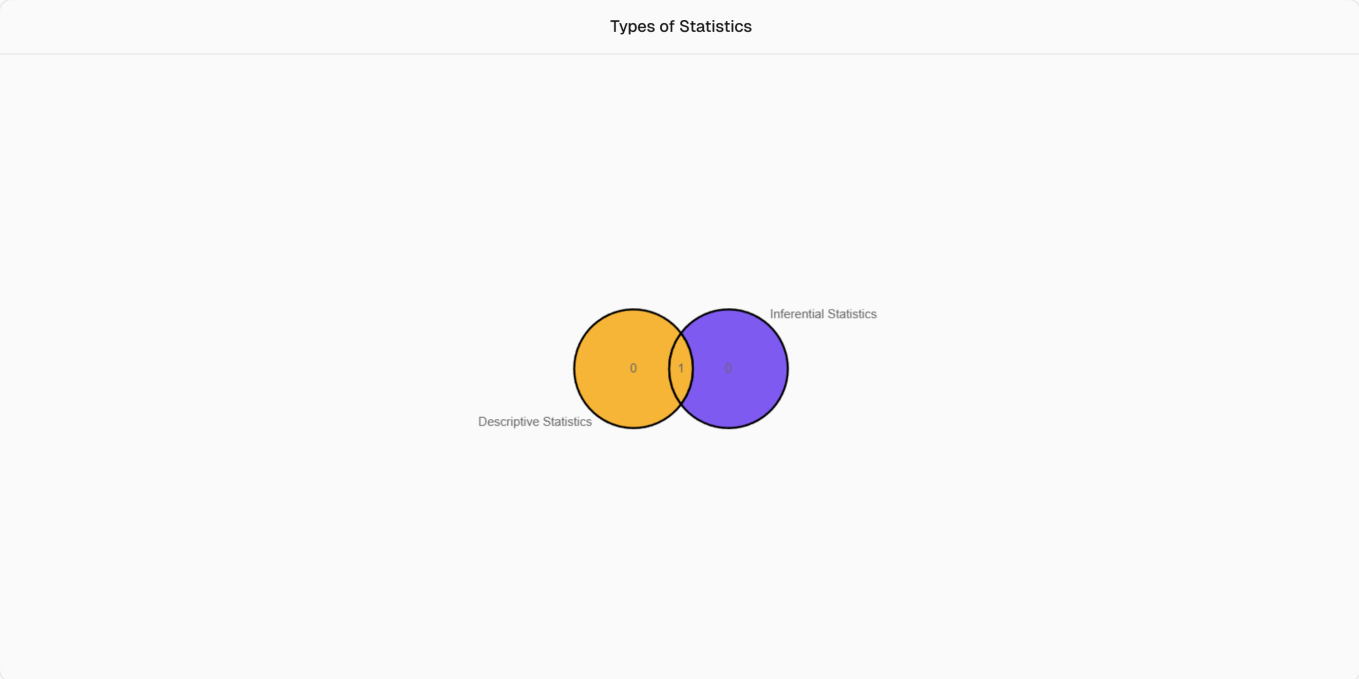
Statistics Roadmap Overview

- This roadmap outlines how to prepare for statistics, covering the entire syllabus for data analysts and data scientists.
- The playlist will include approximately 30-40 videos and will be entirely in Hindi.
- This learning path is beneficial for aspiring data analysts, data scientists, and Business Intelligence (BI) developers.

Types of Statistics

- Statistics is broadly divided into two main fields: **Descriptive Statistics** and **Inferential Statistics**.
- Data analysts primarily work with both types of statistics.



Feature	Descriptive Statistics	Inferential Statistics
Main Goal	Summarize and visualize data	Draw conclusions about a population from sample data
Focus	Describing characteristics of the observed data	Making predictions or inferences about a larger group
Key Activities	Calculating measures of central tendency/dispersion, creating charts	Hypothesis testing, estimation, prediction
Example Topics	Mean, Median, Mode, Variance, Histograms, Box Plots	Z-test, T-test, Chi-square, ANOVA, Hypothesis Testing

Descriptive Statistics

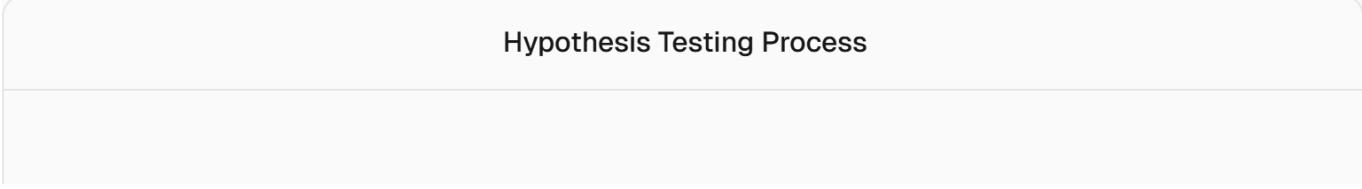
- **Descriptive Statistics** focuses on summarizing and visualizing data.

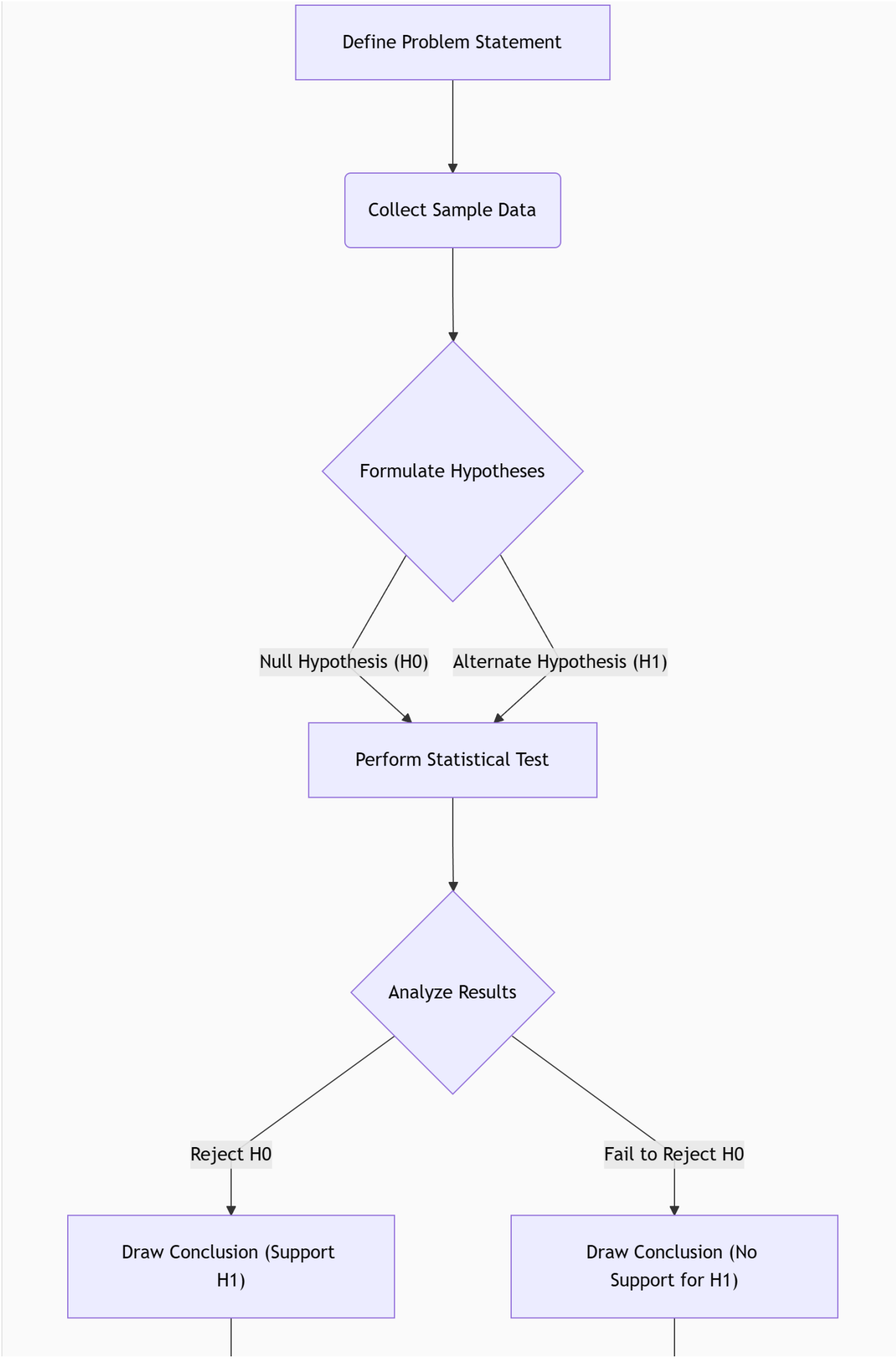
- Key topics include:
    - **Measures of Central Tendency**: Mean, Median, and Mode.
    - **Measures of Dispersion**: Variance and Standard Deviation.
    - **Correlation**: Understanding how one data point relates to another, including Spearman and Pearson correlation, and Covariance.
- Common visualizations and functions covered are:
    - **Histograms** and **Probability Density Function (PDF)**.
    - **Probability Mass Functions (PMF)** and **Cumulative Distribution Functions (CDF)**.
    - **Kernel Density Estimator**.
    - **Box Plots** (also known as Whisker Plots) are specifically used for finding outliers.
    - Other plots like bar graphs and scatter plots are also used, depending on variables and data types.
- It also covers **Univariate** and **Bivariate analysis** to analyze single or two features and their relationships.

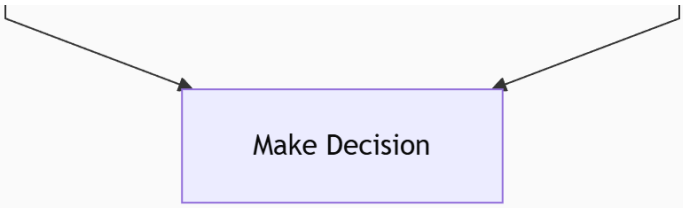
Inferential Statistics

- **Inferential Statistics** involves drawing conclusions about a population based on sample data through experiments and tests.
- A crucial topic in inferential statistics is **Hypothesis Testing**.
    - It involves defining a **Null Hypothesis** and an **Alternate Hypothesis**.
    - This process helps data scientists predict outcomes and make conclusions about population datasets.
- Examples of statistical tests used in inferential statistics include:
    - **Z-test** and **T-test**.
    - **Chi-square test**.
    - **ANOVA** (Analysis of Variance), also known as **F-test**.
- For instance, a data analyst might use hypothesis testing to determine if a new ATM should be opened in a specific location by analyzing transaction data from nearby ATMs.

Here is a flowchart illustrating the general process of hypothesis testing:




Define Problem Statement
Collect Sample Data
Formulate Hypotheses
Null Hypothesis (H0)
Alternate Hypothesis (H1)
Perform Statistical Test
Analyze Results
Reject H0
Fail to Reject H0
Draw Conclusion (Support H1)
Draw Conclusion (No Support for H1)



Importance of Statistics for Decision Making

- Statistical tools are essential for understanding data and making informed decisions.
-

> **Key Insight:** "Data never lies." If something is present in the data, it will be revealed through these tools.

- Sufficient data and comprehensive analysis enhance decision-making skills.

Related Business Intelligence Tools

- The roadmap also touches upon Business Intelligence (BI) tools like **Power BI** and **Tableau**.
- These tools are used to find Key Performance Indicators (KPIs) and create reports and various visualizations.

Day - 2

Introduction to Statistics

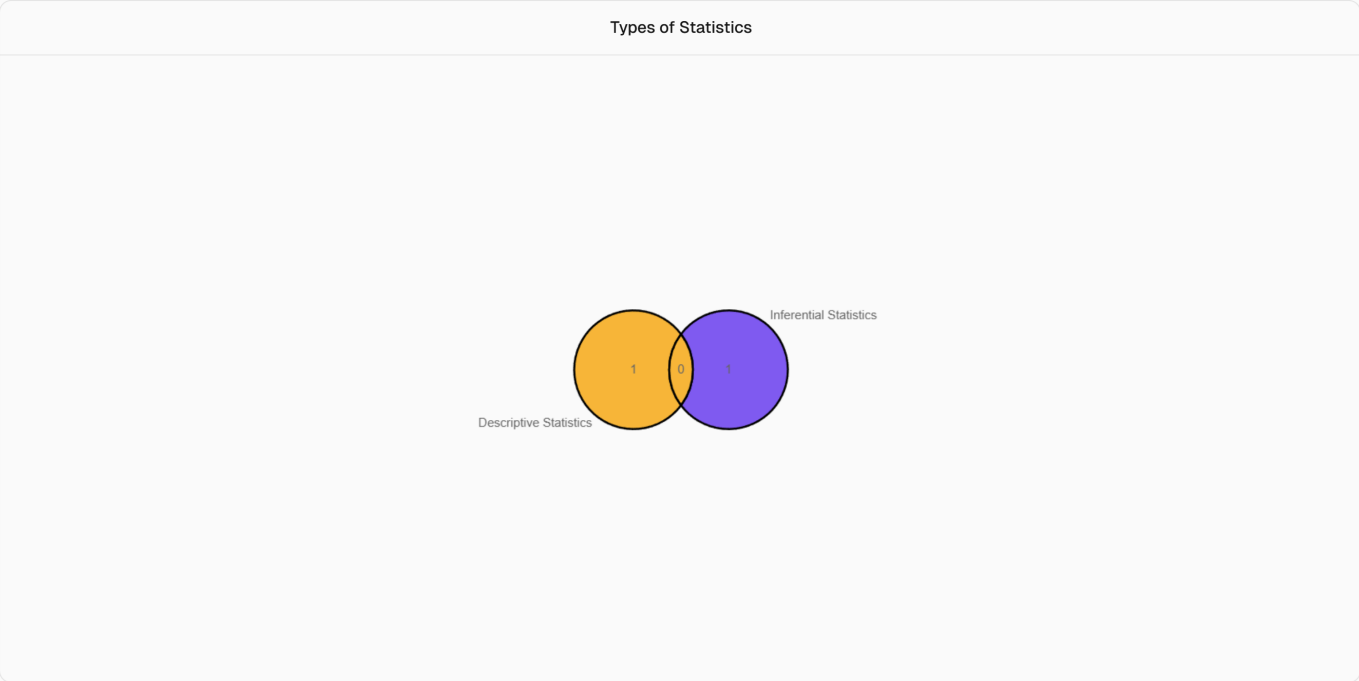
- Statistics is defined as the science of **collecting**, **organizing**, and **analyzing data**.
- It also involves **summarizing** data.
- The basic mathematics required for collecting, organizing, analyzing, and visualizing data falls under statistics.

Understanding Data

- **Data** refers to facts or pieces of information.
- Data is generated from various sources, including smartphone usage, software applications, and social media interactions.
- Examples of data include:
    - Heights of students in a class (e.g., 170 cm, 145 cm).
    - Gender of a person visiting a doctor.

Types of Statistics

- Statistics is broadly divided into two main types: **Descriptive Statistics** and **Inferential Statistics**.



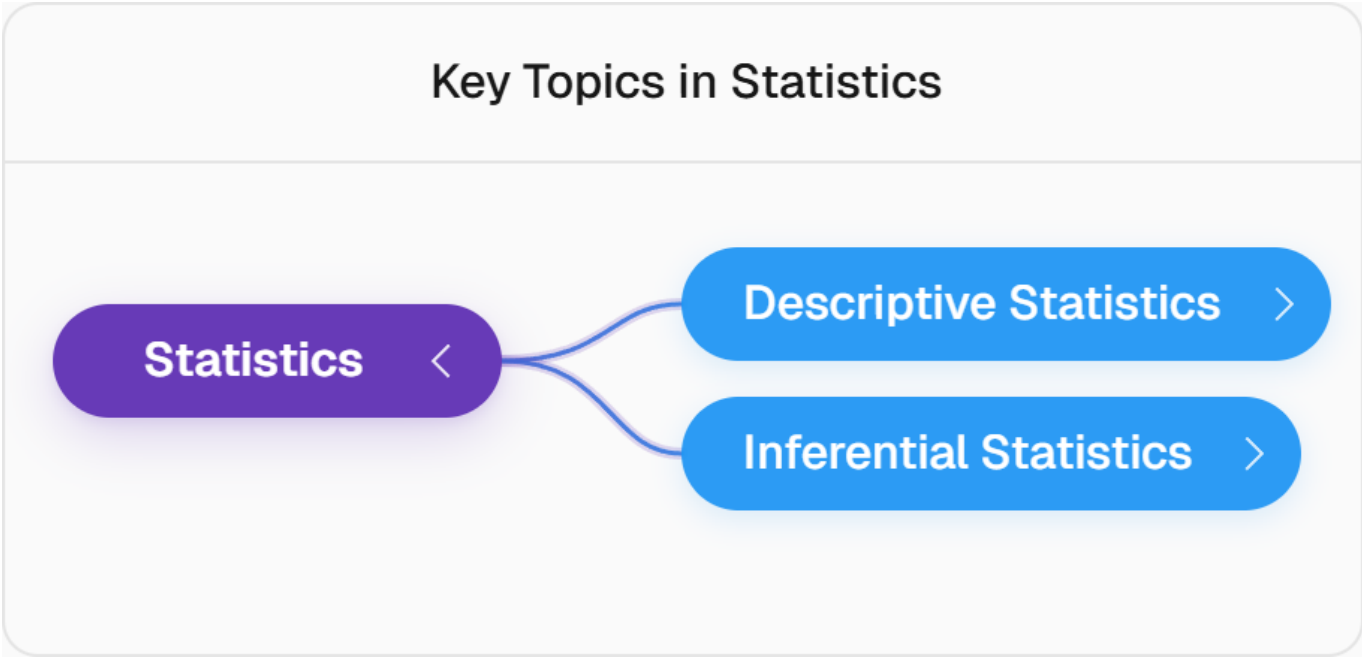
Descriptive Statistics

- **Definition**: Descriptive statistics involves **organizing** and **summarizing data**.
- **Key Topics**:
  - **Measures of Central Tendency**: This includes concepts like **mean**, **median**, and **mode**.
  - **Measures of Dispersion/Variance**: Topics such as **variance** and **standard deviation** are covered here.
  - **Data Distribution**: Understanding different types of data distributions.
  - **Visualization and Analysis Techniques**:
    - **Histograms**: Used to visualize the frequency of elements in data.
    - **Probability Density Function (PDF)**: Helps in understanding the type of data distribution, such as **Normal Distribution** or **Log-Normal Distribution**.

Inferential Statistics

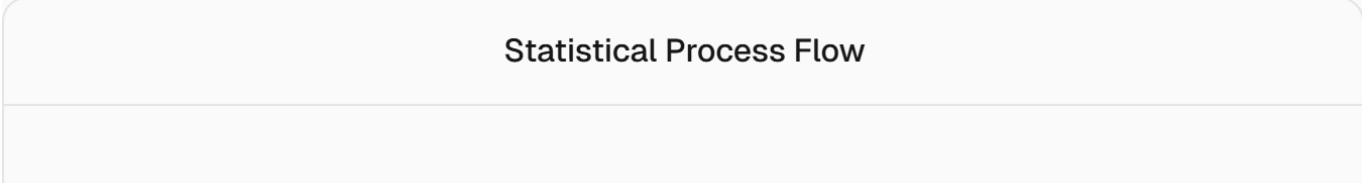
- **Definition**: Inferential statistics involves using existing data (typically a sample) to **form conclusions** about a larger dataset (population).
- It primarily focuses on drawing conclusions about **population data** based on **sample data**.
- **Key Topics**:
  - **Hypothesis Testing**: A core concept that includes various statistical tests.

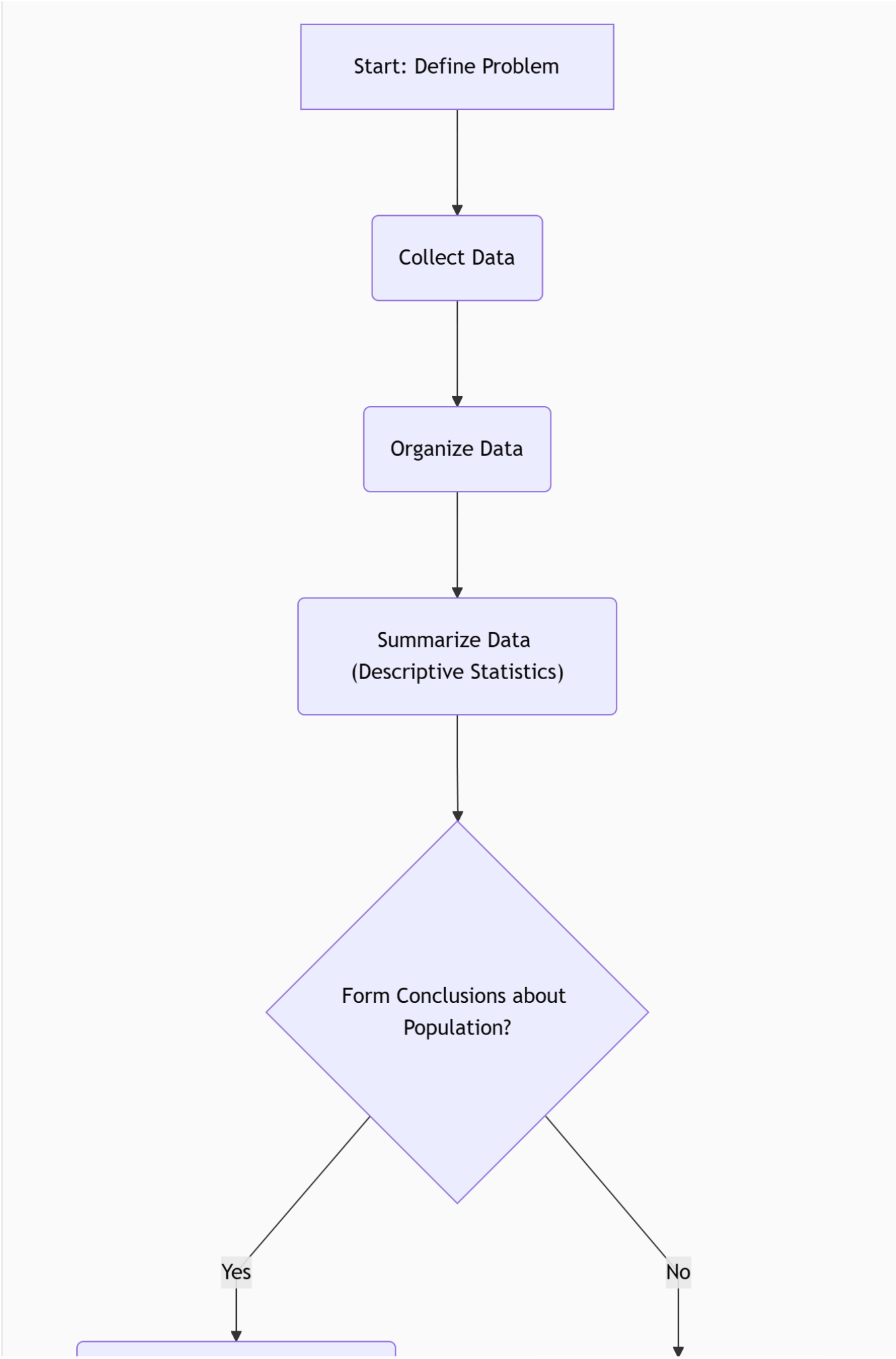
- **Statistical Tests**: Examples include **Z-test**, **T-test**, **Chi-square test**, and **ANOVA test**.
- **Hypothesis Testing Components**: Involves understanding concepts like null hypothesis, alternate hypothesis, **p-value**, and **significance value**.

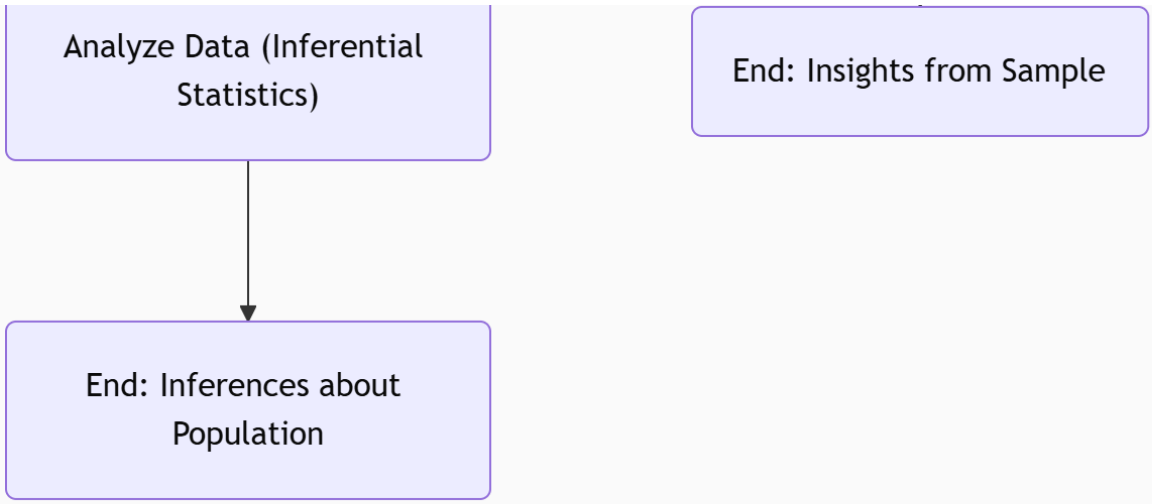


Examples: Descriptive vs. Inferential Statistics

- **Scenario**: Consider a college with 25 classrooms. Ages of students are collected from just one classroom (this is the **sample data**).
- **Descriptive Statistics Example**:
  - **Question**: "What is the common age of students in *this specific class*?".
  - **Approach**: Calculate the **average (mean)** age of students in that class.
  - **Outcome**: This is an example of descriptive statistics because it summarizes information directly from the collected sample data using measures of central tendency.
- **Inferential Statistics Example**:
  - **Question**: "Are the ages of students in *this classroom similar to what you expect from students in the entire university* (the population)?".
  - **Approach**: This involves comparing the sample data (ages from one class) to the broader population (ages of all university students) to draw a conclusion.
  - **Outcome**: This is an example of inferential statistics, as it uses sample data to make inferences or conclusions about a larger population.




Start: Define Problem
Collect Data
Organize Data
Summarize Data
(Descriptive Statistics)
Form Conclusions about
Population?
Yes
No

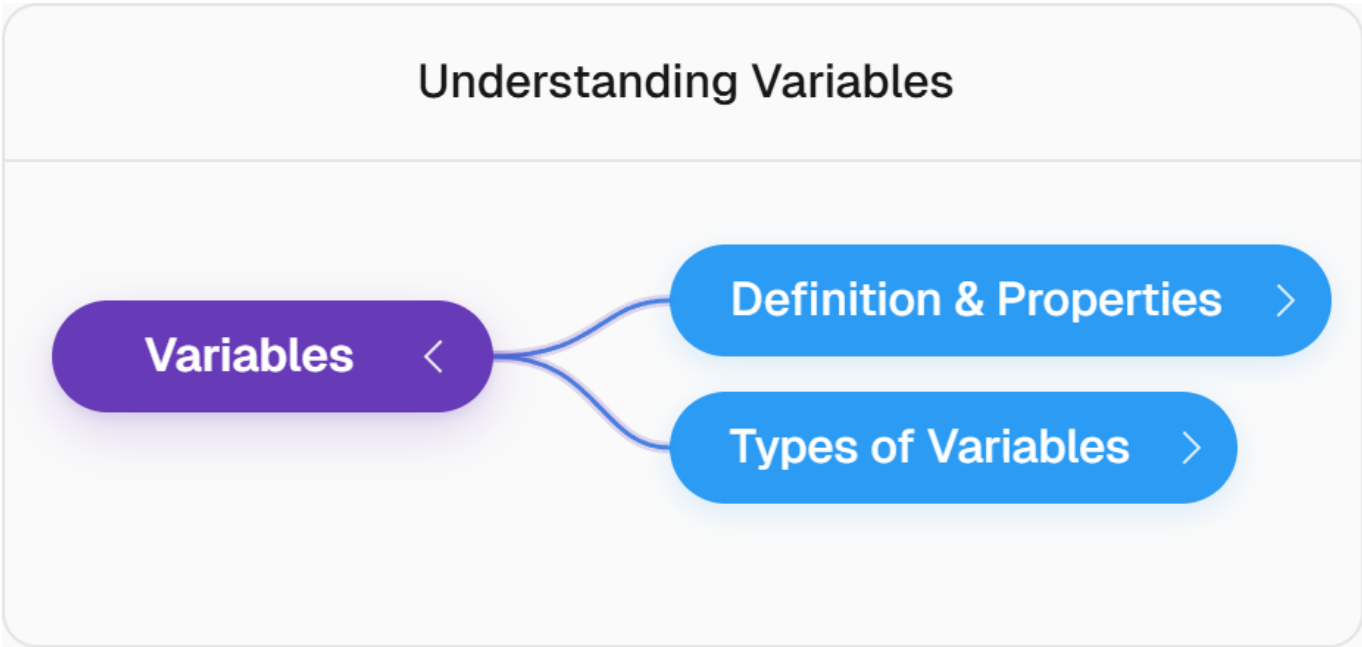


Day - 3

Understanding Variables

- A **variable** is defined as a property that can take on multiple values.
- Examples of variables include **age** (e.g., 12, 30, 40, 100 years), **height** (e.g., 170.2 cm, 180.3 cm), and **weight** (e.g., 72 kg, 72.5 kg).
- It is crucial to remember that a variable is always discussed in its singular form (e.g., "age" not "ages"), as plural forms refer to collections of values, not the variable itself.

Here is a mind map illustrating the definition and types of variables:



Types of Variables

- Variables are broadly categorized into two main types: **Quantitative Variables** and **Qualitative or Categorical Variables**.

Quantitative Variables

- **Quantitative variables** are numerical and can be further divided into two sub-types: **Discrete** and **Continuous**.
    - **Discrete variables** always represent a number that is countable and takes specific, distinct values, typically whole numbers.
        - Examples include the number of bank accounts or the number of children in a family, which cannot be fractional (e.g., 2.5 children).
    - **Continuous variables** can take on an infinite number of values within a given range and are not restricted to whole numbers.
        - Examples are height (e.g., 175.25 cm) or weight (e.g., 72.5 kg), where values can include decimals.

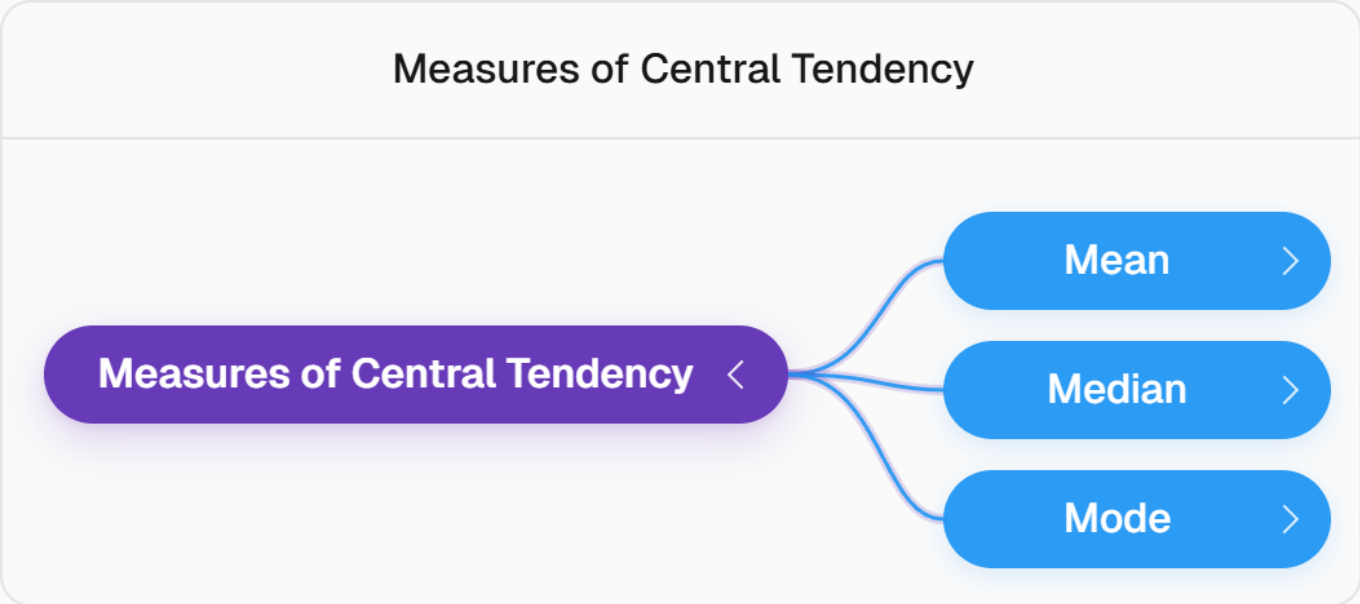
Qualitative (Categorical) Variables

- **Qualitative** or **Categorical variables** classify variables based on certain properties or characteristics.
- These variables involve classification rather than numerical measurement.
    - An example is **gender**, which classifies individuals as male or female based on inherent properties.
    - Another example is the **type of flowers**, such as roses, lilies, or cacti, where classification is based on distinct botanical properties.

Day - 4

Measures of Central Tendency

- **Measures of Central Tendency** are statistical metrics used to determine the center or distribution of a dataset.
- The three main measures of central tendency are **Mean**, **Median**, and **Mode**.
- Understanding these measures is crucial for project discussions and system knowledge, especially when dealing with population and sample data.



Mean

- The **mean** represents the average of all values in a dataset.
- For **population data**, the mean is denoted by μ (mu).
- The formula for the population mean is: $$ \mu = \frac{\sum_{i=1}^{N} x_i}{N} $$ where N is the size of the population data.
- For **sample data**, the mean is denoted by $\bar{x}$ (x-bar).
- The formula for the sample mean is: $$ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} $$ where n is the size of the sample data.
- x_i represents individual elements within the dataset.
- For example, given a population data set `[1, 2, 2, 3, 4, 5]`, the sum is 17 and there are 6 elements, so the mean is $17/6 \approx 2.83$.
- The mean is highly sensitive to **outliers**, meaning extreme values can significantly skew its value.

Median

- The **median** is the central element in a dataset after the data has been sorted.
- To find the median, first arrange all elements in ascending or descending order.
- If the number of elements is **odd**, the median is the single middle element; for example, in `[1, 2, 3, 4, 5]`, the median is 3.
- If the number of elements is **even**, the median is the average of the two central elements; for example, in `[1, 2, 3, 4, 5, 100]`, the median is $(3+4)/2 = 3.5$.
- The median is robust to outliers, meaning it is less affected by extreme values compared to the mean. For instance, adding '100' to a dataset significantly changes the mean but has a much smaller impact on the median.
- It is particularly useful when the data distribution contains outliers, as it helps prevent their impact on the central tendency representation.

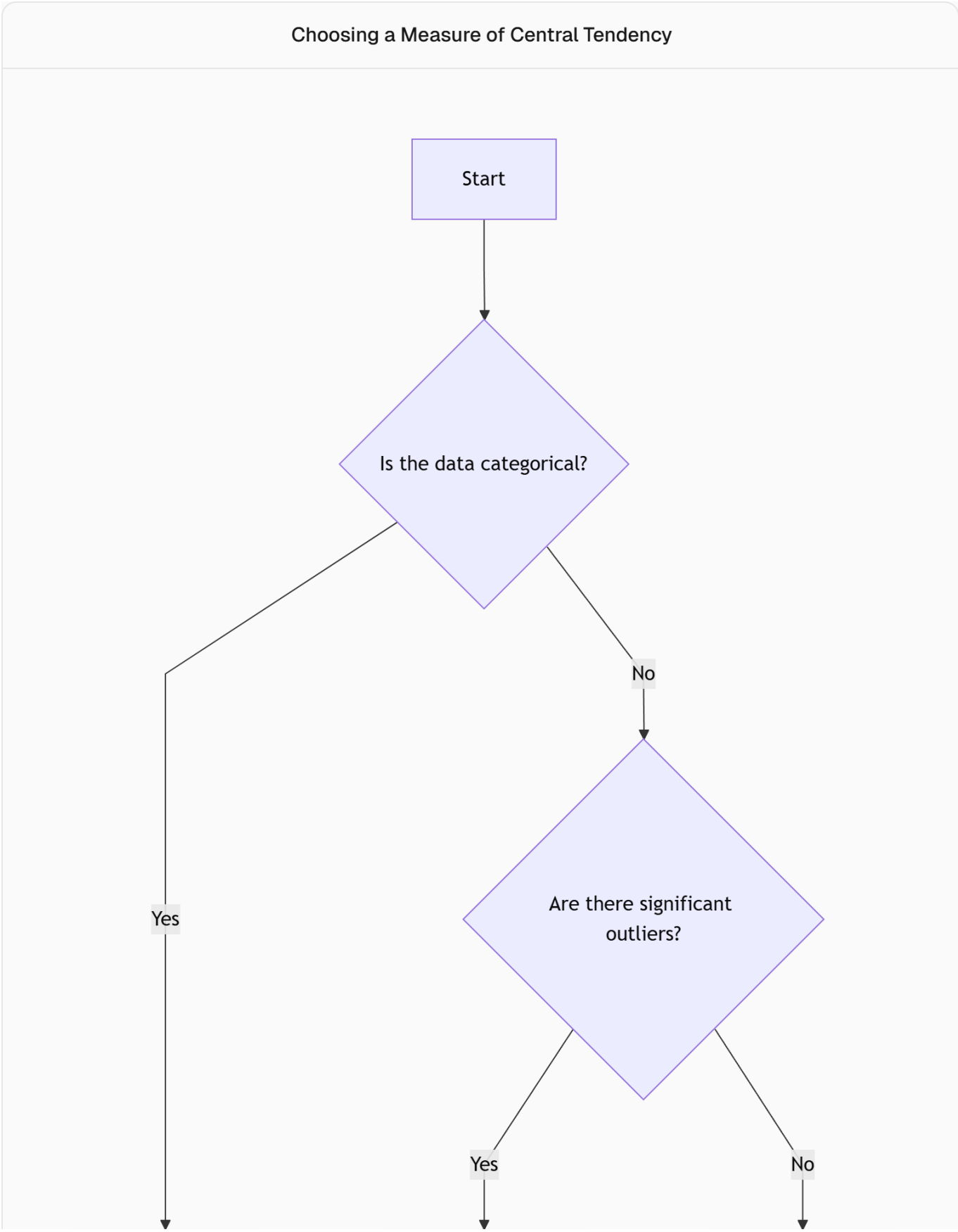
Mode

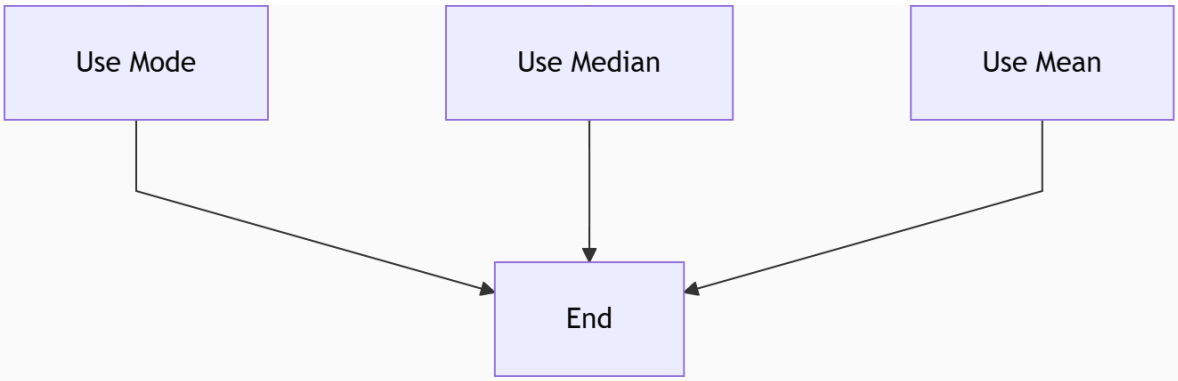
- The **mode** is the most frequently occurring element in a dataset.
- For example, in the dataset `[1, 2, 2, 3, 4, 5, 5, 5, 5]`, the number 5 appears most frequently (four times), making it the mode.

- The mode is primarily used for **categorical features**, such as 'Gender' (Male/Female).
- It can also be used to replace **missing values** in categorical data by filling them with the most frequent category.

When to Use Each Measure

- The choice of central tendency measure depends on the type and distribution of the data.



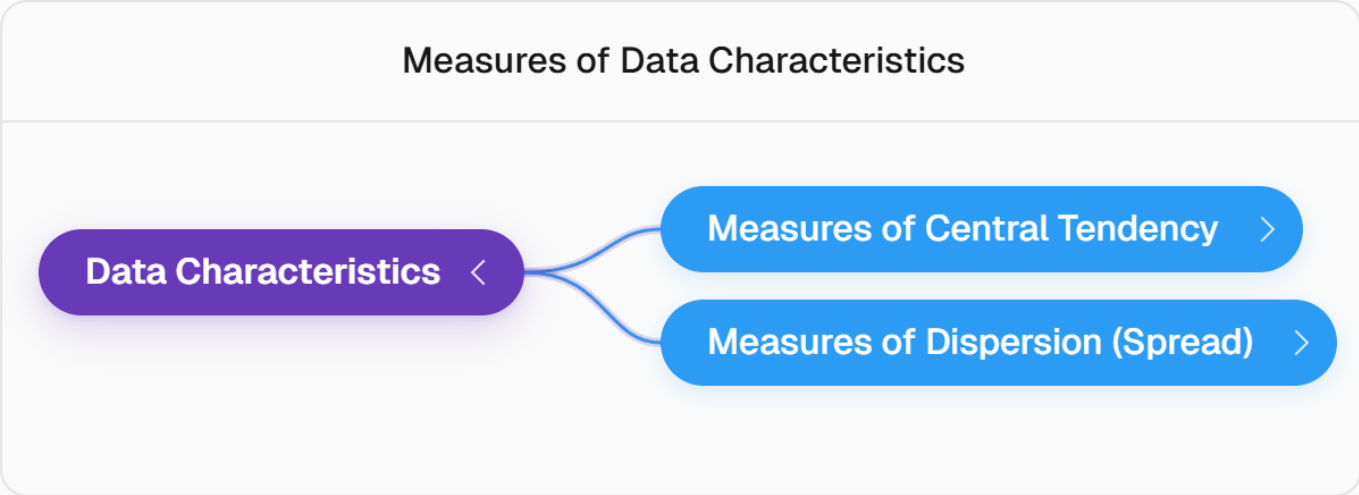


- **Mean** is generally preferred for normally distributed data without significant outliers.
- **Median** is recommended when the dataset contains outliers, as it provides a more representative central value by mitigating their influence.
- **Mode** is best suited for categorical data or when dealing with missing values in categorical features, as it identifies the most common category.

Day - 5

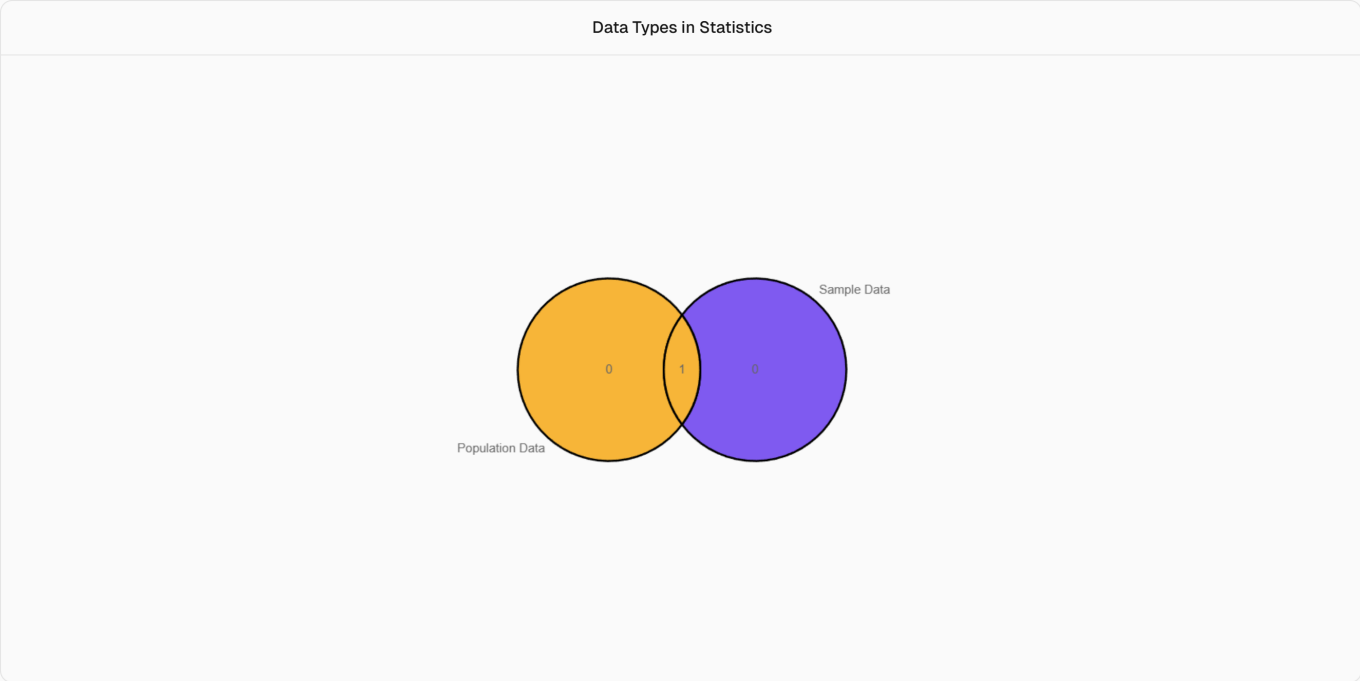
Measures of Dispersion

- Measures of dispersion, also known as **spread**, quantify how spread out data is within a dataset.
- They are distinct from **measures of central tendency** (mean, median, mode), which describe the central element of data.
- Key measures of dispersion include **variance** and **standard deviation**.



Population vs. Sample

- Statistical calculations often differentiate between **population** (capital 'N') and **sample** (small 'n') data.
- For a **population mean** (μ), the formula is: $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$
- For a **sample mean** ($\bar{x}$), the formula is: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



Variance

- **Variance** quantifies the spread of data and is a crucial concept in understanding data distribution.
- **Population variance** (σ^2) is defined as: $$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$
- **Sample variance** (s^2) is defined as: $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$
- The division by `n-1` in sample variance is known as **Bessel's correction** or **degrees of freedom**, which is used to make the sample variance an unbiased estimator of the population variance.
- When the variance is a **large number**, it indicates a **large spread** in the data.
- When the variance is a **small number**, it indicates a **small spread** in the data, resulting in a taller, narrower distribution curve.

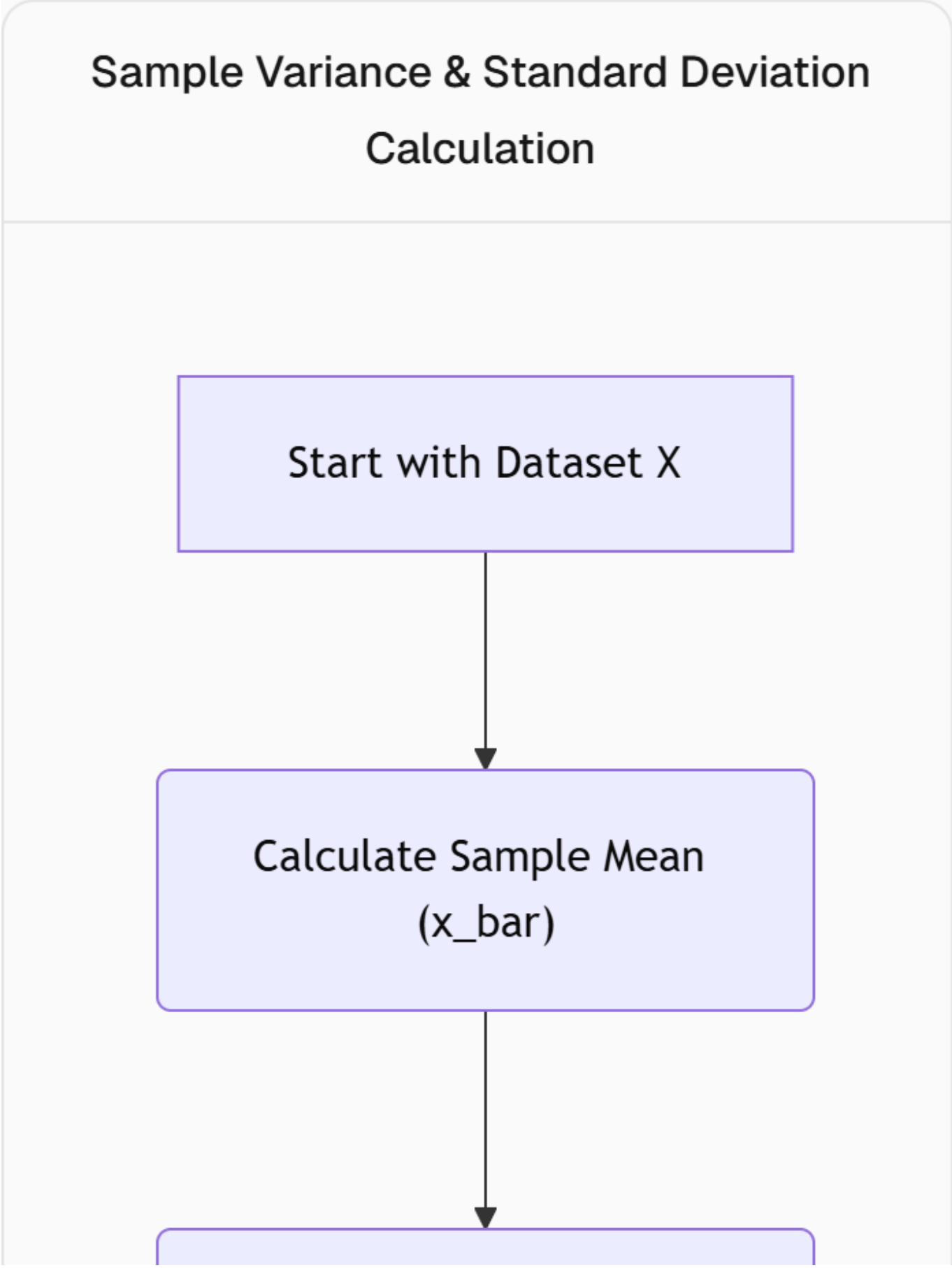
Standard Deviation

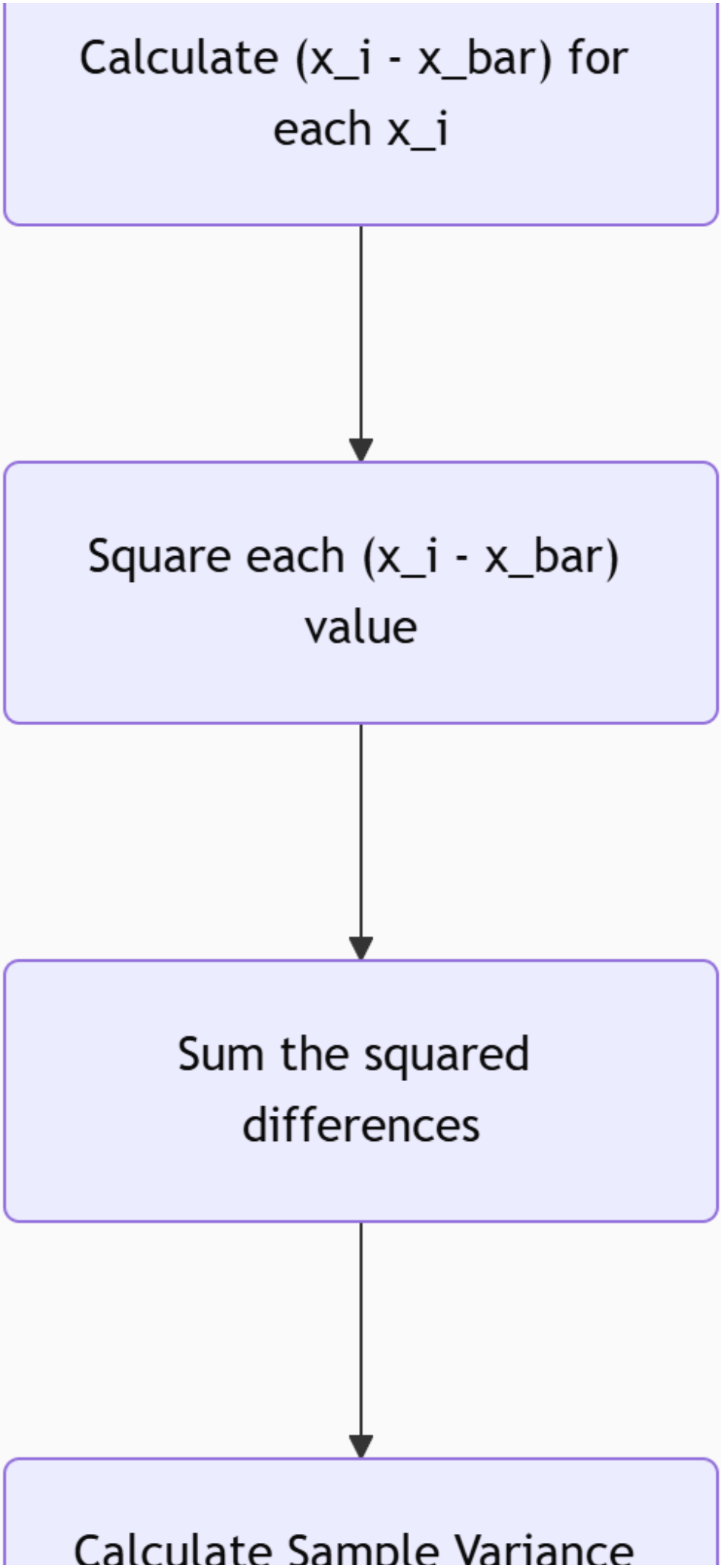
- **Standard deviation** is the square root of the variance.
- It is a direct measure of the average distance of data points from the mean.
- If variance is a large number, standard deviation will also be large, indicating a wide spread.
- If standard deviation is a small number, the data distribution will have a high peak and less spread.

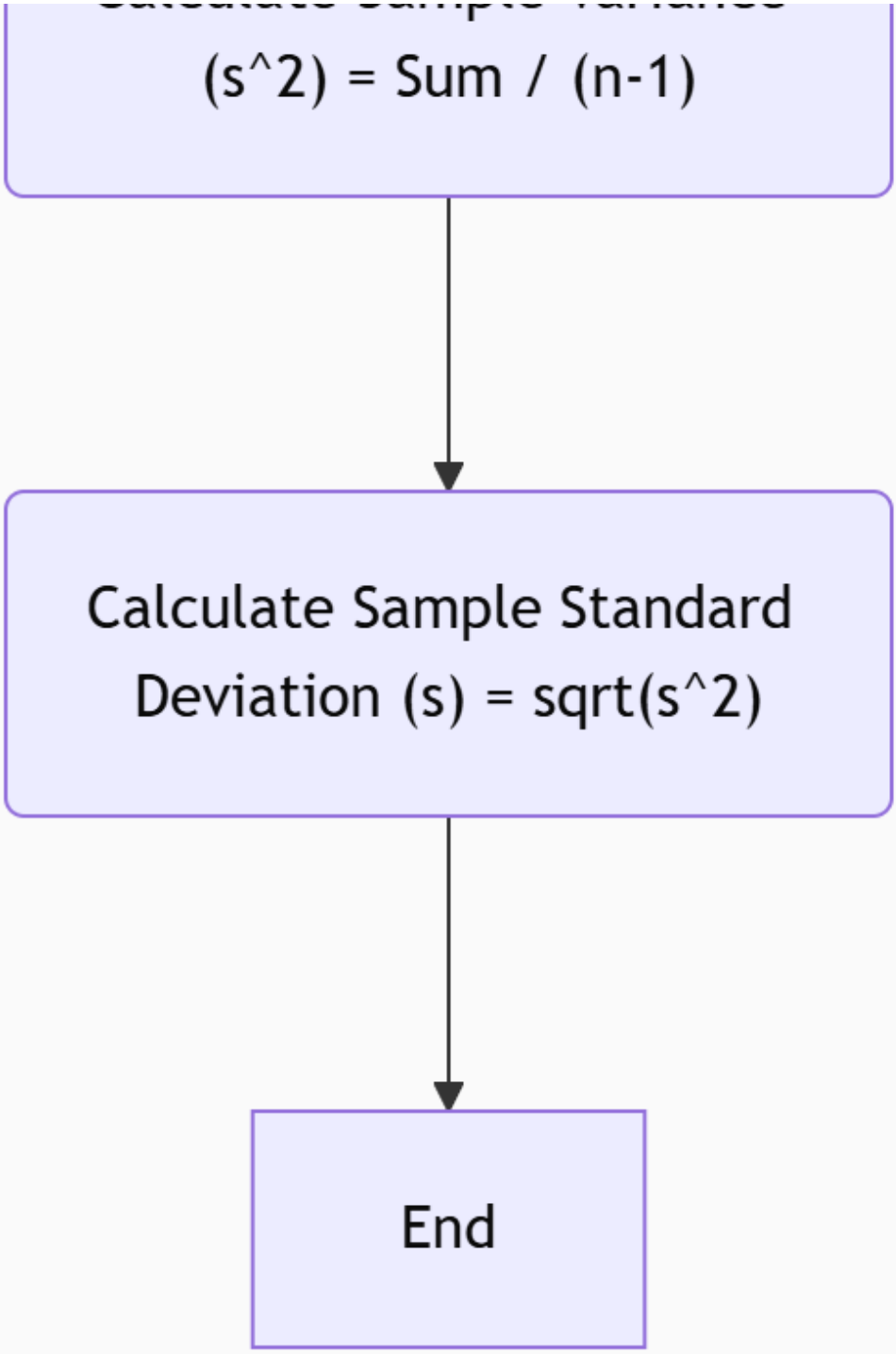
Calculation Example for Sample Variance and Standard Deviation

- Consider a dataset: `X = {1, 2, 2, 3, 4, 5}`.
- **Step 1: Calculate the Mean ($\bar{x}$)**:
  - Sum of elements: `1+2+2+3+4+5 = 17`
  - Number of elements (`n`): `6`
  - Mean: `17 / 6 = 2.83`
- **Step 2: Calculate $(x_i - \bar{x})$ for each data point**.
- **Step 3: Square each $(x_i - \bar{x})$ value**.

- **Step 4: Sum the squared differences** ($\sum (x_i - \bar{x})^2$). For the example, this sum is `10.84`.
- **Step 5: Calculate Sample Variance (s^2)**:
  - $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{10.84}{6-1} = \frac{10.84}{5} = 2.168$
- **Step 6: Calculate Sample Standard Deviation (s)**:
  - $s = \sqrt{s^2} = \sqrt{2.168} \approx 1.472$




Calculate (x_i - x_bar) for each x_i
Square each (x_i - x_bar) value
Sum the squared differences
Calculate Sample Variance



Interpreting Standard Deviation Ranges

- Standard deviation helps define ranges around the mean, indicating where data points fall within a distribution.
- For a mean of `2.83` and a standard deviation of `1.472`:
  - **First Standard Deviation to the Right**: `2.83 + 1.472 = 4.302`. Data points between `2.83` and `4.302` fall within this range.

- **Second Standard Deviation to the Right**: `4.302 + 1.472 = 5.774`. Data points between `4.302` and `5.774` fall within this range.
- **Third Standard Deviation to the Right**: `5.774 + 1.472 = 7.246`. Data points between `5.774` and `7.246` fall within this range.
- **First Standard Deviation to the Left**: `2.83 - 1.472 = 1.358`. Data points between `1.358` and `2.83` fall within this range.
- **Second Standard Deviation to the Left**: `1.358 - 1.472 = -0.114`. Data points between `-0.114` and `1.358` fall within this range.

> 💡 **Key Insight:** A larger variance or standard deviation means the data is more spread out, while a smaller value means the data points are clustered closer to the mean, resulting in a higher peak in the distribution curve. ❗ **Important:** Measures of dispersion are crucial for understanding the **spread of data**, which is why they are also called **measures of dispersion**.

Hypothesis - Testing (Tutorial - 17)

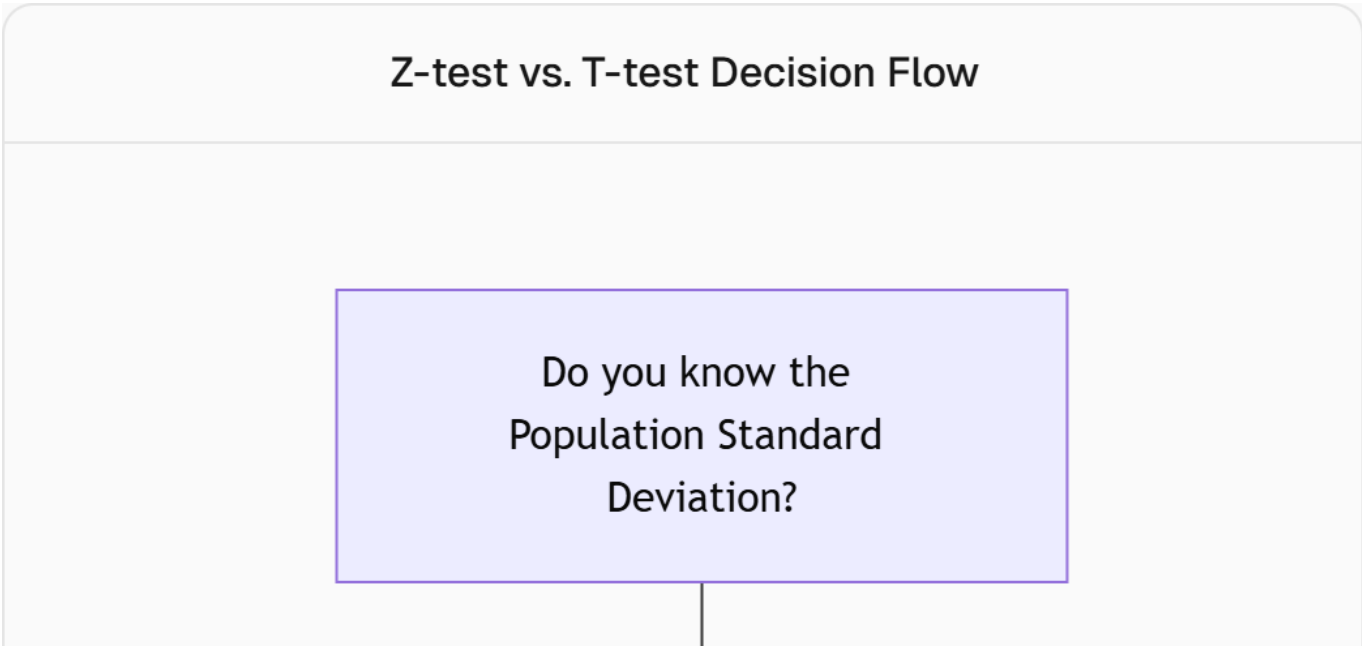
Introduction to Hypothesis Testing and Z-test

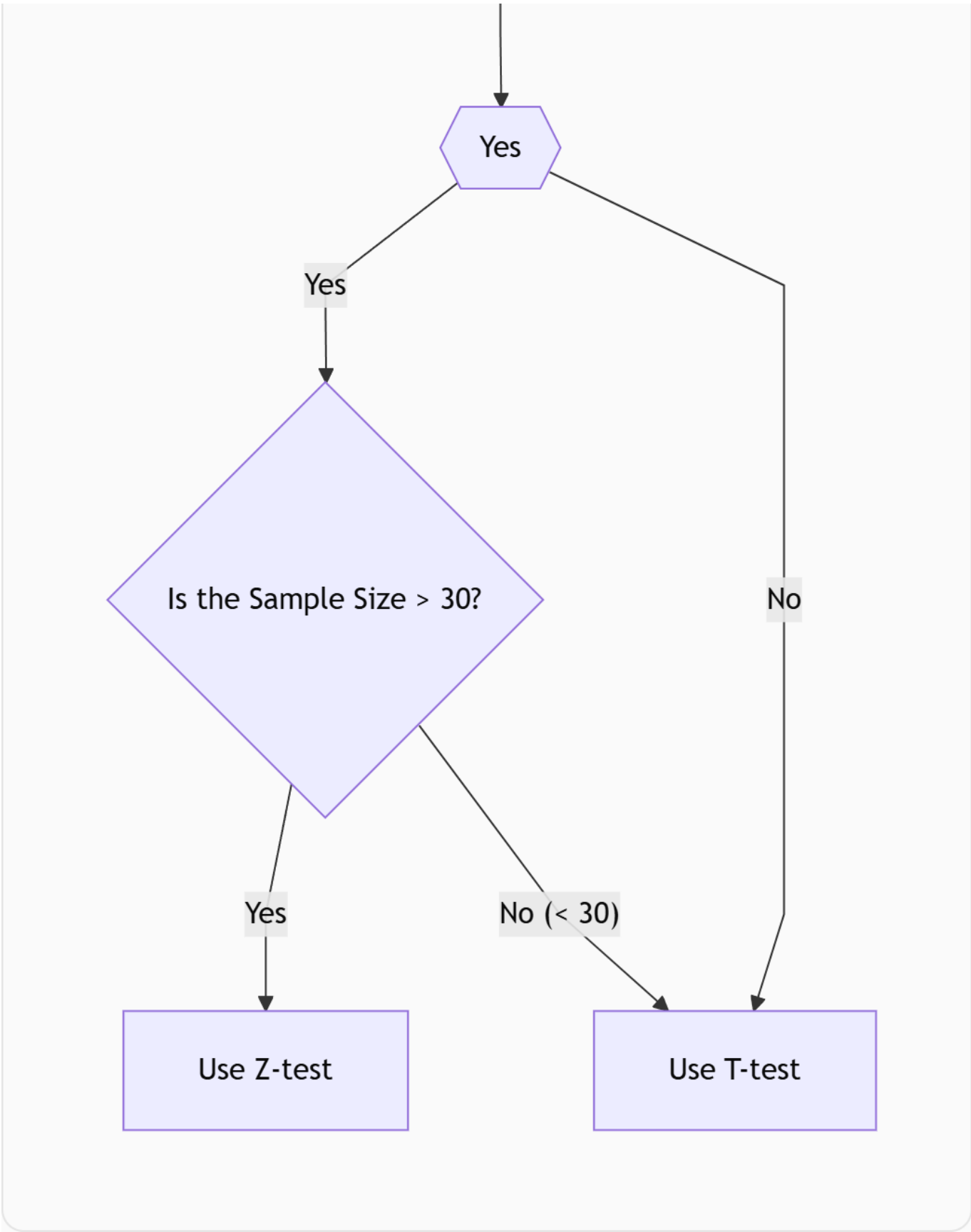
- Hypothesis testing and statistical analysis are crucial for data science aspirants to solve business use cases.
- These techniques are specifically used in **inferential statistics**, which aims to draw conclusions about **population data** using **sample data**.
- This video focuses on understanding the **Z-test**.

When to Use Z-test vs. T-test

- A key interview question is when to use a **Z-test** versus a **T-test**, as many people get confused.
- The decision depends on two main factors: knowing the **population standard deviation** and the **sample size**.

Here is a flowchart to help decide between Z-test and T-test:



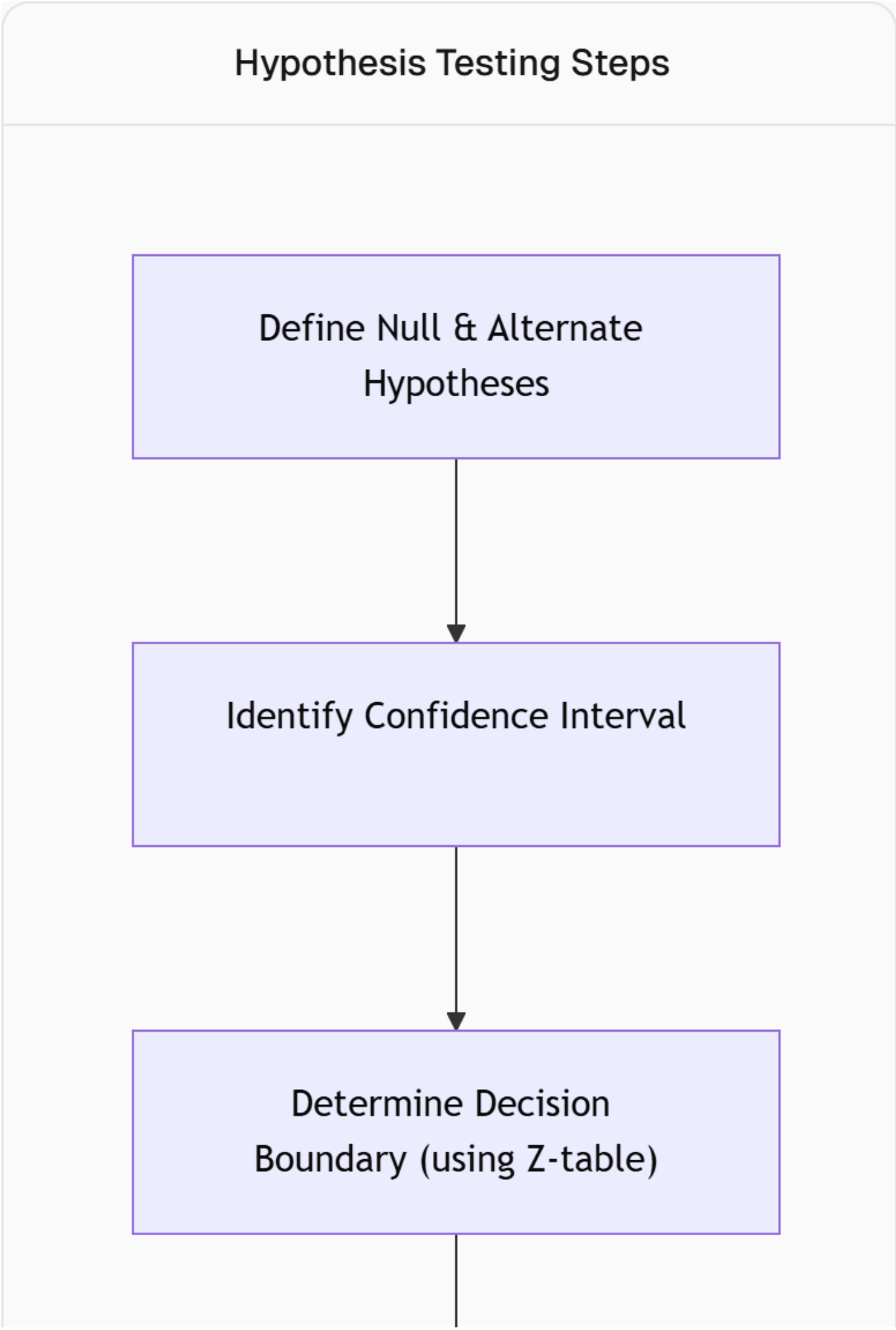


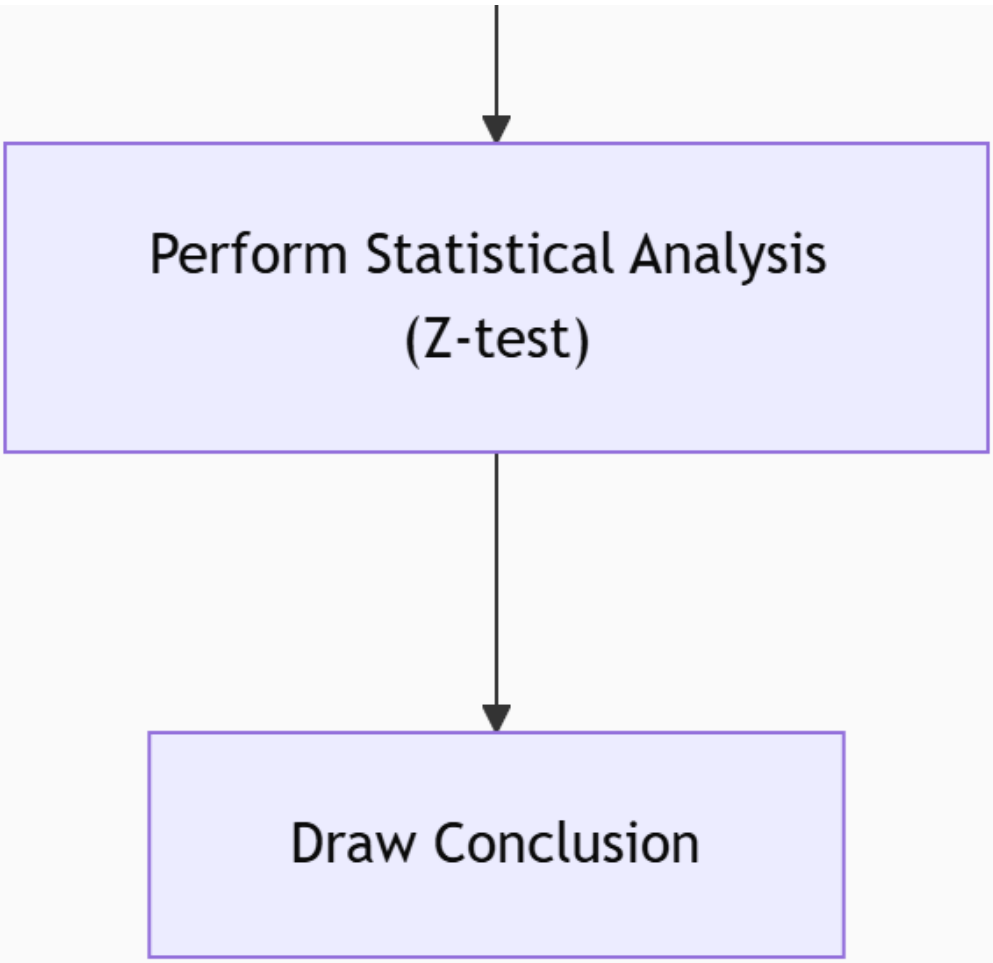
- If the **population standard deviation** is unknown, you should directly use the **T-test**, even if the sample standard deviation is known.

Steps in Hypothesis Testing

- Hypothesis testing involves several steps to draw conclusions from data.

Here is a flowchart outlining the general steps in hypothesis testing:





Example Problem Statement and Data Extraction

- **Problem Statement:** The average height of all residents in a city is 165 cm with a population standard deviation of 3.9 cm. A doctor believes the mean to be different, specifically 168 cm. He measured the height of 36 individuals and found their average height to be 169.5 cm. The analysis should be at a 95% confidence interval.

Here's a summary of the extracted data:

Parameter	Value
Population Mean (μ)	168 cm
Population Std. Dev. (σ)	3.9 cm
Sample Size (n)	36
Sample Mean ($\bar{x}$)	169.5 cm
Confidence Interval (CI)	95% (0.95)

Defining Null and Alternate Hypotheses

- **Null Hypothesis (H_0):** The average height of all residents is 168 cm ($\mu = 168$ cm).
- **Alternate Hypothesis (H_1):** The doctor believes the mean is different, meaning the mean is not equal to 168 cm ($\mu \neq 168$ cm).
- This scenario is a **two-tailed test** because the alternate hypothesis suggests the mean could be greater than or less than 168 cm.

Confidence Interval and Decision Boundary

- The **confidence interval (CI)** is 95% (0.95).
- The **significance value (α)** is calculated as 1 - CI, which is 1 - 0.95 = 0.05.
- For a two-tailed test with a 95% CI, the remaining 5% (0.05) is split into two **rejection areas** of 2.5% (0.025) each on both tails of the distribution.
- The central 95% area is the **acceptance area**.
- To find the **decision boundary** (critical Z-values), we refer to the **Z-table**.
- For the upper bound of the 95% CI, the cumulative area under the curve is 1 - 0.025 = 0.9750.
- Looking up 0.9750 in the Z-table gives a Z-score of **1.96**.
- Due to symmetry, the lower bound is **-1.96**.
- **Decision Rule:** If the calculated Z-test value falls between -1.96 and +1.96, we **fail to reject the null hypothesis**. Otherwise, we **reject the null hypothesis**.

Statistical Analysis (Z-test Calculation)

- The **Z-test formula** is: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ Where:
  - $\bar{X}$ is the sample mean
  - μ is the population mean
  - σ is the population standard deviation
  - n is the sample size
- The term $\frac{\sigma}{\sqrt{n}}$ is known as the **standard error**.
- Plugging in the values from the example: $Z = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}}$ $Z = \frac{1.5}{\frac{3.9}{6}}$ $Z = \frac{1.5}{0.65}$ $Z \approx 2.31$

Conclusion

- The calculated Z-value is **2.31**.
- Comparing this to the decision boundary of -1.96 to +1.96, **2.31 is greater than 1.96**.
- Therefore, we **reject the null hypothesis**.

- This means the doctor's belief that the mean height is different is correct. The sample's average height (169.5 cm) is significantly higher than the assumed population mean (168 cm), falling into the rejection area on the positive side.

Other Statistical Tests

- Other statistical tests include:
    - **Chi-square test:** Used for **categorical data**.
    - **ANOVA (Analysis of Variance):** Used to check if the **variances** of two data sets are the same or different, with respect to sample data.

Practice Problem

- A factory manufactures bulbs with an average warranty of 5 years and a standard deviation of 0.50 years. A worker believes a bulb will malfunction in less than 5 years. A sample of 40 bulbs has an average warranty of 4.8 years. Solve this problem at a 2% significance level.

T - Test

Introduction to T-Test

- The T-test is a statistical test used for hypothesis testing, forming a part of inferential statistics.

T-Test vs. Z-Test

- The choice between a T-test and a Z-test depends on whether the **population standard deviation** is known and the **sample size**.

Condition	Test to Use
Population Standard Deviation Known	Z-Test
Sample Size > 30 (even if population SD unknown)	Z-Test
Population Standard Deviation Unknown	T-Test
Sample Size < 30 (even if population SD known)	T-Test

Problem Statement: Medication Effect on IQ

- A team of researchers wants to test a new medication's effect on intelligence, where the average IQ in the population is 100.
- A sample of 30 participants who took the medication has a mean IQ of 140 with a sample standard deviation of 20.
- The confidence interval for this test is 95%, and the question is whether the medication affected intelligence.

Steps in T-Test Hypothesis Testing

1. Formulating Hypotheses

- The **null hypothesis (H0)** states that the average IQ is 100, implying the medication has no effect.
- The **alternate hypothesis (H1)** states that the mean IQ is not equal to 100, indicating the medication could have a positive or negative effect.
- This scenario, where the effect can be positive or negative, is known as a **two-tailed test**.

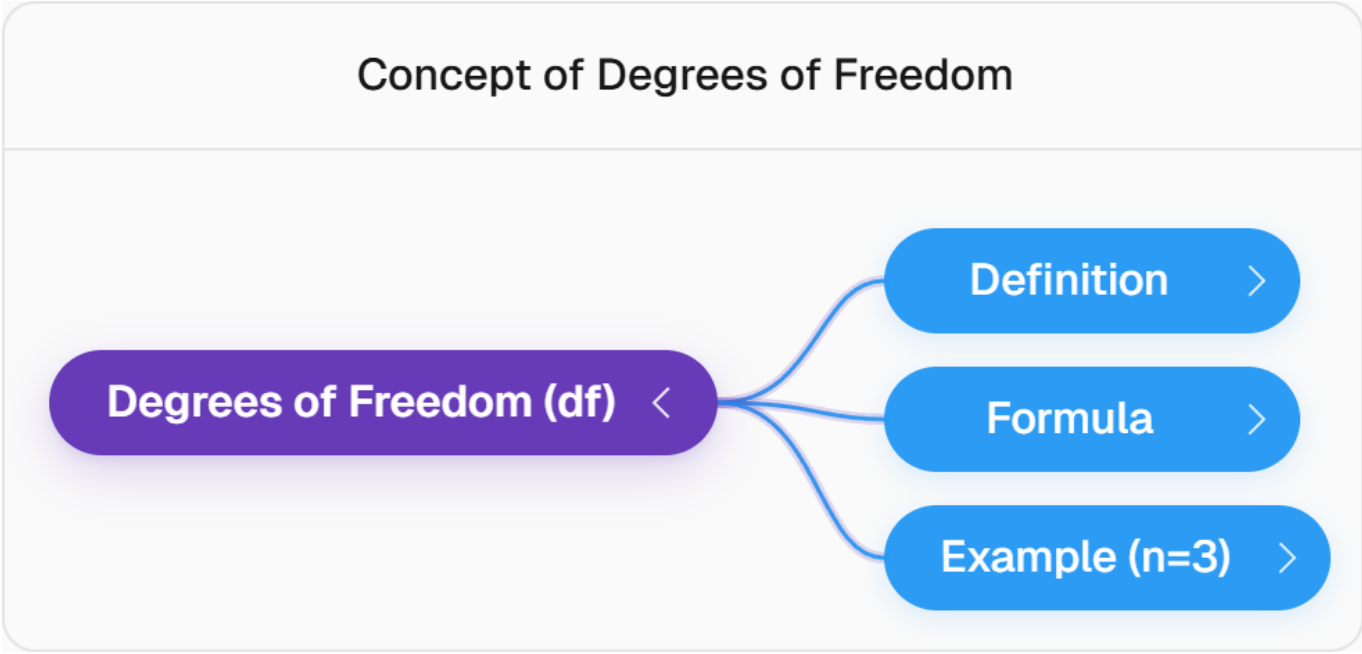
2. Determining Significance Value

- The significance value (α) is calculated as 1 minus the confidence interval.
- For a 95% confidence interval, α = 1 - 0.95 = 0.05.

3. Calculating Degrees of Freedom (df)

- Degrees of freedom are crucial for the T-test and are calculated using the formula: **n - 1**, where 'n' is the sample size.
- For a sample size of 30 participants, the degrees of freedom (df) are 30 - 1 = 29.
- Degrees of freedom represent the number of choices available or the number of values in a calculation that are free to vary.
- For example, with 3 options (n=3), the first person has 3 choices, the second has 2, and the last has only 1, resulting in 3-1=2 degrees of freedom.

Here is a mind map illustrating the concept of Degrees of Freedom:



4. Establishing Decision Boundary

- For a two-tailed test with a 95% confidence interval (α = 0.05), each tail represents 2.5% (0.025) of the distribution.
- Using a T-table, with degrees of freedom (df) = 29 and a two-tailed significance level of 0.05 (or one-tailed 0.025), the critical T-value is found to be 2.045.
- The decision boundary for the acceptance area is between -2.045 and +2.045.
- If the calculated T-statistic falls outside this range (i.e., less than -2.045 or greater than +2.045), the null hypothesis is rejected.

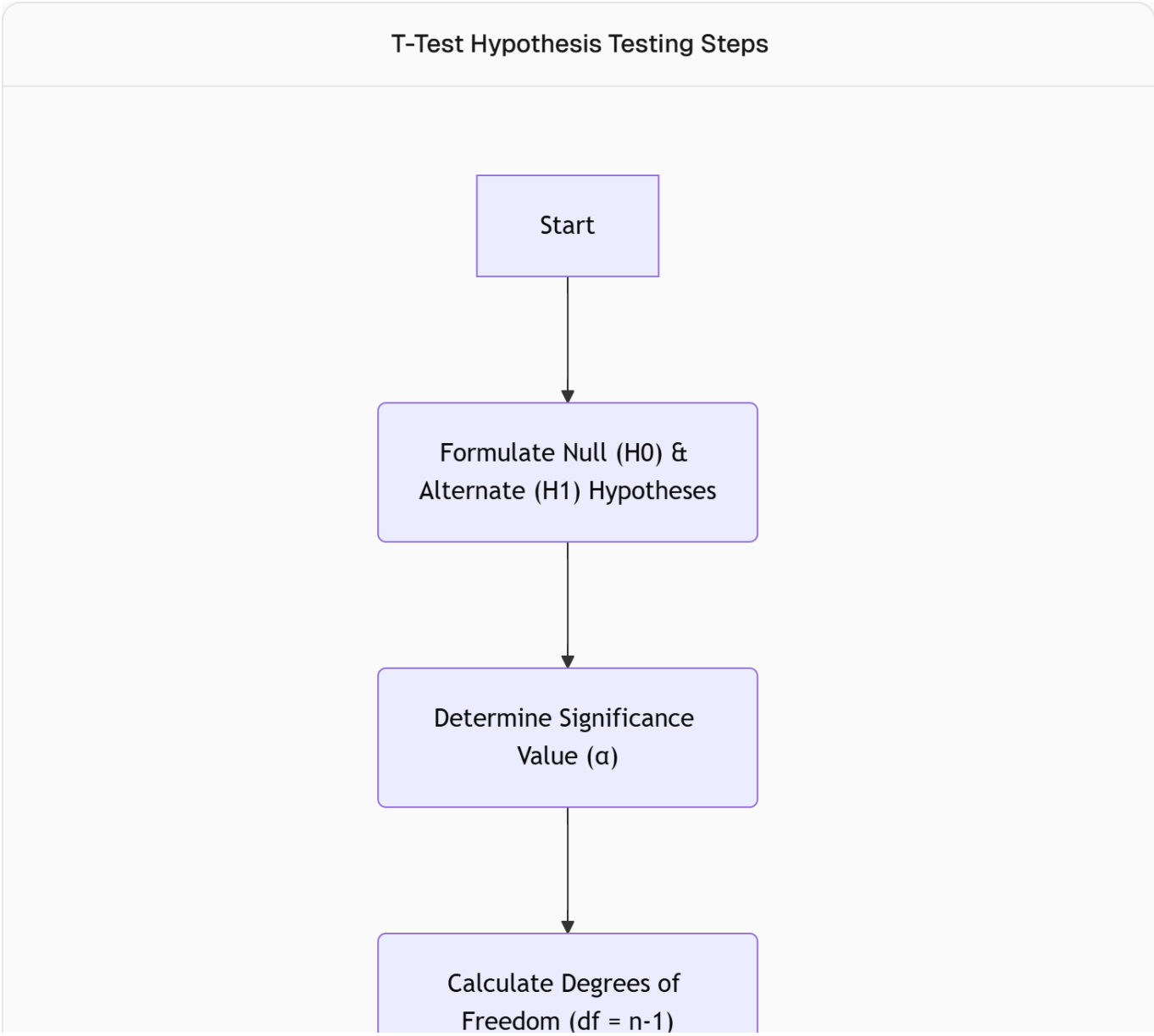
5. Calculating T-Test Statistic

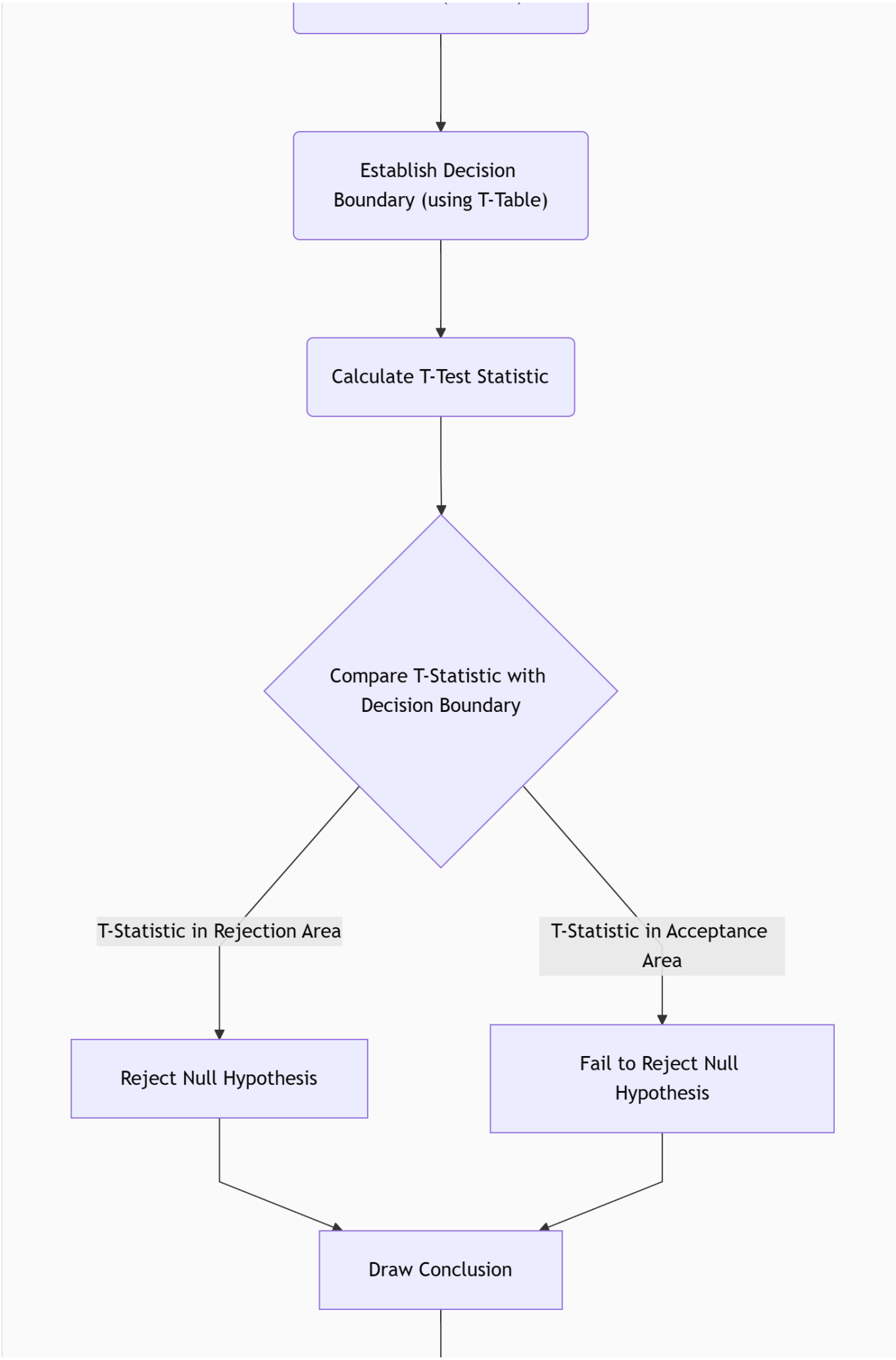
- The T-test statistic is calculated using the formula: $$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$ Where:
- $\bar{X}$ = Sample Mean (140)
- μ = Population Mean (100)
- S = Sample Standard Deviation (20)
- n = Sample Size (30)
- Substituting the values: $$T = \frac{140 - 100}{20 / \sqrt{30}}$$ $$T = \frac{40}{20 / 5.477}$$ $$T \approx \frac{40}{3.65} \approx 10.96$$

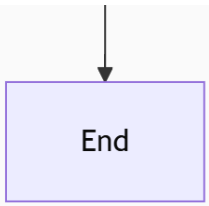
6. Drawing Conclusion

- The calculated T-statistic is 10.96.
- Since 10.96 is greater than the critical T-value of 2.045, it falls into the rejection area.
- Therefore, the null hypothesis is rejected.
- This conclusion indicates that the medication has a positive effect on intelligence, as the T-value is significantly positive.

Here is a flowchart illustrating the steps involved in performing a T-Test:




Establish Decision Boundary (using T-Table)
Calculate T-Test Statistic
Compare T-Statistic with Decision Boundary
T-Statistic in Rejection Area
T-Statistic in Acceptance Area
Reject Null Hypothesis
Fail to Reject Null Hypothesis
Draw Conclusion



Tutorial 19- Type 1 And Type 2 Error In Statistics

Introduction to Type I and Type II Errors

- This tutorial focuses on **Type I** and **Type II errors** in statistics, a crucial topic also applied in machine learning.
- The **confusion matrix** is used in machine learning, especially for **binary classification models**, to evaluate model accuracy, precision, and recall.
- It compares **actual values** (known outputs from the dataset) with **predicted values** (model outputs).
- The confusion matrix categorizes predictions into **true positive**, **false positive**, **false negative**, and **true negative** scenarios.

Actual Value	Predicted Value	Classification
1	**1**	True Positive (TP)
0	**1**	False Positive (FP)
1	**0**	False Negative (FN)
0	**0**	True Negative (TN)

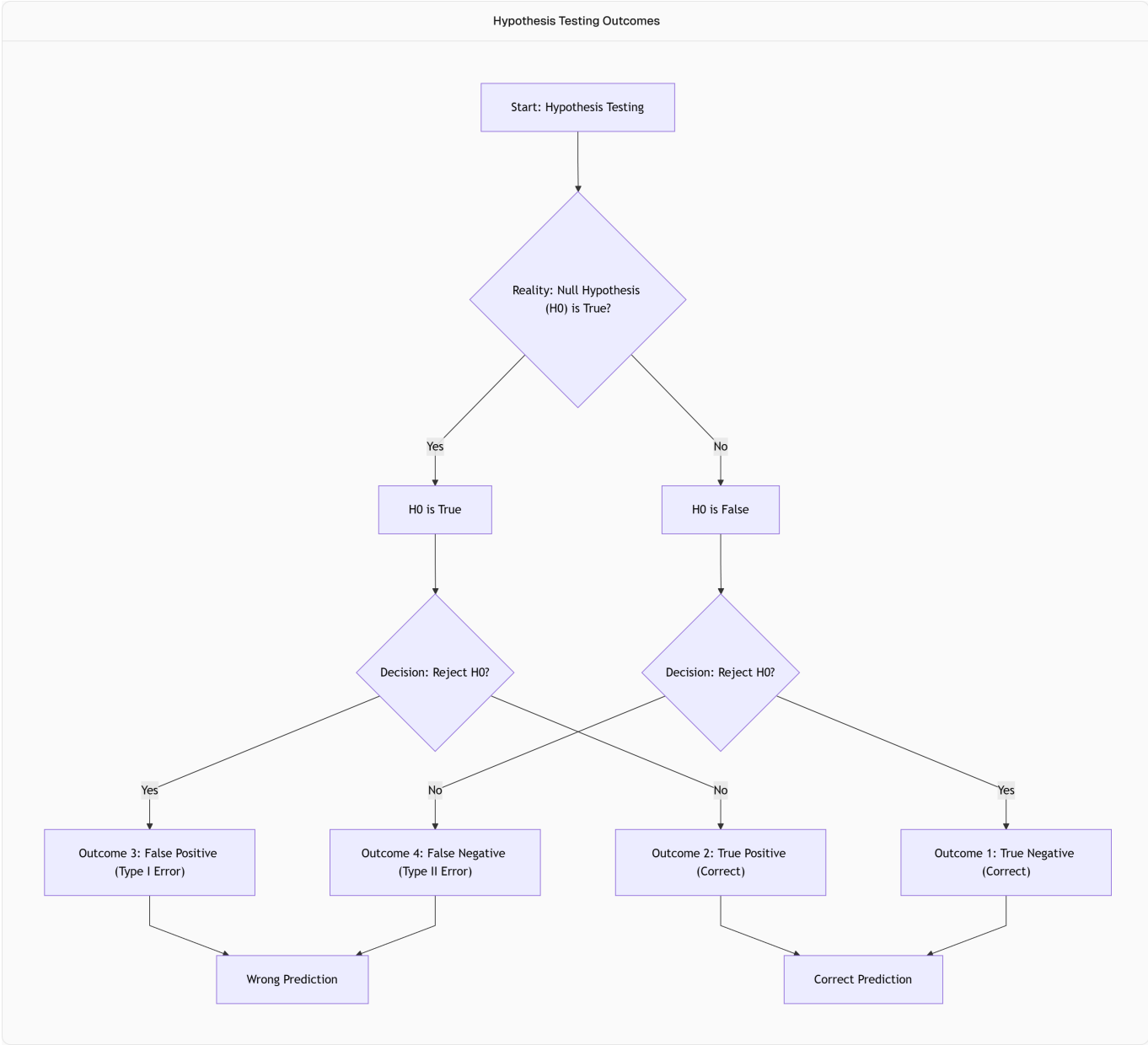
Understanding Type I and Type II Errors

- **False Positive (FP)** is specifically known as a **Type I error**.
    - This occurs when the **actual value is zero**, but the model **predicts one**.
    - It represents a **misclassification** where a null hypothesis is rejected when it is actually true.
- **False Negative (FN)** is specifically known as a **Type II error**.
    - This occurs when the **actual value is one**, but the model **predicts zero**.
    - It represents a scenario where the null hypothesis is failed to be rejected when it is actually false.
- Both Type I and Type II errors are misclassifications that need to be reduced in any model.

Hypothesis Testing Outcomes and Error Relation

- Hypothesis testing involves comparing a **null hypothesis** (H0) against an **alternate hypothesis** (H1).
- The outcomes of hypothesis testing are based on the **reality** of the null hypothesis (true or false) and the **decision** made (fail to reject or reject the null hypothesis).

Here is a flowchart illustrating the four possible outcomes in hypothesis testing:



- **Outcome 1: True Negative**
  - Occurs when the **null hypothesis is rejected** and in **reality, it is false**.
  - This is a **correct classification**, indicating the model or test performed well.
- **Outcome 2: True Positive**
  - Occurs when we **fail to reject the null hypothesis** and in **reality, it is true**.
  - This is also a **correct classification**, aligning with desired model behavior.
- **Outcome 3: False Positive (Type I Error)**

    - Occurs when the **null hypothesis is rejected**, but in **reality, it is true**.
    - This is a **wrong prediction** or **misclassification**, leading to an incorrect decision.
- **Outcome 4: False Negative (Type II Error)**
    - Occurs when we **fail to reject the null hypothesis**, but in **reality, it is false**.
    - This is also a **wrong prediction**, where a true condition is missed.

Importance of Error Reduction

- Reducing **Type I** and **Type II errors** is crucial for any machine learning model or statistical hypothesis testing.
- Understanding these errors is vital for statisticians and is frequently asked in interviews.