***Lecture content created by: Sulaiman Ahmed.***
***Sharing the content without proper credit is prohibited.***
***email: sulaimanahmed013@gmail.com***
***LinkedIn: www.linkedin.com/in/sulaimanahmed***
***Website: www.sulaimanahmed013.wixsite.com/my-site***

# Table of Contents: Understanding Conditional Probability, Bayes' Theorem, and P-value

### I. Introduction to Conditional Probability

- Steps for Calculating Conditional Probability from a Contingency Table
- Important Considerations

**II. Python Implementation: Titanic Dataset**

1. **Dataset and Problem**
   - Dataset Source: Titanic - Machine Learning from Disaster | Kaggle
   - Problem Statement: Survival Probability Based on Passenger Class
2. **Python Code**
   - Importing Libraries
   - Loading the Dataset
   - Creating the Contingency Table
   - Calculating Conditional Probabilities
3. **Explanation**
   - Code Breakdown and Functionality
4. **Output**
   - Contingency Table Output
   - Conditional Probability Results
5. **Interpretation**
   - Analyzing the Findings

**III. Deep Dive: The Seattle Example**

1. **Setting Up the Contingency Table**
   - Representing Weather Data
2. **Calculating Conditional Probability (P(A|B))**
   - Defining Events and the Question
   - Using the Contingency Table for Calculation
3. **Bayes' Theorem**

- The Reverse Question
- Formula: P(B|A) = [P(A|B) * P(B)] / P(A)
- Breaking Down the Formula and Defining Terms
- Applying Bayes' Theorem to the Example

4. **Key Points**
    - Reversing Conditional Probability with Bayes' Theorem
    - Using Contingency Tables for Visualization and Calculation

**IV. Python Implementation: Seattle Example**

1. **Code Walkthrough**
    - Defining Events and Probabilities
    - Calculating Conditional Probability
    - Calculating Bayes' Theorem
    - Outputting Results
2. **Output and Interpretation**

**V. Hypothesis Testing, Confidence Intervals, and Significance Values**

1. **Testing for a Fair Coin: Problem and Intuition**
    - The Coin Flip Example
    - Identifying the Need for Statistical Testing
2. **Hypothesis Testing**
    - Steps in Hypothesis Testing
        - Defining the Null Hypothesis (H0)
        - Defining the Alternative Hypothesis (H1)
        - Performing the Experiment
        - Making a Decision
    - Importance of Not "Accepting" the Null Hypothesis

3. **Significance Value (Alpha) and Confidence Intervals**
    - Significance Value (Alpha)
    - Confidence Interval
    - Relating Alpha, Confidence Interval, and Decision Making (Table)
    - Example and Decision-Making Process
4. **Interpreting Results and Making Decisions**
    - Meaning of Rejecting the Null Hypothesis
    - Meaning of Failing to Reject the Null Hypothesis
5. **Additional Concepts**
    - Type 1 Error (False Positive)
    - Type 2 Error (False Negative) (Table)
    - One-Tailed Test
    - Two-Tailed Test

**VI. Demystifying P-value**

1. **The Mousepad Analogy**
    - Visualizing Probability Distribution
    - High vs. Low Touch Probability
2. **What the P-value Represents**
    - Relating P-value to Probability and the Null Hypothesis
    - Mousepad Examples: High vs. Low P-value
3. **P-value and Decision Making in Hypothesis Testing**
    - Setting a Significance Level (Alpha)
    - Calculating the P-value
    - Comparing P-value to Alpha and Making Decisions
4. **In Summary**
    - The P-value as Evidence Against the Null Hypothesis

- Interpreting P-value in Context

**VII. Conclusion**

- Recap of Key Concepts
- Importance of Understanding Conditional Probability, Bayes' Theorem, and P-value in Data Analysis and Decision Making

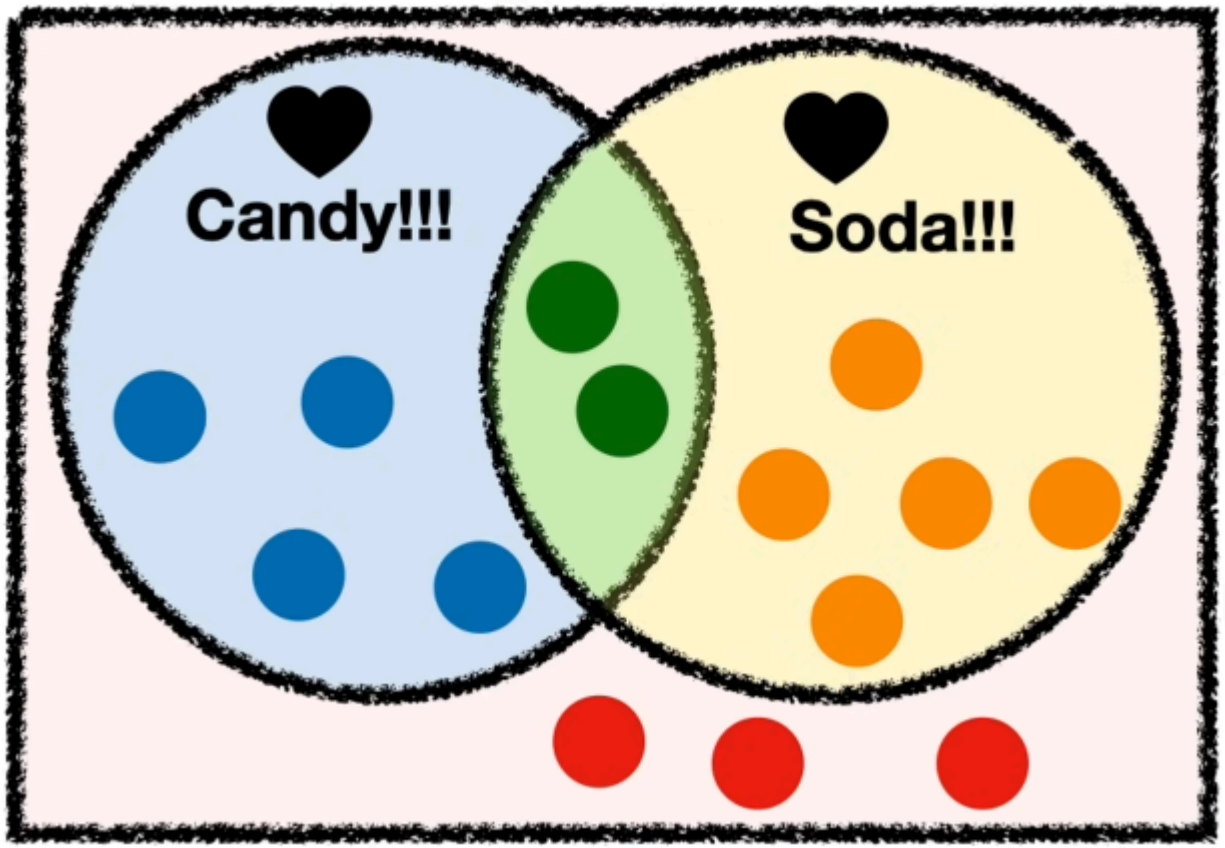
# A Detailed Guide to Conditional Probability Using Contingency Tables

This guide will walk you through understanding and applying conditional probability using contingency tables, illustrated by the example you provided.

**1. Understanding the Basics**

- **Contingency Table:** A contingency table, also known as a two-way table, summarizes data for two categorical variables. Each cell in the table represents the frequency (count) of individuals belonging to a specific combination of categories.
- **Probability:** Probability measures the likelihood of an event occurring. It is expressed as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.
- **Conditional Probability:** Conditional probability is the probability of an event happening *given* that another event has already occurred. It focuses on how the occurrence of one event influences the likelihood of another.

**2. Analyzing the Provided Example**




Let's break down the information presented in your images:

- **Variables:** The two categorical variables are:
  - **Love for Candy:** With categories "Loves Candy" and "Doesn't Love Candy".
  - **Love for Soda:** With categories "Loves Soda" and "Doesn't Love Soda".
- **Contingency Table:** The table summarizes the preferences of 14 individuals:

| | **Loves Candy** | **Doesn't Love Candy** | **Row Total** |
|---|---|---|---|
| Loves Soda | 2 | 5 | 7 |
| Doesn't Love Soda | 4 | 3 | 7 |

| | Loves Candy | Doesn't Love Candy | Row Total |
|---|---|---|---|
| Column Total | 6 | 8 | 14 |

**3. Calculating Conditional Probability**

The formula for conditional probability is:

$P(A|B) = P(A \text{ and } B) / P(B)$

where:

- $P(A|B)$ is the probability of event A happening given that event B has already occurred.
- $P(A \text{ and } B)$ is the probability of both events A and B happening.
- $P(B)$ is the probability of event B happening.

**4. Applying to the Example**

Let's calculate the conditional probability of a person loving candy given that they love soda:

- **Event A:** Loves Candy
- **Event B:** Loves Soda

1. **Find P(A and B):**
   - This is the probability of someone loving both candy and soda.
   - From the table, 2 people love both.
   - $P(A \text{ and } B) = 2/14$
2. **Find P(B):**
   - This is the probability of someone loving soda.
   - From the table, 7 people love soda.
   - $P(B) = 7/14$

3. **Calculate P(A|B):**
    - P(A|B) = P(A and B) / P(B)
    - P(A|B) = (2/14) / (7/14) = 2/7

**Therefore, the probability of a person loving candy given that they love soda is 2/7.**

**5. Generalizing the Approach**

You can apply these steps to calculate any conditional probability from the contingency table by:

1. Identifying the events of interest (A and B).
2. Finding the corresponding cell in the table for P(A and B).
3. Finding the row or column total corresponding to P(B).
4. Applying the conditional probability formula.

**Remember:**

- The event you are conditioning on (event B) will determine which row or column total you use for P(B).
- Always refer back to the table and the definitions of your events to ensure you are using the correct values.

# Python Implementation of Contingency Table and Conditional Probability

This example uses the "Titanic - Machine Learning from Disaster" dataset from Kaggle to demonstrate contingency tables and conditional probability in Python.

**1. Dataset and Problem:**

- **Dataset:** https://www.kaggle.com/competitions/titanic
- **Problem:** We'll explore the relationship between passenger class ('Pclass') and survival ('Survived') on the Titanic. We'll calculate the conditional probability of survival given a passenger's class.

### 2. Python Code:

```
import pandas as pd

# Load the Titanic dataset
df = pd.read_csv('titanic.csv')

# Create a contingency table for 'Pclass' and 'Survived'
contingency_table = pd.crosstab(df['Pclass'], df['Survived'])
print("Contingency Table:\n", contingency_table)

# Calculate conditional probability of survival given each passenger class
for pclass in range(1, 4):
    # Probability of survival given the passenger class
    p_survival_given_class = contingency_table.loc[pclass, 1] / contingency_table.loc[pclass].sum()

    print(f"P(Survival = 1 | Pclass = {pclass}) = {p_survival_given_class:.2f}")
```

### 3. Explanation:

- **Import pandas:** We import the pandas library for data manipulation.
- **Load the dataset:** We load the Titanic dataset from a local CSV file named 'titanic.csv'. Make sure to download the dataset from the provided Kaggle link and place it in the same directory as your Python script.
- **Create contingency table:** `pd.crosstab()` creates a contingency table showing the frequency of each combination of 'Pclass' and 'Survived'.
- **Calculate conditional probability:**
  - We iterate through each passenger class (1, 2, and 3).
  - `contingency_table.loc[pclass, 1]` retrieves the number of survivors (Survived = 1) for the given passenger class.
  - `contingency_table.loc[pclass].sum()` calculates the total number of passengers in that class.
  - We divide these values to get the conditional probability of survival given the passenger class.

**4. Output:**

The code will output the contingency table and the conditional probabilities:

```
Contingency Table:
 Survived    0    1
Pclass
1           80  136
2           97   87
3          372  119

P(Survival = 1 | Pclass = 1) = 0.63
P(Survival = 1 | Pclass = 2) = 0.47
P(Survival = 1 | Pclass = 3) = 0.24
```

**Interpretation:**

The results show that passengers in first class (Pclass = 1) had a higher probability of survival (63%) compared to passengers in second (47%) and third class (24%). This highlights how conditional probability can reveal relationships between variables.

Let's break down the Seattle example and Bayes' Theorem step-by-step, relating them to contingency tables.

# The Seattle Example:

**1. Setting up the Contingency Table:**

We can represent the Seattle weather information using a contingency table:

| | Rainy Tomorrow (A) | Not Rainy Tomorrow (not A) | Row Total |
|---|---|---|---|
| Cloudy Today (B) | 182 | 44 | 226 |
| Not Cloudy Today (not B) | 74 | 65 | 139 |
| Column Total | 256 | 109 | 365 |

**2. Calculating Conditional Probability (P(A|B)):**

- **Question:** Given that it is cloudy today (B), what is the probability that it will rain tomorrow (A)?
- **Focus:** We focus on the row where "Cloudy Today (B)" is true (the first row).
- **Calculation:**
  - P(A and B) = 182/365 (Probability of both cloudy today and rainy tomorrow)
  - P(B) = 226/365 (Probability of cloudy today)
  - P(A|B) = P(A and B) / P(B) = (182/365) / (226/365) = 182/226 = 0.80 or 80%

**Therefore, given that it is cloudy today, there is an 80% chance it will rain tomorrow.**

# Bayes' Theorem:

**1. The Reverse Question:**

Now, we're asking: Given that it rained today (A), what is the probability that yesterday was cloudy (B)?

**2. Bayes' Theorem Formula:**

P(B|A) = [P(A|B) * P(B)] / P(A)

**3. Breaking Down the Formula:**

- **P(B|A):** The probability of event B (cloudy yesterday) given that event A (rain today) has occurred. This is what we want to find.
- **P(A|B):** The probability of event A (rain today) given that event B (cloudy yesterday) has occurred. We know this is 0.50 or 50% from the original problem statement (cloudy days are followed by rain 50% of the time).
- **P(B):** The prior probability of event B (cloudy yesterday). This is the general probability of a cloudy day in Seattle, which is 226/365 or approximately 62%.
- **P(A):** The prior probability of event A (rain today). This is the general probability of a rainy day in Seattle, which is 256/365 or approximately 43%.

**4. Calculation:**

P(B|A) = (0.50 * 0.62) / 0.43 = 0.72 or 72%

**Therefore, given that it rained today, there is a 72% probability that yesterday was cloudy in Seattle.**

**Key Points:**

- **Bayes' Theorem allows us to "reverse" the conditioning in conditional probability.** We can infer the probability of a previous event given a known outcome.
- **Contingency tables provide a visual and organized way to represent the probabilities needed for both conditional probability and Bayes' Theorem calculations.**

Here's a Python implementation of the Seattle weather example, demonstrating both conditional probability and Bayes' Theorem calculations:

```python
# Define the events and their probabilities
cloudy_days = 226 / 365  # P(B) - Probability of a cloudy day
rainy_days = 256 / 365  # P(A) - Probability of a rainy day
cloudy_then_rainy = 182 / 365  # P(A and B) - Probability of cloudy then rainy

# Conditional probability: P(Rain tomorrow | Cloudy today)
prob_rain_given_cloudy = cloudy_then_rainy / cloudy_days

# Bayes' Theorem: P(Cloudy yesterday | Rain today)
prob_cloudy_given_rain = (cloudy_then_rainy / rainy_days) * (cloudy_days / rainy_days)

# Output the results
print(f"Probability of rain tomorrow given cloudy today: {prob_rain_given_cloudy:.2f}")
print(f"Probability of cloudy yesterday given rain today: {prob_cloudy_given_rain:.2f}")
```

**Explanation:**

1. **Define events and probabilities:**
   - We directly define the probabilities of each event based on the information provided in the Seattle example.
   - These probabilities represent the values from the contingency table.
2. **Calculate conditional probability:**
   - We calculate `prob_rain_given_cloudy` using the formula `P(A|B) = P(A and B) / P(B)`.
3. **Calculate Bayes' Theorem:**
   - We calculate `prob_cloudy_given_rain` using the formula `P(B|A) = [P(A|B) * P(B)] / P(A)`.
4. **Output results:**
   - The code prints the calculated probabilities, formatted to two decimal places.

**Output:**

```
Probability of rain tomorrow given cloudy today: 0.80
Probability of cloudy yesterday given rain today: 0.72
```

# Testing for a Fair Coin: A Deep Dive into Hypothesis Testing, Confidence Intervals, and Significance Values

Using the relatable example of determining whether a coin is fair. We'll cover:

1. **The Problem and Intuition**
2. **Hypothesis Testing**
3. **Significance Value (Alpha) and Confidence Intervals**
4. **Interpreting Results and Making Decisions**
5. **Additional Concepts (Type 1 & 2 Errors, One-Tailed vs. Two-Tailed Tests)**

## 1. The Problem and Intuition

Imagine you have a coin, and you want to determine if it's fair. Intuitively, a fair coin should land on heads roughly 50% of the time when flipped. But how can we test this statistically?

**Example:** You decide to flip the coin 100 times.

- **Ideal Scenario:** If you get exactly 50 heads, you might confidently declare the coin fair.
- **Reality:** It's unlikely to get *exactly* 50 heads. What if you get 48? Or 53? When do we start suspecting the coin might be biased?

This is where hypothesis testing comes in.

# 2. Hypothesis Testing

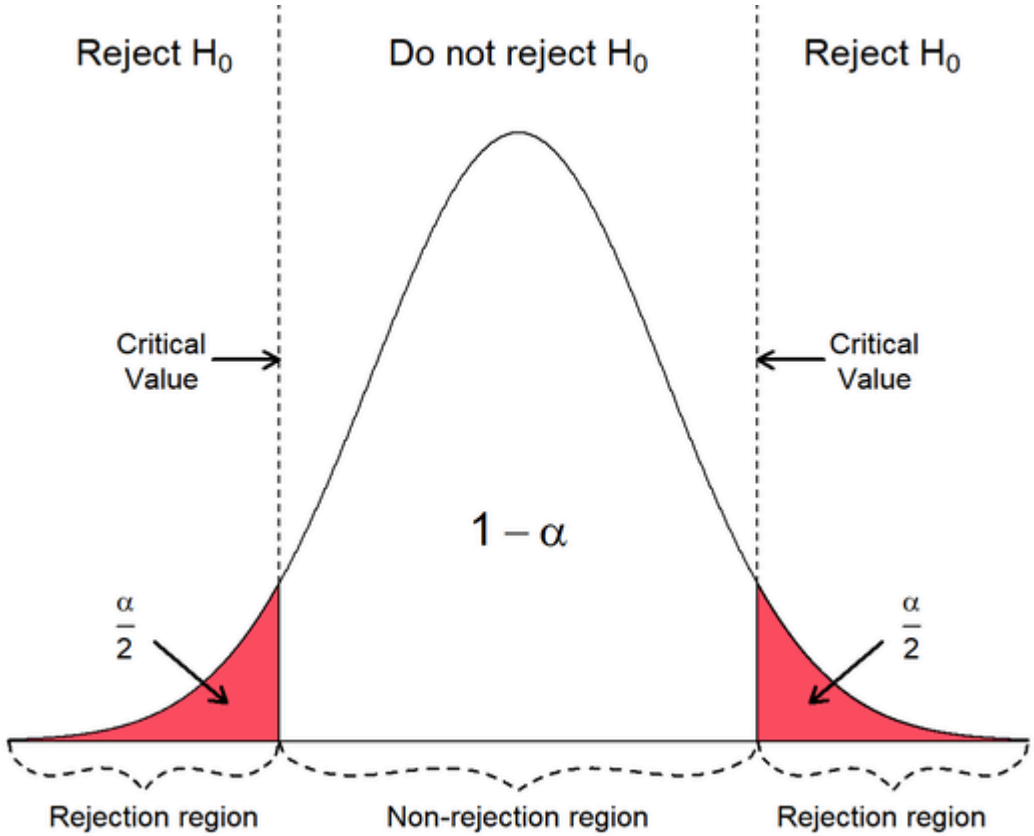
Hypothesis testing provides a structured framework for using data to make inferences about a population. In our case, the "population" is all possible flips of this coin.

**Steps in Hypothesis Testing:**

1. **Define the Null Hypothesis (H0):** The null hypothesis is a statement of "no effect" or "no difference." It's what we're trying to disprove. In our coin example:
    - **H0: The coin is fair (probability of heads = 0.5).**
2. **Define the Alternative Hypothesis (H1):** This is the opposite of the null hypothesis. It's what we're trying to find evidence for.
    - **H1: The coin is not fair (probability of heads ≠ 0.5).**
3. **Perform the Experiment:** Flip the coin 100 times and record the number of heads.
4. **Make a Decision:** Based on the experiment's results, we decide whether to:
    - **Reject the null hypothesis:** We have enough evidence to conclude the coin is likely not fair.
    - **Fail to reject the null hypothesis:** We don't have enough evidence to say the coin is unfair.

**Important Note:** We never "accept" the null hypothesis. We can only find evidence against it or fail to find such evidence.

# 3. Significance Value (Alpha) and Confidence Intervals




To make a decision in hypothesis testing, we need to define how much evidence is "enough" to reject the null hypothesis. This is where the significance value (alpha) comes in.

**Significance Value (Alpha):**

- Typically denoted by α.
- A small probability value (commonly 0.05 or 5%).
- Represents the probability of rejecting the null hypothesis when it's actually true (Type 1 error - more on this later).

**Confidence Interval:**

- Calculated as (1 - alpha) * 100%.
- If alpha = 0.05, the confidence interval is 95%.
- Represents the range within which we are confident the true population parameter (in our case, the probability of heads) lies.

**Relating Alpha, Confidence Interval, and Decision Making:**

| Significance Level (α) | Confidence Interval | Critical Region | Decision if Result Falls in... |
|---|---|---|---|
| 0.05 | 95% | Outer 5% of the distribution | Critical Region: Reject H0 |
| | | Within the central 95% | Fail to Reject H0 |

**Example:**

Let's say alpha = 0.05 (95% confidence interval). We perform our coin flip experiment and get 30 heads.

1. **Critical Region:** The critical region represents the extreme 5% of our expected results if the coin were fair. This would be a range of heads counts significantly far from 50.
2. **Comparison:** 30 heads falls far outside the expected range for a fair coin (the critical region).
3. **Decision:** We reject the null hypothesis (the coin is fair).

# 4. Interpreting Results and Making Decisions

- **Rejecting the null hypothesis:** Does not mean the null hypothesis is definitely false. It means the data provides strong evidence against it.
- **Failing to reject the null hypothesis:** Does not mean the null hypothesis is true. It means we don't have enough evidence to reject it.

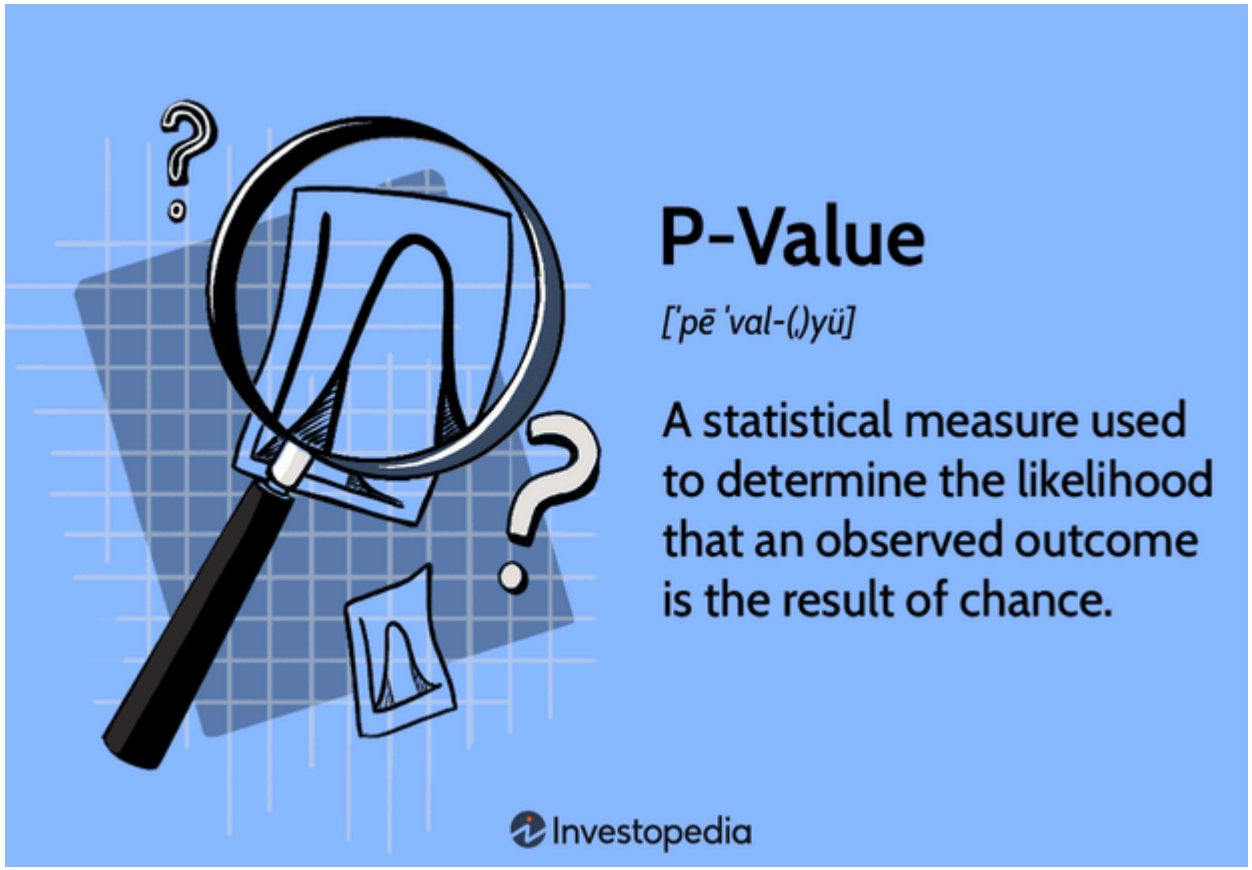
# 5. Additional Concepts

- **Type 1 Error:** Rejecting the null hypothesis when it's actually true (false positive).
- **Type 2 Error:** Failing to reject the null hypothesis when it's actually false (false negative).

| | |
|---|---|
| **Type 1 Error (False Positive)** | **Type 2 Error (False Negative)** |
| Reject H0 when H0 is true | Fail to reject H0 when H0 is false |

- **One-Tailed Test:** Used when the alternative hypothesis is directional (e.g., the coin is biased towards heads).
- **Two-Tailed Test:** Used when the alternative hypothesis is non-directional (e.g., the coin is simply not fair, could be biased towards heads or tails).

# P-value



# Understanding the Concept

The p-value is a fundamental concept in hypothesis testing. It helps us determine the strength of evidence against a null hypothesis.

**What is a Null Hypothesis?**

Before we delve into the p-value, let's briefly discuss the null hypothesis. In hypothesis testing, we start with an assumption called the null hypothesis (denoted as H0). It typically states that there is no effect or no difference between groups.

**The Role of the P-value:**

The p-value comes into play when we collect data and want to see if the data provides enough evidence to reject the null hypothesis. It quantifies the probability of observing our data (or more extreme data) if the null hypothesis were true.

**Interpretation:**

- **Low P-value (typically less than or equal to 0.05):** This suggests that the observed data is unlikely to have occurred by chance alone if the null hypothesis were true. We have strong evidence against the null hypothesis, and we often reject it in favor of an alternative hypothesis.
- **High P-value (typically greater than 0.05):** This indicates that the observed data is reasonably likely to occur even if the null hypothesis were true. We don't have enough evidence to reject the null hypothesis.

**Analogy:**

Imagine you have a bag of 100 coins, and you suspect that it might contain more than just fair coins.

- **Null Hypothesis (H0):** The bag contains only fair coins.
- **Experiment:** You draw 10 coins randomly, and 9 of them come up heads.

This result is quite unusual if the bag contained only fair coins. The p-value would quantify how unusual this result is. A very low p-value would suggest that the bag likely contains some biased coins, leading you to reject the null hypothesis.

**Key Points:**

- The p-value is a probability, ranging from 0 to 1.
- A smaller p-value indicates stronger evidence against the null hypothesis.
- The choice of significance level (alpha, usually 0.05) is arbitrary but commonly used.
- The p-value does not tell us the probability of the null hypothesis being true or false. It only measures the strength of evidence against the null hypothesis based on the observed data.

***Lecture content created by: Sulaiman Ahmed.***
***Sharing the content without proper credit is prohibited.***
***email: sulaimanahmed013@gmail.com***
***LinkedIn: www.linkedin.com/in/sulaimanahmed***
***Website: www.sulaimanahmed013.wixsite.com/my-site***