population mean
Formula	$SD = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$	$SEM = \frac{SD}{\sqrt{n}}$
Affected by sample size?	No	Yes (SEM decreases as n increases)
Use in Exams	Measure data variability	Judge reliability of sample mean

Example:

Suppose you take five readings: 20, 22, 24, 26, 28.

Answer:

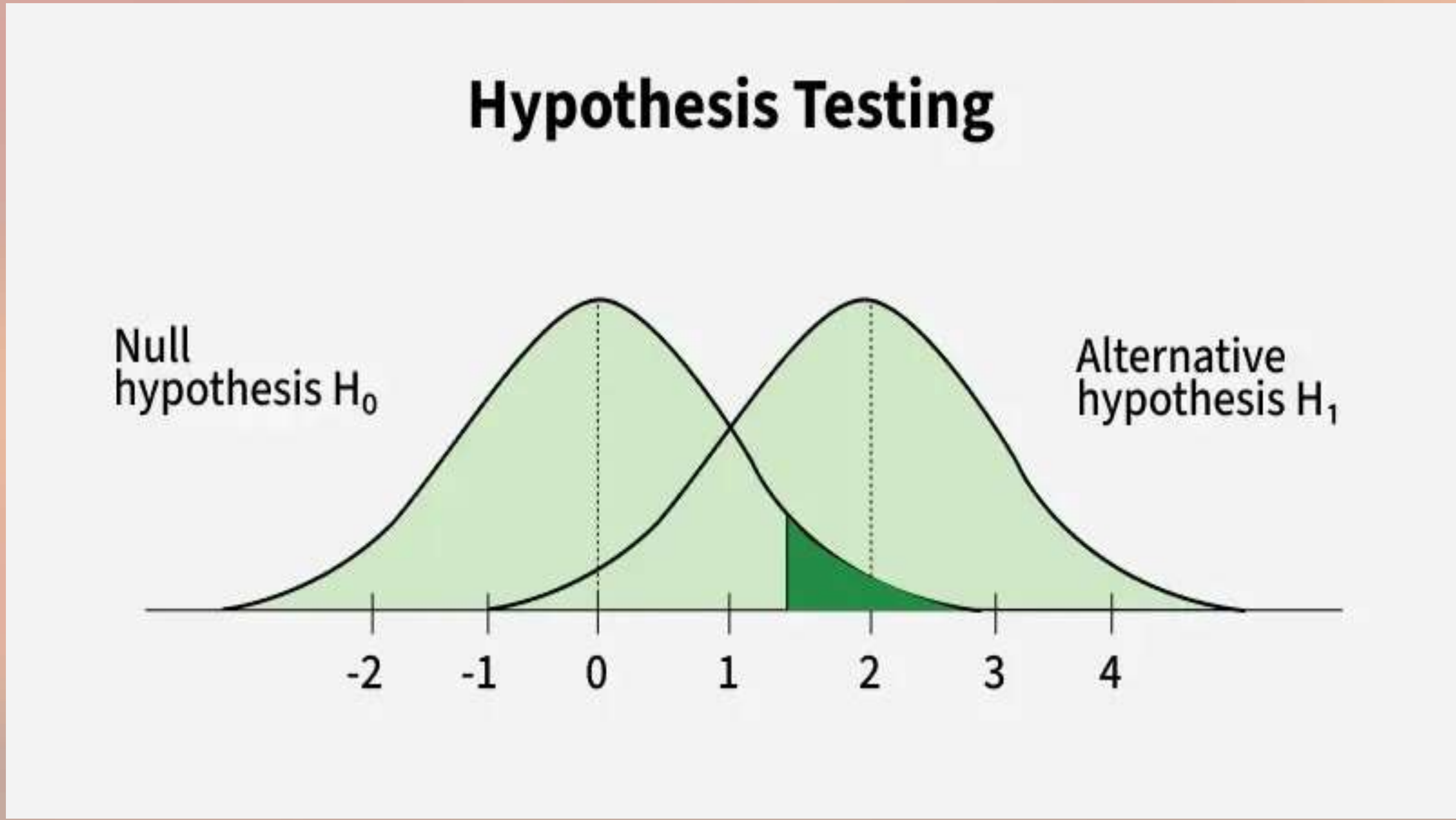
1. Mean = (20+22+24+26+28)/5 = 24
2. $(20-24)^2 = 16$, $(22-24)^2 = 4$, $(24-24)^2 = 0$, $(26-24)^2 = 4$, $(28-24)^2 = 16$
3. Sum of squares = 16 + 4 + 0 + 4 + 16 = 40
4. Variance = 40 / (5-1) = 10
5. SD = $\sqrt{10} \approx 3.16$
6. SEM = $3.16 / \sqrt{5} \approx 1.41$

TESTING:

Hypothesis Testing

- Hypothesis testing compares two opposite ideas about a group of people or things and uses data from a small part of that group (a sample) to decide which idea is more likely true.

- We collect and study the sample data to check if the claim is correct.



For example, if a company says its website gets 50 visitors each day on average, we use hypothesis testing to look at past visitor data and see if this claim is true or if the actual number is different.

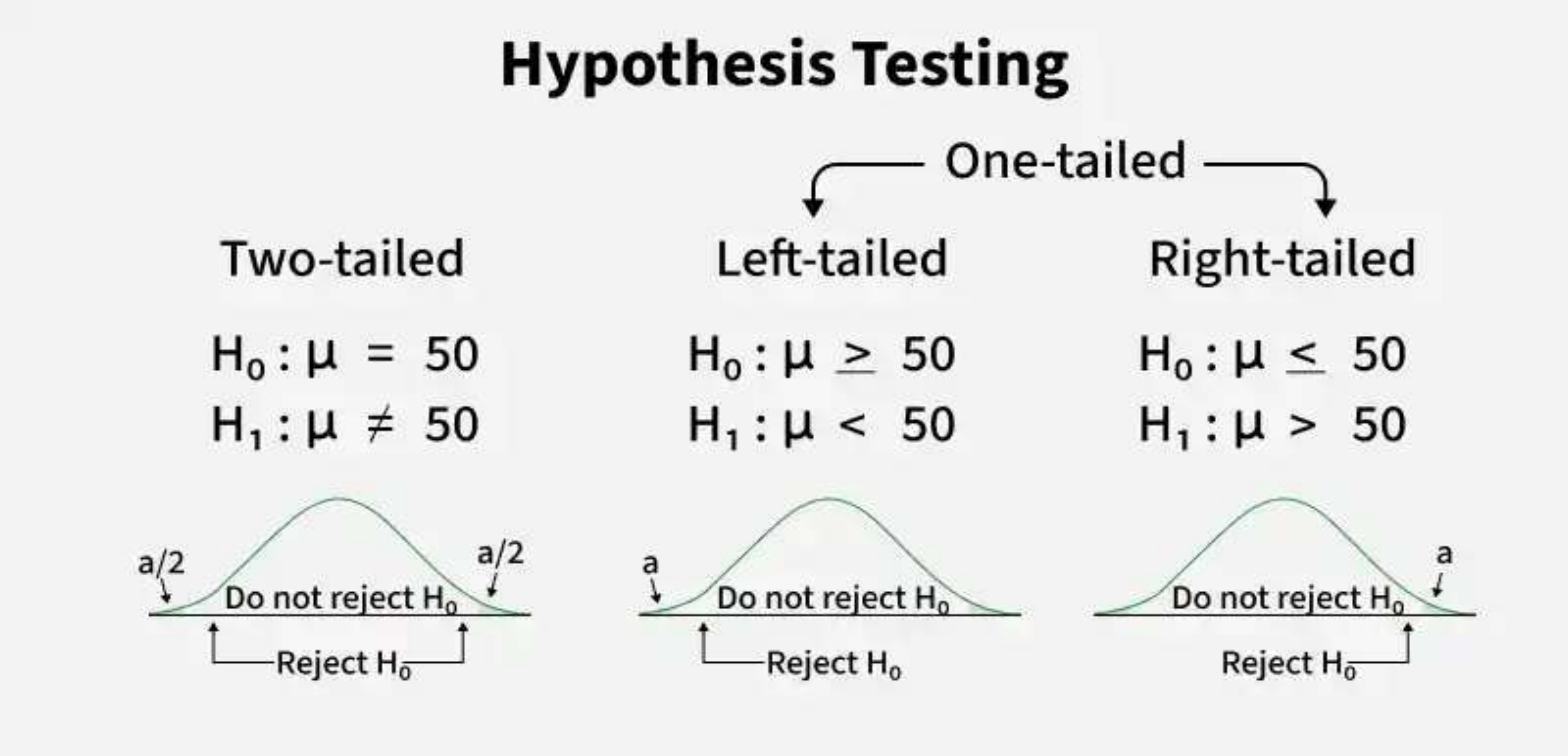
Defining Hypotheses:

Null Hypothesis (H_0): The starting assumption. For example, "The average visits are 50.“

Alternative Hypothesis (H_1): The opposite, saying there is a difference. For example, "The average visits are not 50."

Types of Hypothesis Testing

It involves basically two types of testing:



1. One-Tailed Test

Used when we expect a change in only one direction either up or down, but not both.

For example, if testing whether a new algorithm improves accuracy, we only check if accuracy increases.

2. Two-Tailed Test

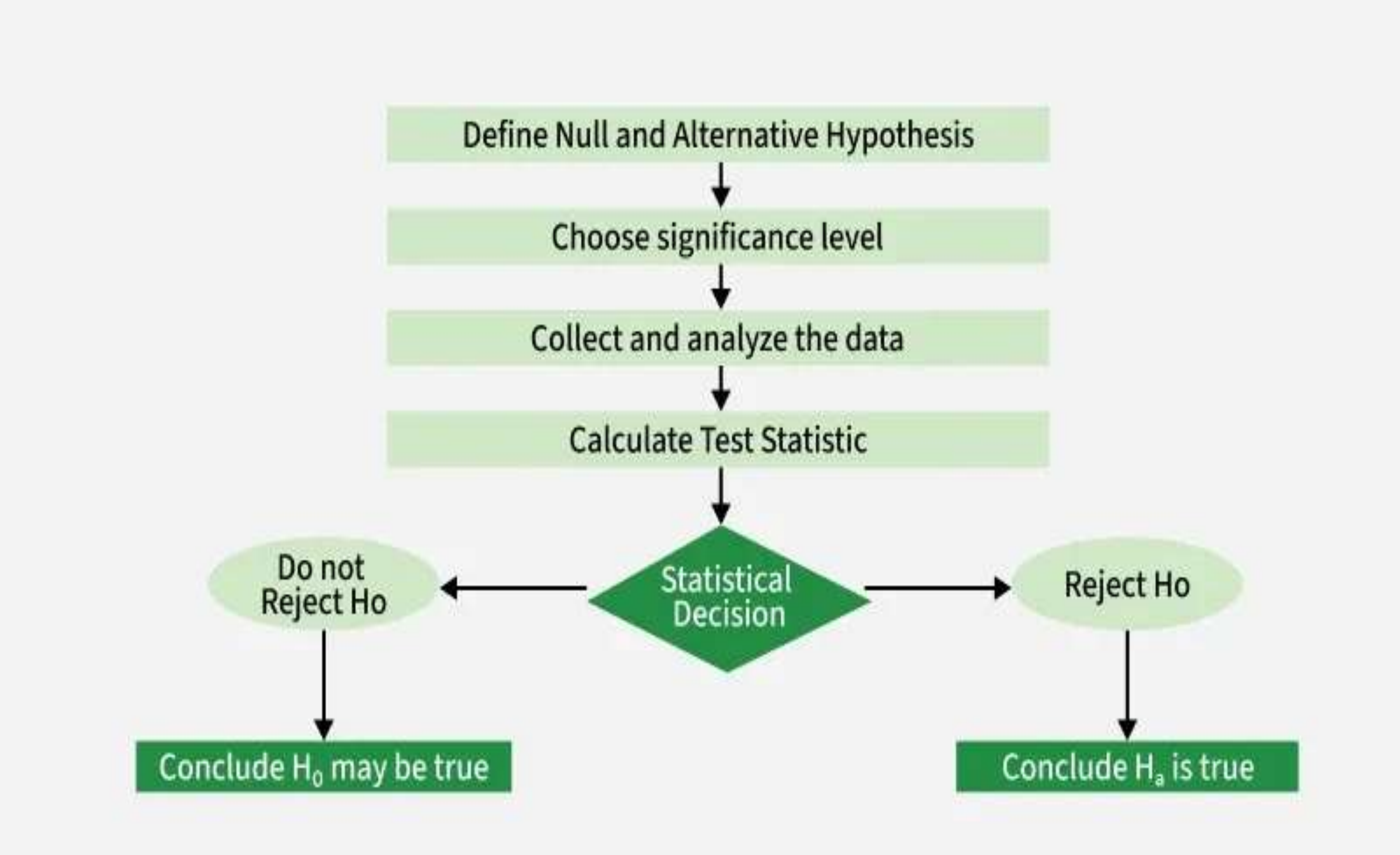
Used when we want to see if there is a difference in either direction higher or lower.

For example, testing if a marketing strategy affects sales, whether it goes up or down

What are Type 1 and Type 2 errors in Hypothesis Testing?

- **Type I error:** When we reject the null hypothesis although that hypothesis was true. Type I error is denoted by alpha(α).
- **Type II errors:** When we accept the null hypothesis but it is false. Type II errors are denoted by beta(β).

How does Hypothesis Testing work?



Step 1: Define Hypotheses:

- **Null hypothesis (H_0):** Assumes no effect or difference.
- **Alternative hypothesis (H_1):** Assumes there is an effect or difference.

Example: Test if a new algorithm improves user engagement.

***Note**: In this we assume that our data is **normally distributed**.*

Step 2: Choose significance level

We select a significance level (usually 0.05). This is the maximum chance we accept of wrongly rejecting the null hypothesis (Type I error). It also sets the confidence needed to accept results.

Step 3: Collect and Analyze data.

•Now we gather data this could come from user observations or an experiment. Once collected we analyze the data using appropriate statistical methods to calculate the test statistic.

Step 4: Calculate Test Statistic

The test statistic measures how much the sample data deviates from what we did expect if the null hypothesis were true. Different tests use different statistics:

Z-test: Used when population variance is known and sample size is large.

T-test: Used when sample size is small or population variance unknown.

Chi-square test: Used for categorical data to compare observed vs. expected counts.

Step 5: Make a Decision

We compare the test statistic to a critical value from a statistical table or use the p-value:

1. Using Critical Value:

If test statistic > critical value → reject H0.

If test statistic ≤ critical value → fail to reject H0.

2. Using P-value:

If p-value ≤ α → reject H0.

If p-value > α → fail to reject H0.

Example: If p-value is 0.03 and α is 0.05, we reject the null hypothesis because 0.03 < 0.05.

Step 6: Interpret the Results

Based on the decision, we conclude whether there is enough evidence to support the alternative hypothesis or if we should keep the null hypothesis.