

Statistics For Data Science

NOTES

HARSH CHOURHARY

Statistics: The Foundation of Data Science

Statistics helps us collect, understand, and make sense of data. From spotting trends to making predictions, statistics gives us the tools to turn raw numbers into useful insights. In data science, whether you are building models or making decisions, statistics is there at every step. Learning statistics is the first step to thinking clearly and solving problems with data.

Basic Statistical Terms

1. Data: Data refers to facts, numbers, or observations collected for analysis. It can be anything from customer purchase records to temperature readings. Data is the raw material that statisticians and data scientists work with to uncover patterns and insights.

2. Variable : Variables are the building blocks of statistical analysis. They help us define what we're measuring and how we'll analyze it. Variables are classified into two main types:

- **Quantitative Variables:** Numerical data that can be measured (e.g., age, income, temperature).
- **Qualitative Variables:** Categorical data that describes qualities (e.g., gender, color, product type).

3. Population: Complete set of individuals, objects, or data points of interest in a study.

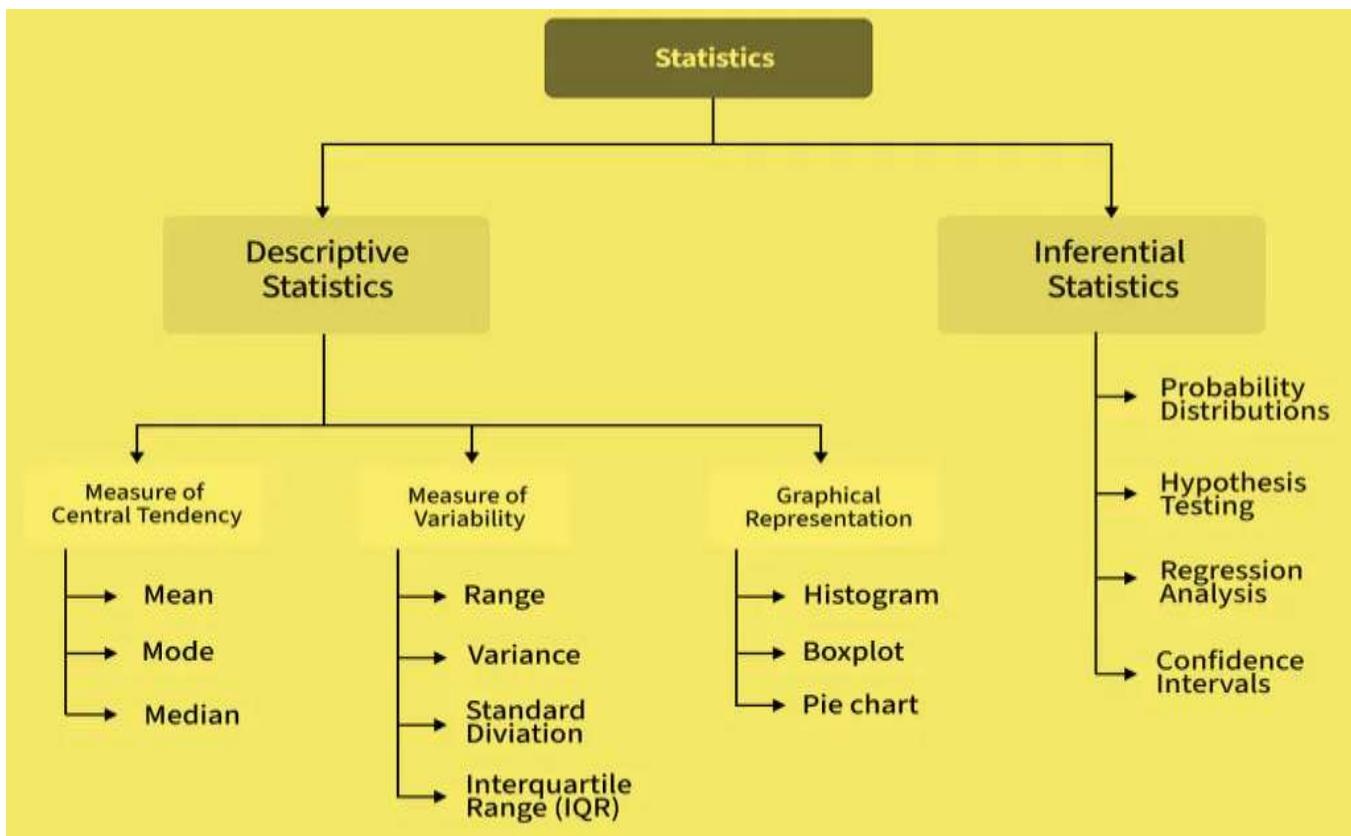
4. Sample : Subset of the population selected for analysis. It's used when studying the entire population is impossible or unnecessary. For instance, instead of measuring the height of every adult in a country, you might measure the height of 1,000 adults and use that data to infer information about the entire population.

5. Parameter: Numerical value that describes a characteristic of a population. For example, the average income of all households in a city is a parameter. Parameters are often unknown and are estimated using sample data.

6. Statistic: Numerical value that describes a characteristic of a sample. For example, the average income of 100 households surveyed in a city is a statistic. Statistics are used to estimate parameters and make inferences about populations.

Types of Statistics

Flow chart of type of statistics



1. Descriptive Statistics

Descriptive statistics summarize and describe the main features of a dataset. They provide simple summaries about the sample and help us understand the data's central tendency, variability, and distribution. Key measures include:

- **Measures of Central Tendency:** Mean, median, and mode.
- **Measures of Variability:** Range, variance, and standard deviation.
- **Measures of Frequency Distribution:** Histograms, frequency tables.

Descriptive statistics are essential for organizing and simplifying data, making it easier to interpret.

2. Inferential Statistics

Inferential statistics allow us to make predictions or inferences about a population based on sample data. They help us generalize findings from a sample to a larger population. Inferential statistics are crucial for drawing conclusions and making data-driven decisions.

Types of Data

1. Quantitative Data

Quantitative data consists of numerical values that can be measured. It is further divided into:

- **Discrete Data:** Countable values that cannot be divided into smaller parts (e.g., number of students in a class, number of cars in a parking lot).

- **Continuous Data:** Measurable values that can take any value within a range (e.g., height, weight, temperature).

2. Qualitative Data

Qualitative data describes qualities or characteristics and is non-numerical. It is further divided into:

- **Nominal Data:** Categories without any inherent order (e.g., gender, color, types of fruits).
- **Ordinal Data:** Categories with a meaningful order or ranking (e.g., education levels, customer satisfaction ratings).

Qualitative data is often used for categorization and is analyzed using frequency counts or percentages.

Levels of Measurement Explained

The level of measurement determines how data can be analyzed and what statistical techniques are appropriate. There are four levels:

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Four level of measurement

1. Nominal Level

Nominal data is the simplest level of measurement. It involves categorizing data into distinct groups or labels **without any order or ranking**. Examples include:

- Types of fruits (apple, banana, orange).
- Colors (red, blue, green).

Nominal data is analyzed using **frequency counts** (e.g., how many apples vs. bananas) or the **mode** (the most frequently occurring category).

2. Ordinal Level

Ordinal data builds on nominal data by introducing **order or ranking**. While the categories can be ranked, the differences between them are not measurable or meaningful. Examples include:

- Education levels (high school, bachelor's, master's).
- Customer satisfaction ratings (poor, fair, good, excellent).

Ordinal data can be summarized using the **median** (middle value) or **mode**, but not the mean (average), because the intervals between ranks are not consistent.

3. Interval Level

Interval data is numerical and the differences between values are meaningful. However, it lacks a **true zero point** meaning zero doesn't indicate the absence of the characteristic being measured. Examples include:

- Difference between 10°C and 20°C is the same as between 30°C and 40°C
- IQ scores.

Zero doesn't mean "none." For instance, 0°C doesn't mean the absence of temperature—it's just a point on the scale.

Interval data allows for addition and subtraction but not multiplication or division because the zero point is arbitrary.

4. Ratio Level

Ratio data is the most advanced level of measurement. It has all the properties of interval data, plus a **true zero point**, which allows for a full range of mathematical operations.

Zero indicates the complete absence of the characteristic being measured. For example, 0 kg means no weight, and 0 income means no earnings.

Examples include:

- Height, weight, income.
- Number of children in a family.

Ratio data allows for all mathematical operations, making it the most versatile level of measurement.

Summary Table for Clarity

Level of Measurement	Examples	Mathematical Operations
Nominal	Colours, types of fruits	Frequency counts, mode
Ordinal	Education levels, satisfaction ratings	Median, mode (no mean)
Interval	Temperature, IQ scores	Addition, subtraction
Ratio	Height, weight, income	All operations (+, -, ×, ÷)

How is Statistics Used in Data Analysis

Statistics is the backbone of data analysis as it transforms raw numbers into actionable business insights. Instead of making decisions based on gut feelings, statistics helps us summarize data, spot patterns and make predictions. It helps in:

- Summarize large datasets quickly (averages, percentages, trends)
- To compare groups or categories

- To spot outliers or trends in user behavior
- To make predictions or recommendations

Let's walk through a simple example.

Example: Predicting Customer Churn

A telecom company wants to find out why some customers are leaving and how to reduce it. Here's a small sample of the dataset:

CustomerID	MonthlyCharges	Tenure	Contract	Churn
1001	70	2	Month-to-Month	Yes
1002	35	30	One year	No
1003	55	10	Month-to-Month	Yes
1004	40	12	Month-to-Month	No
1005	80	1	Month-to-Month	Yes

Now, Let's Apply Statistics

1. Churn Rate

- Total customers = 5
- Churned = 3
- Churn Rate = $(3 / 5) \times 100 = 60\%$

2. Average Tenure of Churned Customers

- Tenure = 2, 10, 1
- Average = $(2 + 10 + 1) / 3 = 4.33$ months

3. Average Monthly Charges

- Churned: $(70 + 55 + 80)/3 = 68.33$
- Not Churned: $(35 + 40)/2 = 37.5$

4. Churn by Contract Type

Contract	Churned	Total	Churn Rate
Month-to-Month	3	4	75%
One year	0	1	0%

What Can We Infer from These Stats?

- Customers with **Month-to-Month** contracts are more likely to leave — 75% churn rate.

- People who leave usually do so **within the first few months** (average tenure is 4.33).
- Churned users have **higher monthly charges** than others.

From this, a data analyst can suggest actions like offering better rates to new customers or encouraging long-term contracts. These decisions are based on clear statistical evidence, not guesswork.

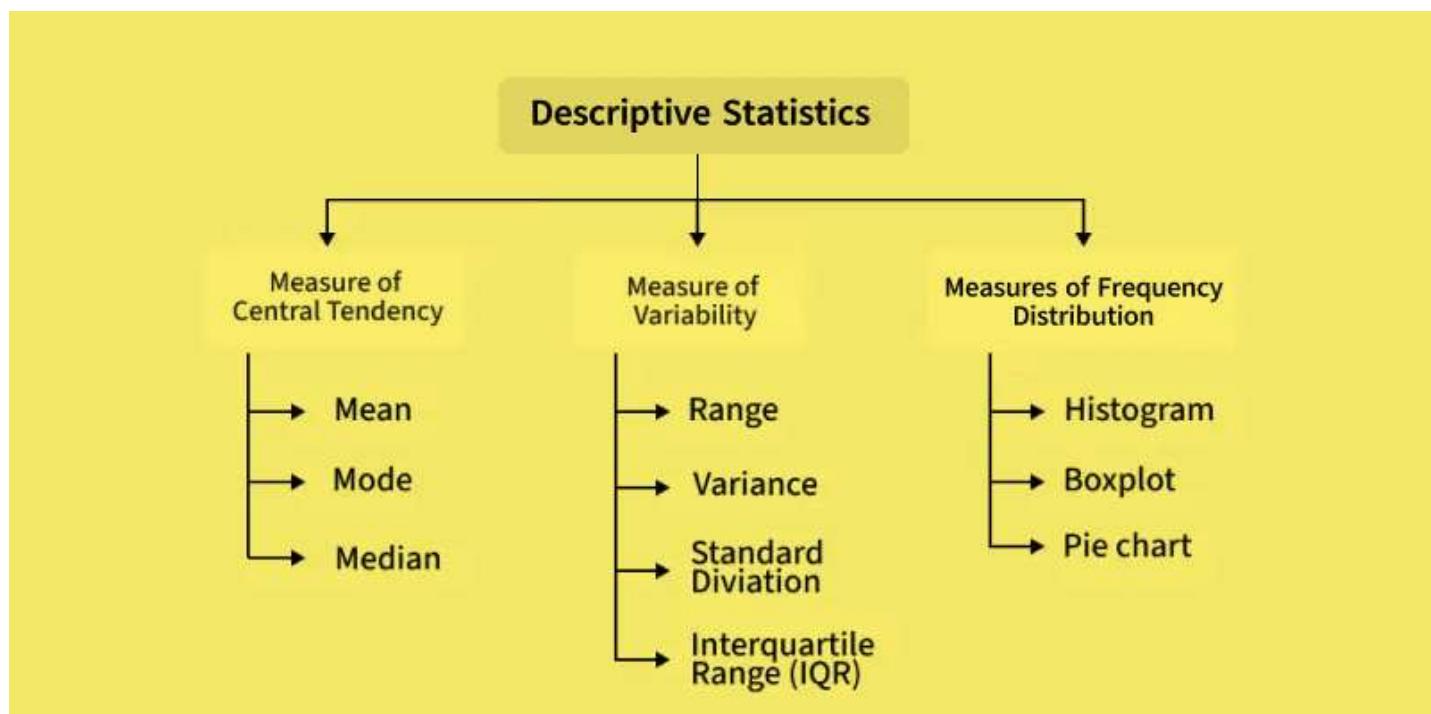
While applying these statistical methods in data analysis, we typically use Python libraries like **NumPy**, **Pandas**, **math** and **scipy** as they help us perform calculations, summarize data and handle tabular datasets efficiently.

Common Statistical Tools Used in Data Analysis

Tool/Concept	Use in Data Analysis
Mean, Median, Mode	Measure central tendency of data
Standard Deviation	Measure spread/variability
Percentages and Ratios	Compare parts of a whole
Correlation	Check relationships between two variables
Regression	Predict values and understand influence
Hypothesis Testing	Validate assumptions about data
Frequency Tables & Charts	Visualize distributions and categories

Descriptive Statistics

Statistics is the foundation of data science. Descriptive statistics are simple tools that help us understand and summarize data. They show the basic features of a dataset, like the average, highest and lowest values and how spread out the numbers are. It's the first step in making sense of information.



Descriptive Statistic

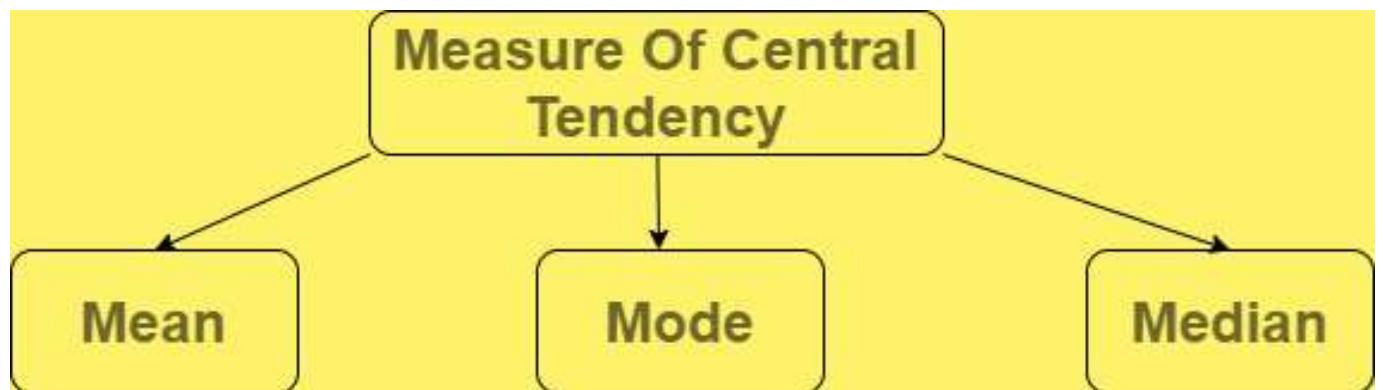
Types of Descriptive Statistics

There are three categories for standard classification of descriptive statistics methods, each serving different purposes in summarizing and describing data. They help us understand:

1. Where the data centers (Measures of Central Tendency)
2. How spread out the data is (Measure of Variability)
3. How the data is distributed (Measures of Frequency Distribution)

1. Measures of Central Tendency

Statistical values that **describe the central position within a dataset**. There are three main measures of central tendency:



Measures of Central Tendency

Mean: is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

$$\bar{x} = \frac{1}{n} \sum x$$

where,

- x = Observations
- n = number of terms

Let's look at an example of how can we find the mean of a data set using python code implementation. Before its implementation we should have some basic knowledge about numpy and scipy.

```
import numpy as np
```

```
# Sample Data  
arr = [5, 6, 11]
```

```
# Mean  
mean = np.mean(arr)
```

```
print("Mean = ", mean)
```

Output

```
Mean = 7.333333333333333
```

Mode: The most frequently occurring value in the dataset. It's useful for categorical data and in cases where knowing the most common choice is crucial.

```
import scipy.stats as stats
```

```
# sample Data  
arr = [1, 2, 2, 3]
```

```
# Mode  
mode = stats.mode(arr)  
print("Mode = ", mode)
```

Output:

```
Mode = ModeResult(mode=array([2]), count=array([2]))
```

Median: The median is the middle value in a sorted dataset. If the number of values is odd, it's the center value, if even, it's the average of the two middle values. It's often better than the mean for skewed data.

```
import numpy as np
```

```
# sample Data  
arr = [1, 2, 3, 4]
```

```
# Median  
median = np.median(arr)
```

```
print("Median = ", median)
```

Output

```
Median = 2.5
```

2. Measure of Variability

Knowing not just where the data centers but also how it spreads out is important. Measures of variability, also called measures of dispersion, help us spot the spread or distribution of observations in a dataset. They identifying outliers, assessing model assumptions and understanding data variability in relation to its mean. The key measures of variability include:

1. Range : describes the difference between the largest and smallest data point in our data set. The bigger the range, the more the spread of data and vice versa. While easy to compute **range is sensitive to outliers**. This measure can provide a quick sense of the data spread but should be complemented with other statistics.

```
Range = Largest data value - smallest data value
```

```
import numpy as np
```

```
# Sample Data  
arr = [1, 2, 3, 4, 5]
```

```
# Finding Max
```

```

Maximum = max(arr)
# Finding Min
Minimum = min(arr)

# Difference Of Max and Min
Range = Maximum-Minimum
print("Maximum = {}, Minimum = {} and Range = {}".format(
    Maximum, Minimum, Range))

```

Output

```
Maximum = 5, Minimum = 1 and Range = 4
```

2. Variance: is defined as an **average squared deviation from the mean**. It is calculated by finding the difference between every data point and the average which is also known as the mean, squaring them, adding all of them and then dividing by the number of data points present in our data set.

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

where,

- **x** -> Observation under consideration
- **N** -> number of terms
- **mu** -> Mean

```
import statistics
```

```
# sample data
```

```
arr = [1, 2, 3, 4, 5]
```

```
# variance
```

```
print("Var = ", (statistics.variance(arr)))
```

Output

```
Var = 2.5
```

3. Standard deviation: Standard deviation is widely used to measure the **extent of variation or dispersion in data**. It's especially important when assessing model performance (e.g., residuals) or comparing datasets with different means.

It is defined as the square root of the variance. It is calculated by finding the mean, then **subtracting each number from the mean which is also known as the average and squaring the result**. Adding all the values and then dividing by the no of terms followed by the square root.

$$\sigma = \sqrt{\frac{1}{N} \sum (x - \mu)^2}$$

where,

- **x** = Observation under consideration
- **N** = number of terms
- **mu** = Mean

```
import statistics
```

```
arr = [1, 2, 3, 4, 5]
```

```
print("Std = ", (statistics.stdev(arr)))
```

Output

Std = 1.5811388300841898

Variability measures are important in residual analysis to check how well a model fits the data.

3. Measures of Frequency Distribution

Frequency distribution table is a powerful summarize way to show **how data points are distributed across different categories or intervals**. Helps identify **patterns, outliers and the overall structure of the dataset**. It is often the first step in understand the dataset before applying more advanced analytical methods or creating visualizations like histograms or pie charts.

Frequency Distribution Table Includes measure like:

- Data intervals or categories
- Frequency counts
- Relative frequencies (percentages)
- Cumulative frequencies when needed

What are Inferential Statistics

Inferential statistics is an important tool that allows us to make predictions and conclusions about a population based on sample data. Unlike descriptive statistics, which only summarize data, inferential statistics let us test hypotheses, make estimates, and measure the uncertainty about our predictions. These tools are essential for evaluating models, testing assumptions, and supporting data-driven decision-making.

For example, instead of surveying every voter in a country, we can survey a few thousand and still make reliable conclusions about the entire population's opinion. Inferential statistics provides the tools to do this systematically and mathematically.

Why Do We Need Inferential Statistics?

In real-world scenarios, analyzing an entire population is often impossible. Instead, we collect data from a sample and use inferential statistics to:

- Conclude the whole population.
- Test claims or hypotheses.
- Calculate confidence intervals and p-values to measure uncertainty.
- Make predictions with statistical models.

Techniques in Inferential Statistics

Inferential statistics offers several key methods for testing hypotheses, estimating population parameters, and making predictions. Here are the major techniques:

1. Confidence Intervals: It gives us a range of values that likely includes the true population parameter. It helps quantify the uncertainty of an estimate. The formula for calculating a confidence interval for the mean is:

$$CI = \bar{x} \pm Z\alpha/2 \times n\sigma$$

Where:

- \bar{x} is the sample mean
- $Z\alpha/2$ is the Z-value from the standard normal distribution (e.g., 1.96 for a 95% confidence interval)
- σ is the population standard deviation
- n is the sample size

For example, if we measure the average height of 100 people, a 95% confidence interval gives us a range where the true population mean height is likely to fall. This helps gauge the precision of our estimate and compare models (like in A/B testing).

2. Hypothesis Testing: Hypothesis testing is a formal procedure for testing claims or assumptions about data. It involves the following steps:

- **Null Hypothesis (H_0):** The default assumption, such as “there’s no difference between two models.”
- **Alternative Hypothesis (H_1):** The claim you aim to prove, such as “Model A performs better than Model B.”

We collect data and compute a test statistic (such as Z for a Z-test or t for a T-test):

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Where:

- \bar{x} is the sample mean
- μ_0 is the hypothesized population mean
- σ is the population standard deviation
- n is the sample size

After calculating the test statistic, we compare it with a critical value or use a p-value to decide whether to reject or accept the null hypothesis. If the p-value is smaller than the significance level α (usually 0.05), we reject the null hypothesis.

$$p\text{-value} = 2 \cdot P(Z > |z_{obs}|)$$

Where z_{obs} is the observed test statistic? A small p-value suggests strong evidence against the null hypothesis.

3. Central Limit Theorem: It states that the distribution of the sample mean will approximate a normal distribution as the sample size increases, regardless of the original population distribution. This is crucial because many statistical methods assume that data is normally distributed. The CLT can be mathematically expressed as:

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

Where:

- μ is the population mean
- σ is the population standard deviation

- n is the sample size

This theorem allows us to apply normal distribution-based methods even when the original data is not normally distributed, such as in cases with skewed income or shopping behavior data.

Errors in Inferential Statistics

In hypothesis testing, Type I Error and Type II Error are key concepts:

- **Type I Error** occurs when we wrongly reject a true null hypothesis. The probability of making a Type I error is denoted by α (the significance level).
- **Type II Error** occurs when we fail to reject a false null hypothesis. The probability of making a Type II error is denoted by β and the power of the test is given by $1-\beta$.

The goal is to minimize these errors by carefully selecting sample sizes and significance levels.

Parametric and Non-Parametric Tests

Statistical tests help decide if the data support a hypothesis. They calculate a test statistic that shows how much the data differs from the assumption (null hypothesis). This is compared to a critical value or p-value to accept or reject the null.

1. **Parametric Tests:** These tests assume that the data follows a specific distribution (often normal) and has consistent variance. They are typically used for continuous data. Examples include the Z-test, T-test, and ANOVA. These tests are effective for comparing models or measuring performance when the assumptions are met.
2. **Non-Parametric Tests:** Non-parametric tests do not assume a specific distribution for the data, making them ideal for small samples or non-normal data, including categorical or ranked data. Examples include the Chi-Square test, Mann-Whitney U test, and Kruskal-Wallis test. They are useful when data is skewed or categorical, such as customer ratings or behaviors.

Example: Evaluating a New Delivery Algorithm Using Inferential Statistics

A quick commerce company wants to check if a new delivery algorithm reduces delivery times compared to the current system.

Experiment Setup:

- 100 orders split into two groups: 50 with the new algorithm, 50 with the current system.
- Delivery times for both groups are recorded.

Steps

Hypotheses:

- **Null (H_0):** The New algorithm does not reduce delivery time.
- **Alternative (H_1):** New algorithm reduces delivery time.

Significance Level:

Set at 0.05 (5% risk of wrongly rejecting H_0).

- **Type I error:** Thinking the new system is better when it isn't.

- **Type II error:** Missing a real improvement.

Test Statistic: Compare average delivery times between the two groups

Analysis:

- Calculate means and differences.
- Check if the data is roughly normal.

Perform a t-test or z-test.

If p-value < 0.05, reject H₀ and conclude the new algorithm is better. Otherwise, no clear improvement.

Confidence Interval: For example, a range of -5 to -2 minutes means deliveries are 2 to 5 minutes faster with the new system.

Covariance and Correlation

Covariance and correlation are the two key concepts in Statistics that help us analyze the relationship between two variables. Covariance measures how two variables change together, indicating whether they move in the same or opposite directions.

Relationship between Independent and dependent variables

To understand this relationship better, consider factors like sunlight, water and soil nutrients (as shown in the image), which are independent variables that influence plant growth our dependent variable. Covariance measures how these variables change together, indicating whether they move in the same or opposite directions.

What is Covariance?

Covariance is a statistical which measures the relationship between a pair of random variables where a change in one variable causes a change in another variable. It assesses how much two variables change together from their mean values. Covariance is calculated by taking the average of the product of the deviations of each variable from their respective means. Covariance helps us understand the direction of the relationship but not how strong it is because the number depends on the units used. It's an important tool to see how two things are connected.

1. It can take any value between - infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
2. It is used for the linear relationship between variables.
3. It gives the direction of relationship between variables.

Covariance Formula

1. Sample Covariance

$$CovS(X, Y) = n - 1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Where:

- X_i : The i^{th} value of the variable X in the sample.
- Y_i : The i^{th} value of the variable Y in the sample.

- X : The sample mean of variable X (i.e., the average of all X_i values in the sample).
- Y : The sample mean of variable Y (i.e., the average of all Y_i values in the sample).
- n : The number of data points in the sample.
- Σ : The summation symbol means we sum the products of the deviations for all the data points.
- $n-1$: This is the degrees of freedom. When working with a sample, we divide by $n-1$ to correct for the bias introduced by estimating the population covariance based on the sample data. This is known as Bessel's correction.

2. Population Covariance

$$\text{Cov}_P(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_X)(Y_i - \mu_Y)$$

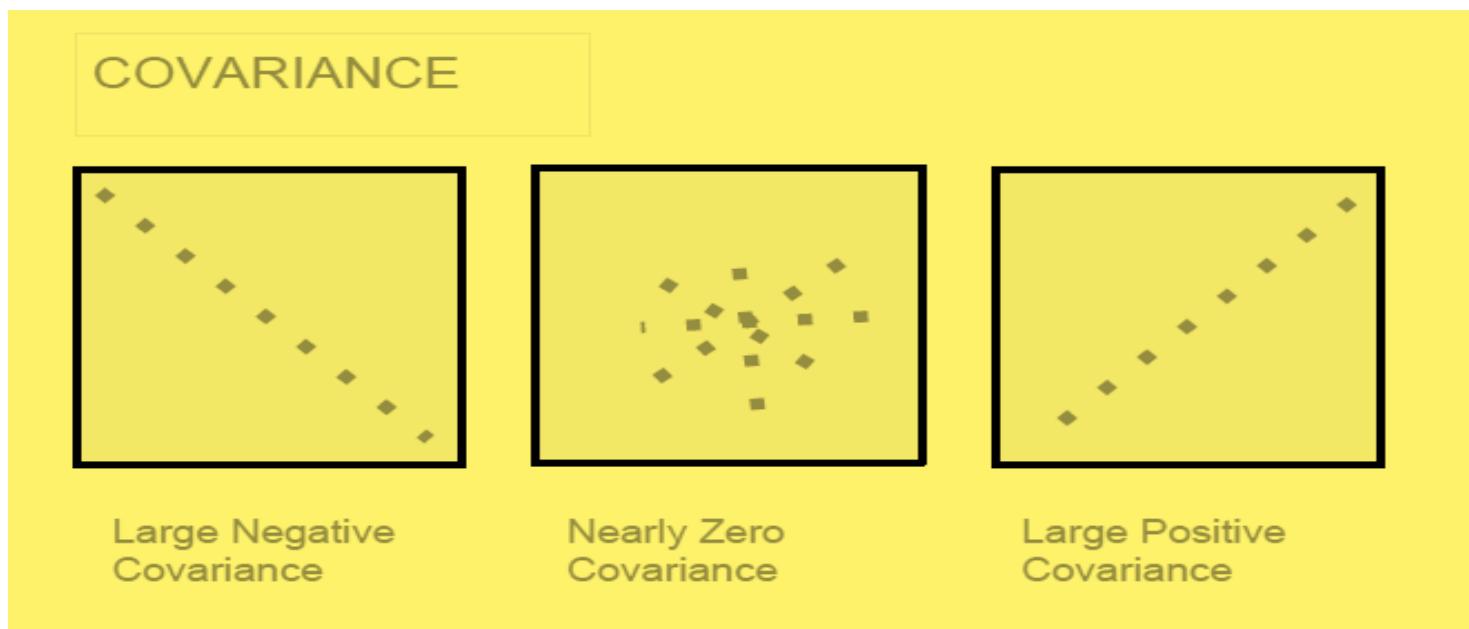
Where:

- X_i : The i th value of the variable X in the population.
- Y_i : The i th value of the variable Y in the population.
- μ_X : The population mean of variable X (i.e., the average of all X_i values in the population).
- μ_Y : The population mean of variable Y (i.e., the average of all Y_i values in the population).
- n : The total number of data points in the population.
- Σ : The summation symbol means we sum the products of the deviations for all the data points.
- n : In the case of population covariance, we divide by n because we are using the entire population data. There's no need for Bessel's correction since we're not estimating anything.

Types of Covariance

- **Positive Covariance**: When one variable increases, the other variable tends to increase as well and vice versa.
- **Negative Covariance**: When one variable increases, the other variable tends to decrease.
- **Zero Covariance**: There is no linear relationship between the two variables; they move independently of each other.

Example



What is Correlation?

Correlation is a standardized measure of the strength and direction of the linear relationship between two variables. It is derived from covariance and ranges between -1 and 1. Unlike covariance, which only indicates the direction of the relationship, correlation provides a standardized measure.

- **Positive Correlation (close to +1):** As one variable increases, the other variable also tends to increase.
- **Negative Correlation (close to -1):** As one variable increases, the other variable tends to decrease.
- **Zero Correlation:** There is no linear relationship between the variables.

The correlation coefficient ρ (rho) for variables X and Y is defined as:

1. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
2. In this variable are indirectly related to each other.
3. It gives the direction and strength of relationship between variables.

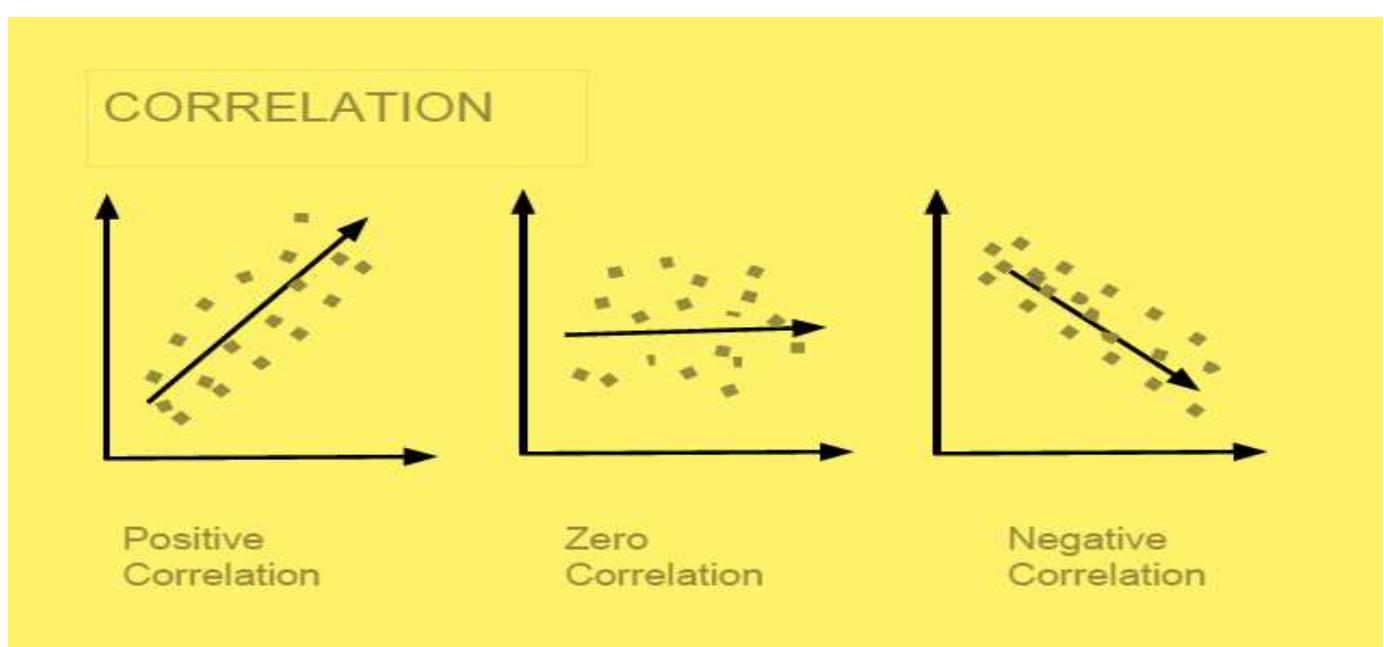
Correlation Formula

$$\text{Corr}(x, y) = \frac{\sum_{i=1}^n (x_i - x') (y_i - y')}{\sqrt{\sum_{i=1}^n (x_i - x')^2 \sum_{i=1}^n (y_i - y')^2}}$$

Here,

- x' and y' = mean of given sample set
- n = total no of sample
- x_i and y_i = individual sample of set

Example



Difference between Covariance and Correlation

This table shows the difference between Covariance and Correlation:

Covariance	Correlation
Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.
Involves the relationship between two variables or data sets	Involves the relationship between multiple variables as well
Lies between -infinity and +infinity	Lies between -1 and +1
Measure of correlation	Scaled version of covariance
Provides direction of relationship	Provides direction and strength of relationship
Dependent on scale of variable	Independent on scale of variable
Have dimensions	Dimensionless

The key difference is that Covariance shows the direction of the relationship between variables, while correlation shows both the direction and strength in a standardized form.

Applications of Covariance and Correlation

Applications of Covariance

- **Portfolio Management in Finance:** Covariance is used to measure how different stocks or financial assets move together, aiding in portfolio diversification to minimize risk.
- **Genetics:** In genetics, covariance can help understand the relationship between different genetic traits and how they vary together.
- **Econometrics:** Covariance is employed to study the relationship between different economic indicators, such as the relationship between GDP growth and inflation rates.
- **Signal Processing:** Covariance is used to analyze and filter signals in various forms, including audio and image signals.
- **Environmental Science:** Covariance is applied to study relationships between environmental variables, such as temperature and humidity changes over time.

Applications of Correlation

- **Market Research:** Correlation is used to identify relationships between consumer behavior and sales trends, helping businesses make informed marketing decisions.
- **Medical Research:** Correlation helps in understanding the relationship between different health indicators, such as the correlation between blood pressure and cholesterol levels.
- **Weather Forecasting:** Correlation is used to analyze the relationship between various meteorological variables, such as temperature and humidity, to improve weather predictions.
- **Machine Learning:** Correlation analysis is used in feature selection to identify which variables have strong relationships with the target variable, improving model accuracy.

Conditional Probability

Conditional probability defines the probability of an event occurring based on a given condition or prior knowledge of another event.

It is **the likelihood of an event occurring**, given that another event has already occurred. In probability, this is denoted as A given B, expressed as $P(A | B)$, indicating the probability of event A when the event B has already occurred.

Explanation of the above carousel:

Question: What are the chances that it's raining given that you carry an umbrella?

Given:

- It rains **30% of the time**
- You carry an umbrella **50% of the time**
- When it rains, you carry an umbrella **80% of the time**

Implies (for 10 days scenario):

- **3 out of 10 days** are rainy. (since 30% of 10 = 3)
- **5 out of 10 days** you carry an umbrella. (since 50% of 10 = 5)
- On rainy days, you carry it **2 out of 3 days** (since 80% of 3 ~ 2)

Total Umbrella Days:

- **Rainy days:** 2 out of 3
- **Sunny days:** 3 out of 7.
- **Total umbrella days = 5** (matching the 50% probability).

Conditional Probability:

Out of all umbrella days (5), only 2 days were rainy.
Thus, $P(\text{Rain} | \text{Carry Umbrella}) = 2/5 = 40\%$.

Conditional Probability Formula

Let's consider two events A and B, then the **formula for the conditional probability** of B when A has already occurred is given by:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Where,

- $P(A \cap B)$ represents the probability of both events A and B occurring simultaneously.
- $P(A)$ represents the probability of event A occurring.

Steps to Find Probability of One Event Given Another Has Already Occurred

To **calculate** the conditional probability, we can use the following step-by-step method:

- Step 1:** Identify the Events. Let's call them Event A and Event B.
- Step 2:** Determine the Probability of Event A i.e., $P(A)$
- Step 3:** Determine the Probability of Event B i.e., $P(B)$
- Step 4:** Determine the Probability of Event A and B i.e., $P(A \cap B)$.
- Step 5:** Apply the Conditional Probability Formula and calculate the required probability.

Conditional Probability Examples

There are various examples of conditional probability as in real life all the events are related to each other and happening any event affects the probability of another event. For **example**, if it rains, the probability of road accidents increases as roads have less friction. Let's consider some problem-based examples here:

1) Tossing a Coin

Let's consider two events in tossing two coins,

- A: Getting a head on the first coin.
- B: Getting a head on the second coin. Sample space for tossing two coins is:
 $S = \{\text{HH, HT, TH, TT}\}$

The conditional probability of getting a head on the second coin (B) given that we got a head on the first coin (A) is = $P(B|A)$.

Since the coins are independent (one coin's outcome does not affect the other), $P(B|A) = P(B) = 0.5$ (50%), which is the probability of getting a head on a single coin toss.

2) Drawing Cards

In a deck of 52 cards where two cards are being drawn, then let's consider the events be.

- A: Drawing a red card on the first draw, and
- B: Drawing a red card on the second draw.

The conditional probability of drawing a red card on the second draw (B) given that we drew a red card on the first draw (A) is = $P(B|A)$

After drawing a red card on the first draw, there are 25 red cards and 51 cards remaining in the deck. So, $P(B|A) = 25/51 \approx 0.49$ (approximately 49%).

Properties of Conditional Probability

Some of the common properties of conditional property are:

1: Let's consider an event A in any sample space S of an experiment. $P(S | A) = P(A | A) = 1$

2: For any two events A and B of a sample space S, and an event X such that $P(X) \neq 0$, $P((A \cup B) | X) = P(A | X) + P(B | X) - P((A \cap B) | X)$

3: The order of sets or events is important in conditional probability, i.e., $P(A | B) \neq P(B | A)$

4: The complement formula for probability only holds conditional probability if it is given in the context of the first argument in conditional probability i.e., $P(A' | B) = 1 - P(A | B)$ $P(A | B') \neq 1 - P(A | B)$

5: For any two or three independent events, the intersection of events can be calculated using the following formula:

- For the intersection of two events A and B $P(A \cap B) = P(A) P(B)$
- For the intersection of three events A, B, and C $P(A \cap B \cap C) = P(A) P(B) P(C)$

Conditional Probability and Independent Events

With the help of conditional probability, we can tell apart dependent and independent events. When the probability of one event happening doesn't influence the probability of any other event, then events are called independent, otherwise dependent events.

Conditional Probability of Independent Events

When two events are independent, those conditional probability is the same as the probability of the event individually i.e., $P(A | B)$ is the same as $P(A)$ as there is no effect of event B on the probability of event A. For independent events, A and B, the conditional probability of A and B concerning each other is given as follows:

- $P(B | A) = P(B)$
- $P(A | B) = P(A)$

Conditional Probability vs Joint Probability vs Marginal Probability

The difference between Conditional Probability, **Joint Probability**, and **Marginal Probability** is given in the following table:

Parameter	Conditional Probability	Joint Probability	Marginal Probability
Definition	The probability of an event occurring is given. That another event has already occurred.	The probability of two or more events occurring simultaneously.	The probability of an event occurring without considering any other events.
Calculation	$P(A B)$	$P(A \cap B)$	$P(A)$
Variables involved	Two or more events	Two or more events	Single event.

Note: Conditional probability is widely used for Bayes' theorem where we update probabilities based on new evidence, for more details you can refer to: [Bayes' Theorem](#)

Multiplication Rule of Probability

Multiplication Rule of Probability, when applied in the context of conditional probability, helps us calculate the probability of the intersection of two events when the probability of one event depends on the occurrence of the other event. **This rule is crucial in understanding the joint probability of events under specific conditions.**

In the context of conditional probability, the Multiplication Rule is often stated as follows:

$$P(A \cap B) = P(A) \times P(B | A)$$

Here's what each term represents:

- **P(A∩B)**: This denotes the probability that both events A and B occur simultaneously.
- **P(A)**: This represents the probability of event A happening.
- **P(B|A)**: This is the conditional probability of event B occurring given that event A has already occurred.

How to Apply the Multiplication Rule?

To apply the Multiplication Rule in the context of conditional probability, we can use the following steps:

- First, we calculate the probability of event A occurring.
- Then, we compute the probability of event B occurring given that event A has occurred.

- Multiplying these probabilities together gives us the joint probability of both events happening under the specified conditions.
- This rule is particularly useful when dealing with events that are not independent, meaning that the occurrence of one event affects the probability of the other event.

Applications of Conditional Probability

Various applications of conditional probability are,

Finance and Risk Management

- **Example:** Assessing the probability of default for a borrower given certain financial indicators.
- **Application:** Banks and financial institutions use conditional probability to evaluate the risk associated with loans and investments.

Healthcare and Diagnostics

- **Example:** Determining the probability of a patient having a specific disease given the results of diagnostic tests.
- **Application:** Conditional probability is crucial in medical diagnoses and decision-making, helping healthcare professionals make informed decisions based on test results.

Marketing and Customer Relationship Management (CRM)

- **Example:** Predicting the probability of a customer making a purchase based on their past buying behavior.
- **Application:** Businesses use conditional probability to tailor marketing strategies, optimize customer experiences, and personalize product recommendations.

Machine Learning and Artificial Intelligence

- **Example:** Predicting the likelihood of a user clicking on a particular ad based on their online behavior.
- **Application:** Conditional probability is fundamental in machine learning algorithms for tasks such as classification, recommendation systems, and natural language processing.

Weather Forecasting

- **Example:** Estimating the probability of rain tomorrow given today's weather conditions.
- **Application:** Meteorologists use conditional probability to make weather predictions based on historical data and current atmospheric conditions.

Bayes' Theorem

Bayes' Theorem is a mathematical formula used to determine the **conditional probability** of an event based on prior knowledge and new evidence.

It adjusts probabilities when new information comes in and helps make better decisions in uncertain situations.

Bayes' Theorem helps us update probabilities based on prior knowledge and new evidence. In this case, knowing that the pet is quiet (new information), we can use Bayes' Theorem to calculate the updated probability of the pet being a cat or a dog, based on how likely each animal is to be quiet.

Bayes Theorem and Conditional Probability

Bayes' theorem (also known as the Bayes Rule or Bayes Law) is used to determine the conditional probability of event A when event B has already occurred.

The general statement of Bayes' theorem is "The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the probability of B, given A, and the probability of A divided by the probability of event B." i.e.

For example, if we want to find the probability that a white marble drawn at random came from the first bag, given that a white marble has already been drawn, and there are three bags each containing some white and black marbles, then we can use Bayes' Theorem.

Bayes Theorem Formula

For any two events A and B, **Bayes's** formula for the Bayes theorem is given by:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Formula for the Bayes theorem

Where,

- **P(A)** and **P(B)** are the probabilities of events A and B; also, **P(B)** is never equal to zero.
- **P(A|B)** is the probability of event A when event B happens,
- **P(B|A)** is the probability of event B when A happens.

Bayes Theorem Statement

Bayes' Theorem for n sets of events is defined as,

Let E_1, E_2, \dots, E_n be a set of events associated with the sample space S, in which all the events E_1, E_2, \dots, E_n have a non-zero probability of occurrence. All the events E_1, E_2, \dots, E form a partition of S. Let A be an event in space S for which we have to find the probability, then according to Bayes theorem,

$$P(E_i|A) = \sum_{k=1}^n P(E_k) \cdot P(A|E_k) P(E_i) \cdot P(A|E_i)$$

for $k = 1, 2, 3, \dots, n$

Bayes Theorem Derivation

The proof of Bayes' Theorem is given as, according to the conditional probability formula,
 $P(E_i|A) = P(A)P(E_i \cap A)$(i)

Then, by using the multiplication rule of probability, we get
 $P(E_i \cap A) = P(E_i) \cdot P(A|E_i)$(ii)

Now, by the total probability theorem,
 $P(A) = \sum_{k=1}^n P(E_k) \cdot P(A|E_k)$(iii)

Substituting the value of $P(E_i \cap A)$ and $P(A)$ from eq (ii) and eq(iii) in eq(i) we get,

$$P(E_i|A) = \frac{\sum_{k=1}^n P(E_k) \cdot P(A|E_k)P(E_i)}{\sum_{k=1}^n P(E_k)P(A|E_k)}$$

Bayes' theorem is also known as the formula for the probability of "causes". As we know, the E_i 's are a partition of the sample space S, and at any given time, only one of the events E_i occurs. Thus, we conclude that the Bayes theorem formula gives the probability of a particular E_i , given that event A has occurred.

Terms Related to Bayes' Theorem

After learning about **Bayes** theorem in detail, let us understand some important terms related to the concepts we covered in the formula and derivation.

Hypotheses

- Hypotheses refer to possible events or outcomes in the sample space; they are denoted as E_1, E_2, \dots, E_n .
- Each hypothesis represents a distinct scenario that could explain an observed event.

Prior Probability

- Prior Probability $P(E_i)$ is the initial probability of an event occurring before any new data is taken into account.
- It reflects existing knowledge or assumptions about the event.
- **Example:** The probability of a person having a disease before taking a test.

Posterior Probability

- Posterior probability ($P(E_i|A)$) is the updated probability of an event after considering new information.
- It is derived using the Bayes Theorem.
- **Example:** The probability of having a disease given a positive test result.

Conditional Probability

- The probability of an event A based on the occurrence of another event B is termed conditional Probability.
- It is denoted as $P(A|B)$ and represents the probability of A when event B has already happened.

Joint Probability

- When the probability of two or more events occurring together and at the same time is measured, it is marked as Joint Probability.
- For two events A and B, it is denoted by joint probability is denoted as $P(A \cap B)$.

Random Variables

- Real-valued variables whose possible values are determined by random experiments are called random variables.
- The probability of finding such variables is the experimental probability.

Bayes Theorem Applications

Bayesian inference is very important and has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc., and Bayesian inference is directly derived from Bayes theorem.

Some of the Key Applications are:

- **AI & Machine Learning** → Used in **Naïve Bayes classifiers** to predict outcomes.
- **Medical Testing** → Finding the real probability of having a disease after a positive test.
- **Spam Filters** → Checking if an email is spam based on keywords.
- **Weather Prediction** → Updating the chance of rain based on new data.

Difference Between Conditional Probability and Bayes Theorem

The difference between Conditional Probability and Bayes's theorem can be understood with the help of the table given below.

Bayes Theorem	Conditional Probability
Bayes's Theorem is derived using the definition of conditional probability. It is used to find the reverse probability.	Conditional Probability is the probability of event A when event B has already occurred.
Formula: $P(A B) = [P(B A)P(A)] / P(B)$	Formula: $P(A B) = P(A \cap B) / P(B)$
Purpose: To update the probability of an event based on new evidence.	Purpose: To find the probability of one event based on the occurrence of another.
Focus: Uses prior knowledge and evidence to compute a revised probability.	Focus: Direct relationship between two events.

Theorem of Total Probability

Let E_1, E_2, \dots, E_n be **mutually exclusive and exhaustive events** of a sample space S , and let E be any event that occurs with some E_i . Then, prove that :

$$P(E) = \sum_{i=1}^n P(E/E_i) \cdot P(E_i)$$

Proof:

Let S be the sample space.

Since the events E_1, E_2, \dots, E_n are mutually exclusive and exhaustive, we have:

$S = E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n$ and $E_i \cap E_j = \emptyset$ for $i \neq j$.

Now, consider the event E : $E = E_i \cup E_j \cup \dots \cup E_n$

Substituting S with the union of E_i 's:

$\boxed{E} = E \cap (E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n)$

Using distributive law:

$\boxed{E} = (E \cap E_1) \cup (E \cap E_2) \cup \dots \cup (E \cap E_n)$

Since the events E_i are mutually exclusive, the intersections $E \cap E_i$ are also **mutually exclusive**. Therefore:

$P(E) = P\{(E \cap E_1) \cup (E \cap E_2) \cup \dots \cup (E \cap E_n)\}$

$\boxed{P(E)} = P(E \cap E_1) + P(E \cap E_2) + \dots + P(E \cap E_n)$

{Therefore, $(E \cap E_1), (E \cap E_2), \dots, (E \cap E_n)$ are pairwise disjoint}

$\boxed{P(E)} = P(E/E_1) \cdot P(E_1) + P(E/E_2) \cdot P(E_2) + \dots + P(E/E_n) \cdot P(E_n)$ [by multiplication theorem]

$\boxed{P(E)} = \sum_{i=1}^n P(E/E_i) \cdot P(E_i)$

Solved Examples of Bayes' Theorem

Example 1: A person has undertaken a job. The probabilities of completion of the job on time with and without rain are 0.44 and 0.9, **and** 5, respectively. If the probability that it will rain is 0.45, then determine the probability that the job will be completed on time.

Solution:

Let E_1 be the event that the mining job will be completed on time and E_2 be the event that it rains. We have,

$$P(A) = 0.45,$$

$$P(\text{no rain}) = P(B) = 1 - P(A) = 1 - 0.45 = 0.55$$

By multiplication law of probability,
 $P(E_1) = 0.44$, and $P(E_2) = 0.95$

Since, events A and B form partitions of the sample space S , by total probability theorem, we have

$$\begin{aligned} P(E) &= P(A) + P(B) \\ \boxed{P(E)} &= 0.45 \times 0.44 + 0.55 \times 0.95 \\ \boxed{P(E)} &= 0.198 + 0.5225 = 0.7205 \end{aligned}$$

So, the probability that the job will be completed on time is 0.7205

Example 2: There are three urns containing 3 white and 2 black balls, 2 white and 3 black balls, and 1 black and 4 white balls, respectively. There is an equal probability of each urn being chosen. One ball is chosen at random. What is the probability that a white ball will be drawn?

Solution:

Let E_1 , E_2 , and E_3 be the events of choosing the first, second, and third urn respectively. Then,
 $P(E_1) = P(E_2) = P(E_3) = 1/3$

Let E be the event that a white ball is drawn. Then,
 $P(E/E_1) = 3/5$, $P(E/E_2) = 2/5$, $P(E/E_3) = 4/5$

By theorem of total probability, we have
 $P(E) = P(E/E_1) \cdot P(E_1) + P(E/E_2) \cdot P(E_2) + P(E/E_3) \cdot P(E_3)$

$$\begin{aligned} \text{② } P(E) &= (3/5 \times 1/3) + (2/5 \times 1/3) + (4/5 \times 1/3) \\ \text{③ } P(E) &= 9/15 = 3/5 \end{aligned}$$

Example 3: A card from a pack of 52 cards is lost. From the remaining cards of the pack, two cards are drawn and are found to be both hearts. Find the probability of the lost card being a heart.

Solution:

Let E_1, E_2, E_3 , and E_4 be the events of losing a card of hearts, clubs, spades, and diamonds respectively. Then $P(E_1) = P(E_2) = P(E_3) = P(E_4) = 13/52 = 1/4$.

Let E be the event of drawing 2 hearts from the remaining 51 cards. Then,

$$\begin{aligned} P(E|E_1) &= \text{probability of drawing 2 hearts, given that a card of hearts is missing} \\ \text{④ } P(E|E_1) &= {}^{12}C_2 / {}^{51}C_2 = (12 \times 11)/2! \times 2!/(51 \times 50) = 22/425 \end{aligned}$$

$$\begin{aligned} P(E|E_2) &= \text{probability of drawing 2 clubs, given that a card of clubs is missing} \\ \text{⑤ } P(E|E_2) &= {}^{13}C_2 / {}^{51}C_2 = (13 \times 12)/2! \times 2!/(51 \times 50) = 26/425 \end{aligned}$$

$$\begin{aligned} P(E|E_3) &= \text{probability of drawing 2 spades, given that a card of hearts is missing} \\ \text{⑥ } P(E|E_3) &= {}^{13}C_2 / {}^{51}C_2 = 26/425 \end{aligned}$$

$$\begin{aligned} P(E|E_4) &= \text{probability of drawing 2 diamonds, given that a card of diamonds is missing} \\ \text{⑦ } P(E|E_4) &= {}^{13}C_2 / {}^{51}C_2 = 26/425 \end{aligned}$$

Therefore,

$P(E_1|E)$ = probability of the lost card is being a heart, given the 2 hearts are drawn from the remaining 51 cards

$$\begin{aligned} \text{⑧ } P(E_1|E) &= P(E_1) \cdot P(E|E_1)/P(E_1) \cdot P(E|E_1) + P(E_2) \cdot P(E|E_2) + P(E_3) \cdot P(E|E_3) + P(E_4) \cdot P(E|E_4) \\ \text{⑨ } P(E_1|E) &= (1/4 \times 22/425) / \{(1/4 \times 22/425) + (1/4 \times 26/425) + (1/4 \times 26/425) + (1/4 \times 26/425)\} \\ \text{⑩ } P(E_1|E) &= 22/100 = 0.22 \end{aligned}$$

Hence, The required probability is 0.22.

How is Probability Used in Data Science

Probability helps data scientists make decisions when outcomes are uncertain. It gives a mathematical way to estimate how likely an event is, based on existing data. Whether it's classifying emails, forecasting demand or handling missing values probability is at the core of many data tasks.

Where Probability Fits in Data Science

In data science, probability is used for:

- Estimating the chances of different outcomes
- Building models that predict behavior (e.g., spam detection, churn prediction)
- Making decisions with incomplete or noisy data
- Measuring how confident we are in predictions
- Updating results as new data becomes available (Bayesian methods)

Example 1: Are Emails with "Offer" More Likely to Be Spam?

Suppose you're analyzing email data to detect spam. You notice that many spam emails contain the word "offer".

Here's a sample:

EmailID	Contains "Offer"	Spam
001	Yes	Yes
002	Yes	No
003	No	No
004	Yes	Yes

Calculation Using Bayes' Theorem

$$P(\text{Spam}|\text{Offer}) = P(\text{Offer})P(\text{Offer}|\text{Spam}) \cdot P(\text{Spam})$$

From the table:

- $P(\text{Spam})=42$
- $P(\text{Offer})=43$
- $P(\text{Offer}|\text{Spam})=22=1$

$$P(\text{Spam}|\text{Offer}) = 43 \cdot 42 = 32 \approx 0.67$$

So, there's a 67% chance an email is spam if it contains "offer".

In this example, Data Science uses past email data to find that the word "offer" often appears in spam. By applying probability, it estimates the chance an email is spam, helping build simple but effective spam detection systems.

Example 2: How Likely is Product Demand to Rise?

An e-commerce company wants to know if advertising during festivals increases product sales. By collecting past data and applying **probability distributions** (like Poisson or Normal), you can model the likelihood of a sales spike during specific times.

Common Probability Tools in Data Science

Task	Probability Concept
Classifying outcomes	Conditional probability
Measuring uncertainty	Probability distributions
Making predictions	Naive Bayes, Logistic Regression
Decision-making under uncertainty	Bayesian inference
Simulating random events	Monte Carlo simulations

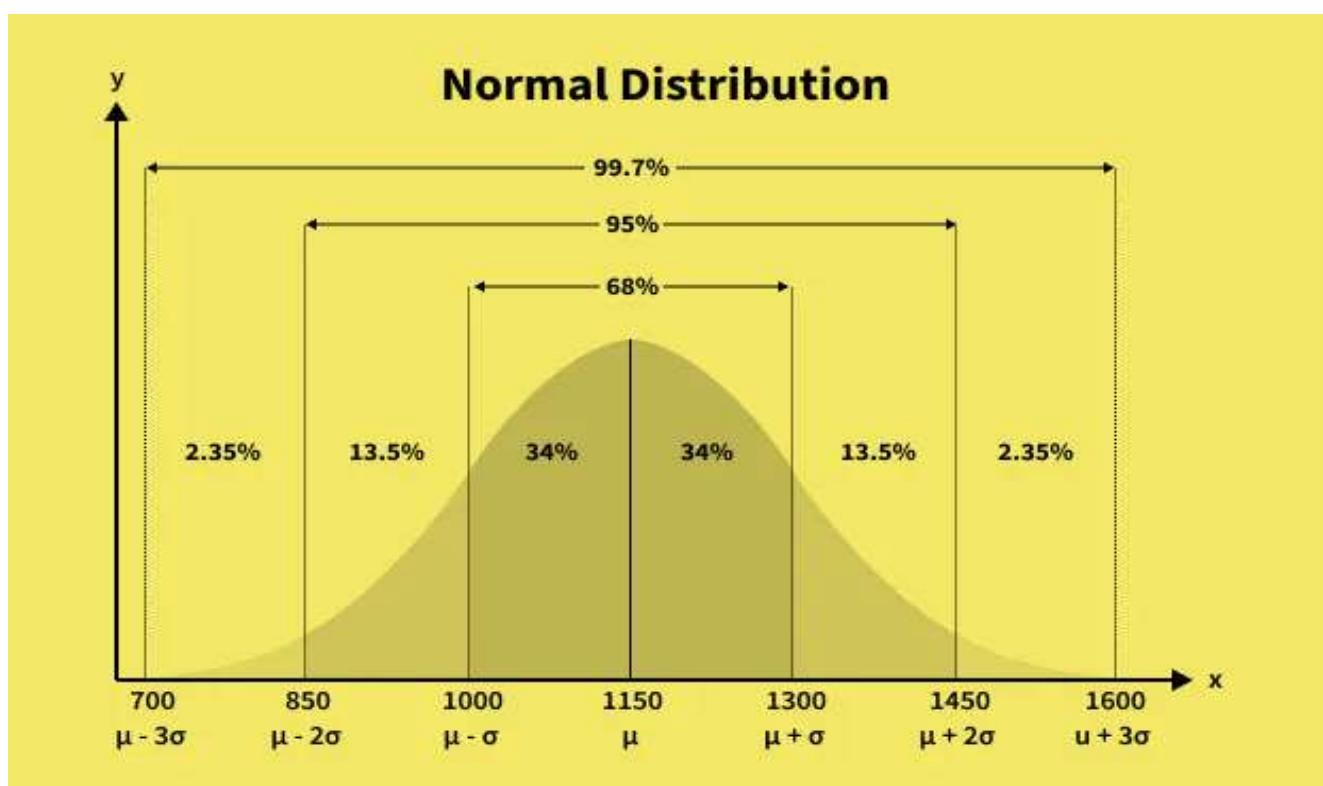
Probability helps turn raw data into insights when we're not 100% sure of the outcome. It's what allows data scientists to work with risks, patterns and predictions in a logical and structured way.

Technical Tools and Libraries Used

Probability in Data Science is implemented using libraries like NumPy for numerical operations and random simulations, pandas for organizing data and **SciPy** for working with probability distributions and statistical tests. For Bayesian analysis, tools like PyMC or scikit-learn (for Naive Bayes models) are often used. Visualizations are done using matplotlib and seaborn to make probabilistic insights easier to interpret.

Normal Distribution in Data Science

Normal Distribution also known as the Gaussian Distribution or Bell-shaped Distribution is one of the widely used probability distributions in statistics. It plays an important role in probability theory and statistics basically in the Central Limit Theorem (CLT). It is characterized by its bell-shaped curve which is symmetric around the mean (μ). This symmetry shows that values equally distant from the mean. The probability of an event decreases as we move further away from the mean with most events clustering around the center. In this article, we will see the normal distribution and its core concepts.



It can be observed in the above image that the distribution is symmetric about its center which is the mean (0 in this case). This makes the probability of events at equal deviations from the mean equally probable. The density is highly centered around the mean which translates to lower probabilities for values away from the mean.

Probability Density Function (PDF)

The probability density function of the normal distribution defines the likelihood of a random variable taking a particular value. The formula for the PDF is given by:

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

where:

- μ (mu) is the mean of the distribution. It represents the central value of the distribution.
- σ (sigma) is the standard deviation which measures the spread or dispersion of the distribution.
- x is the specific value for which we're calculating the probability.

While the formula might seem complex at first time lets break it down to simplify it. The z-score is a measure shows how many standard deviations a data point is from the mean. Mathematically, it's defined as:

$$z\text{-score} = \frac{x - \mu}{\sigma}$$

The exponent in the formula involves the square of the z-score multiplied by 2–1 which aligns with the observation that values farther from the mean are less likely. Larger z-scores (representing values farther from the mean) result in smaller probabilities due to the negative exponent. On the other hand, values closer to the mean result in smaller z-scores and higher probabilities.

This behavior is reflected in the 68-95-99.7 rule which states that:

- 68% of values lie within 1 standard deviation from the mean,
- 95% lie within 2 standard deviations and
- 99.7% lie within 3 standard deviations.

The figure given below shows this rule:

68-95-99.7 rule

Expectation ($E[X]$), Variance and Standard Deviation

The expectation or expected value $E[X]$ of a random variable gives us a measure of the "center" of the distribution. For a normally distributed random variable X with parameters μ (mean) and σ^2 (variance), the expectation is calculated by integrating the product of the random variable and its probability density function (PDF) over all possible values.

Mathematically, the expected value $E[X]$ is:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

For the normal distribution, the formula becomes:

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

We can simplify this by breaking it into two parts:

- The first part involves integrating $(x-\mu)$ which is symmetric about the mean and its result is zero because the distribution is symmetric.
- The second part involves multiplying the mean μ by the total probability which equals 1 (since the area under the normal curve is always 1).

Thus we find: $E[X] = \mu$

This tells us that the expected value of a normal distribution is simply the mean μ .

Variance and Standard Deviation

The variance of a normal distribution is the square of the standard deviation denoted as σ^2 . It measures how spread out the values of the distribution are from the mean.

The standard deviation σ is simply the square root of the variance:

$$\text{Variance} = \sigma^2$$

$$\text{Standard Deviation} = \sigma$$

Standard Normal Distribution

In the General Normal Distribution, if the Mean is set to 0 and the Standard Deviation is set to 1 then resulting distribution is called the Standard Normal Distribution. The formula for the Probability Density Function (PDF) of the standard normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

where:

- $\mu=0$ (mean)
- $\sigma=1$ (SD).

The Standard Normal Distribution is symmetric around the mean and its PDF defines the shape of the bell curve.

Cumulative Distribution Function (CDF)

1. The Cumulative Distribution Function (CDF) of the normal distribution does not have a closed-form expression. As a result, precomputed values from standard normal tables are used to find cumulative probabilities. These tables specifically provide cumulative probabilities for the standard normal distribution.

2. For a general normal distribution, the first step is to standardize the distribution by converting it into a z-score. Once standardized, the cumulative probability is calculated using the standard normal distribution tables.

3. This process has two key benefits:

- Only one table is needed to calculate probabilities for all normal distributions regardless of the specific mean and standard deviation.
- The table size is manageable containing 40 to 50 rows and 10 columns.

This is due 68-95-99.7 rule which says that values within 3 standard deviations of the mean account for 99.7% probability. So beyond $X=3$ ($\mu+3\sigma=0+3*1=3$) will have very small probabilities which are

approximately

x/0.0	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79670	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

Example: Finding Probabilities

Problem: Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with a mean of 10 milliamperes and a variance of four milliamperes². What is the probability that a measurement exceeds 13 milliamperes?

Solution:

1. Let X denote the current in milliamperes. We are tasked with finding $P(X > 13)$.

2. Standardize X by converting it to a z-score:

$$Z = \sigma X - \mu = 13 - 10 = 3 = 1.5$$

3. Now $P(X > 13)$ becomes equivalent to $P(Z > 1.5)$ in the standard normal distribution.

4. From the standard normal table, find the value of $P(Z \leq 1.5) = 0.93319$

5. So $P(Z \geq 1.5) = 1 - P(Z \leq 1.5) = 1 - 0.93319 = 0.06681$

Thus the probability that the current exceeds 13 milliamperes is approximately 0.06681, or 6.7%.

Implementation of Normal Distribution in Python

Here we will be using Numpy, Matplotlib and Seaborn libraries for the implementation.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

mean = 10

```

std_dev = 2
size = 1000

data = np.random.normal(loc=mean, scale=std_dev, size=size)

sns.histplot(data, kde=True, stat="density", bins=30, color="skyblue", linewidth=0.8)

plt.title(f'Normal Distribution ( $\mu$ ={mean},  $\sigma$ ={std_dev})')
plt.xlabel('Value')
plt.ylabel('Density')
plt.show()

Output:

```

Result

Applications of Normal Distribution

The normal distribution is incredibly versatile and is used across a variety of fields:

1. **Scientific Research:** Measurement errors are normally distributed helps in making this distribution important in experimental design and hypothesis testing.
2. **Finance:** In stock market analysis, returns of stock prices follow a normal distribution. This helps in risk assessment and portfolio optimization.
3. **Engineering:** Manufacturing processes such as the dimensions of parts produced can be modeled using normal distribution.
4. **Psychometrics:** Test scores and IQ scores are assumed to follow a normal distribution helps in aiding in standardized testing and education.
5. **Healthcare:** Certain biological measurements (e.g blood pressure) tend to follow normal distributions which helps in identifying outliers or abnormal conditions.

Binomial Distribution in Data Science

Binomial Distribution is used to calculate the probability of a specific number of successes in a fixed number of independent trials where each trial results in one of two outcomes: success or failure. It is used in various fields such as quality control, election predictions and medical tests to make decisions based on probability. In this article, we'll see the more about Binomial Distribution and its core concepts.

Key Concepts of Binomial Distribution

1. **Bernoulli Trial:** A Bernoulli trial is an experiment that results in one of two outcomes: success or failure. The trials are independent means the outcome of one trial does not affect the others. Example: Tossing a coin where heads = success and tails = failure.
2. **Number of Trials (n):** This refers to the fixed number of trials performed in the experiment. For example if we flip a coin 5 times, n = 5.

3. Success Probability (p): The probability of success in each trial is denoted by p. This probability is constant across all trials. Example: For a fair coin the probability of heads (success) on each flip is $p = 0.5$.

4. Failure Probability (q): The probability of failure is denoted by q and it is calculated as $q = 1 - p$. Since each trial results in either success or failure, we always have $p + q = 1$. Example: For a fair coin, $p=0.5$ for heads so $q = 1 - 0.5 = 0.5$ for tails.

Binomial Distribution Formula

Binomial Distribution calculates the probability of getting exactly x successes in n independent trials. The formula for the Probability Mass Function (PMF) is:

$$P(X=x) = (xn)p^x(1-p)^{n-x}$$

Where:

- $P(X=x)$ is the probability of getting exactly x successes.
- n is the number of trials.
- x is the number of successes we want to calculate the probability for.
- p is the probability of success in each trial.
- (xn) is the **binomial coefficient** which represents the number of ways to arrange x successes in n trials. It is calculated as:

$$(xn) = x!(n-x)!n!$$

Probability Mass Function (PMF)

The **Probability Mass Function** defines the probability of a specific number of successes occurring in the Binomial Distribution. It provides the likelihood of getting exactly x successes out of n trials. The formula for the PMF is as follows:

$$P(X=x) = (xn)p^x(1-p)^{n-x}$$

This formula tells us the probability of achieving exactly x successes in n trials. The binomial coefficient (xn) accounts for all the different ways in which x successes can occur in n trials.

Example: Tossing a Coin

Let's say we flip a coin 4 times ($n= 4$) and want to know the probability of getting exactly 2 heads (successes). Since the probability of heads is $p= 0.5$ for a fair coin, the probability of tails (failure) is $q = 1-p = 0.5$.

We calculate the probability using the PMF formula: $P(X=2) = (24)(0.5)^2(0.5)^{4-2}$

First, calculate the binomial coefficient: $(24) = (4-2)!2!4! = 2 \times 1 \times 3 = 6$

Now substitute the values into the PMF formula: $P(X=2) = 6 \times (0.5)^2 \times (0.5)^2 = 6 \times 0.25 \times 0.25 = 6 \times 0.0625 = 0.375$

Thus, the probability of getting exactly 2 heads in 4 tosses is **0.375** or **37.5%**.

Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of the Binomial Distribution gives the probability of obtaining at most x successes in n trials. It's the sum of the probabilities from $P(X=0)$ to $P(X=x)$.

The CDF is defined as:

$$F(x) = P(X \leq x) = \sum_{k=0}^x P(X=k)$$

In simpler terms, the CDF tells us the cumulative probability of getting zero, one, two or more successes in n trials. It is helpful when we want to know the probability of getting a certain number of successes or fewer.

Example:

If we want to know the probability of getting 3 or fewer heads in 5 coin tosses ($n=5$) we would calculate $P(X \leq 3)$ by summing the probabilities:

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

Expected Value of the Binomial Distribution

The expected value (mean) of a Binomial Distribution represents the average number of successes we expect after performing n trials. The expected value is calculated as:

$$E[X] = n \cdot p$$

For example if we flip a fair coin 5 times, the expected number of heads would be:

$$E[X] = 5 \times 0.5 = 2.5$$

So we would expect to get 2.5 heads on average after 5 flips of the coin.

Variance and Standard Deviation

1. Variance: The variance of a Binomial Distribution measures how much the number of successes varies from the expected value. It is given by:

$$\text{Var}[X] = n \cdot p \cdot (1-p)$$

2. Standard Deviation: The standard deviation is the square root of the variance which gives us a measure of how much the number of successes is likely to differ from the expected value on average:

$$\sigma = \sqrt{n \cdot p \cdot (1-p)}$$

Practical Example: Airline Ticket Sales

Let's apply the Binomial Distribution in a real-life scenario. Consider an airline that sells 65 tickets for a flight with a capacity of 60 passengers. The probability that a passenger does not show up for the flight is $q=0.1$ means the probability that a passenger shows up is $p=0.9$. The airline wants to know the probability that 60 or fewer passengers will show up so they don't need to reschedule tickets.

Step 1: Define Random Variable

Here the random variable X represents the number of passengers who show up. We need to calculate $P(X \leq 60)$ the probability that 60 or fewer passengers show up.

Step 2: Calculate Probability of More Than 60 Passengers

We first calculate the probability that more than 60 passengers show up which is:

$$P(X \geq 61) = P(X=61) + P(X=62) + \dots + P(X=65)$$

Step 3: Using Binomial Formula

Here we calculate the probabilities for $X=61, 62, \dots, 65$. We then subtract this from 1 to find $P(X \leq 60)$:

$$P(X \leq 60) = 1 - (P(X=61) + P(X=62) + P(X=63) + P(X=64) + P(X=65))$$

Step 4: Result

After performing the calculation we find:

$$P(X \leq 60) \approx 0.7909$$

After performing the calculation we find that the probability of 60 or fewer passengers showing up is approximately 79.09%. This means there is a 79.09% chance that the airline will not need to rebook any passengers.

Python Implementation for Binomial Distribution

Now let's implement the Binomial Distribution in Python to find the probabilities, visualize outcomes and calculate both the PMF and CDF. We'll be using Numpy, SciPy and Matplotlib libraries for this.

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import binom

n = 10
p = 0.5
x = np.arange(0, n+1)
pmf = binom.pmf(x, n, p)

plt.figure(figsize=(8, 6))
plt.bar(x, pmf, color='skyblue', edgecolor='black')
plt.title('Binomial Distribution PMF (n=10, p=0.5)', fontsize=14)
plt.xlabel('Number of successes (x)', fontsize=12)
plt.ylabel('Probability', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

cdf = binom.cdf(x, n, p)

plt.figure(figsize=(8, 6))
plt.plot(x, cdf, color='purple', marker='o', linestyle='-', linewidth=2)
plt.title('Binomial Distribution CDF (n=10, p=0.5)', fontsize=14)
plt.xlabel('Number of successes (x)', fontsize=12)
```

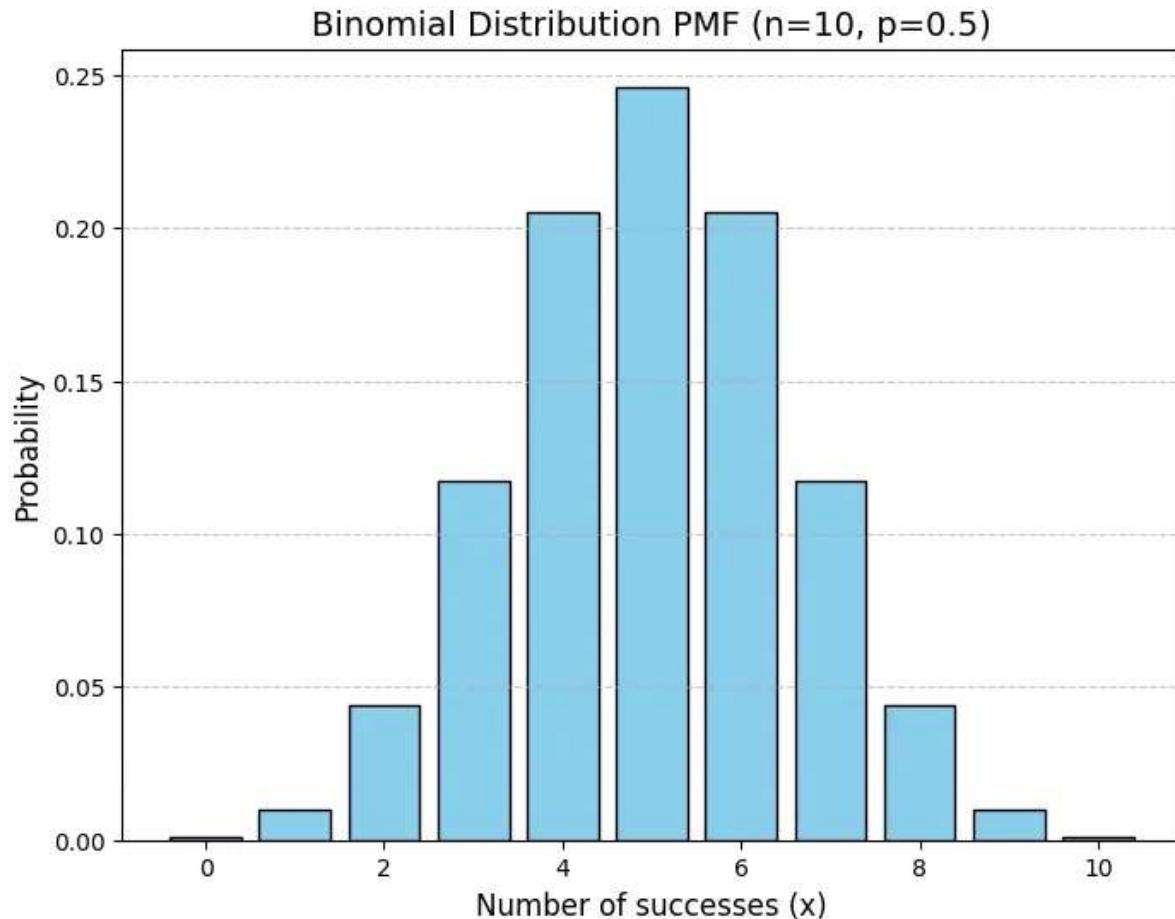
```
plt.ylabel('Cumulative Probability', fontsize=12)
```

```
plt.grid(True)
```

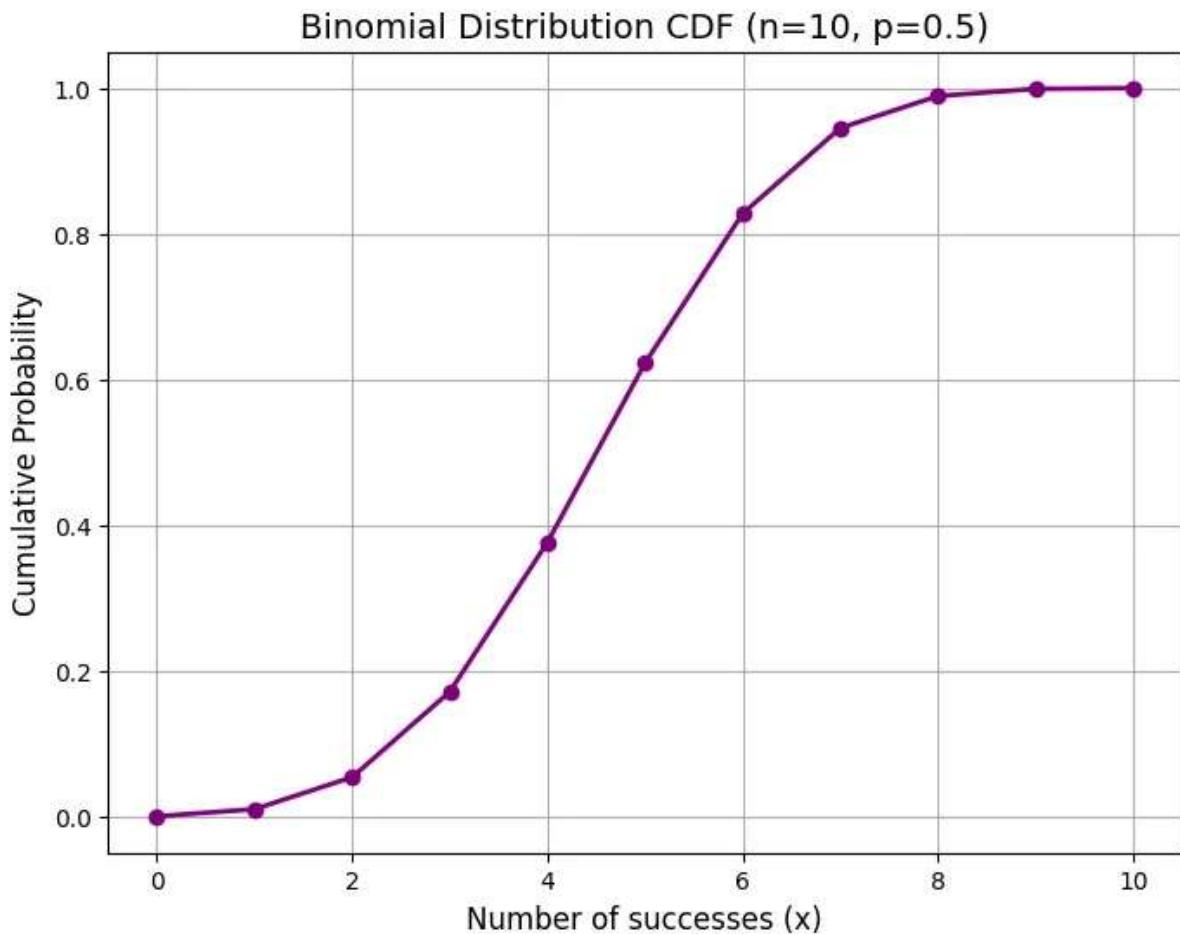
```
plt.show()
```

```
probability_3_heads = binom.pmf(3, n, p)
print(f'Probability of exactly 3 heads: {probability_3_heads:.4f}')
```

Output:



Result



Result

Probability of exactly 3 heads: 0.1172

Applications of the Binomial Distribution

Binomial Distribution has numerous applications in real-life scenarios:

1. **Quality Control:** In manufacturing the Binomial Distribution is used to model the number of defective items in a batch. For Example if a factory produces 100 items and has a 5% defect rate, the distribution can help estimate the probability of finding a certain number of defective items.
2. **Election Predictions:** During elections it can model the probability of a candidate receiving a certain number of votes. If each voter's decision is independent and the probability of voting for a particular candidate is known it can help to predict the likelihood of the candidate winning.
3. **Medical Testing:** In medical testing it is useful for predicting the number of positive test results out of a fixed number of tests. If the probability of a test result being positive is known the distribution can be used to calculate the chances of a certain number of positive results.
4. **Customer Behavior:** In retail, businesses can use the Binomial Distribution to model customer behaviors such as the probability that a customer will buy a product or the likelihood of a certain number of sales occurring in a fixed period.

Uniform Distribution in Data Science

Uniform Distribution also known as the Rectangular Distribution is a type of Continuous Probability Distribution where all outcomes in a given interval are equally likely. Unlike Normal Distribution which

have varying probabilities across their range, Uniform Distribution has a constant probability density throughout the interval which results in a "flat" distribution.

Key Concepts of the Uniform Distribution

1. Events and Interval

Uniform Distribution applies to events that are equally likely within a fixed interval $[a,b]$. This interval can represent time, space or any continuous measurement. The events within this interval are random and independent but they all have the same likelihood of occurring.

Example: If a fair die is rolled each side (1 to 6) is equally likely to appear which represents a discrete uniform distribution. In contrast, a continuous uniform distribution could model the outcome of a randomly chosen time within a 24-hour day where every moment has the same probability.

2. Probability Density Function (PDF)

PDF for a Uniform Distribution is a constant value across the entire interval $[a,b]$. This is because every point in the interval is equally likely to be chosen. Formula for the Uniform PDF is:

$$f(x)=\frac{1}{b-a}, a \leq x \leq b$$

For any x outside the interval $[a,b]$ the PDF is zero:

$$f(x)=0, \text{for } x < a \text{ or } x > b$$

Example: For a uniform distribution on the interval $[0,25]$ the PDF is:

$$f(x)=\frac{1}{25-0}=0.04, 0 \leq x \leq 25$$

This means every value between 0 and 25 has the same probability of occurring.

3. Cumulative Distribution Function (CDF)

CDF provides the probability that the random variable is less than or equal to a specific value x . For a Uniform Distribution, the CDF is calculated as the cumulative sum of the PDF over the range from a to x and its formula is:

$$F(x)=\begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases}$$

Example: For a Uniform Distribution between $a = 0$ and $b = 25$ the CDF is:

$$F(x)=\begin{cases} 0, & \text{if } x < 0 \\ 0.25x, & \text{if } 0 \leq x \leq 25 \\ 1, & \text{if } x > 25 \end{cases}$$

This shows that for any value of x , the probability that $X \leq x$ is simply the proportion of the interval up to x .

Properties of the Uniform Distribution

1. Expected Value (Mean)

The expected value of a Uniform Distribution represents the central tendency of the distribution. It is the average of the lower and upper bounds of the interval which is calculated as:

$$E[X] = a + b$$

Example: For a Uniform Distribution with $a=0$ and $b=25$, the expected value is:

$$E[X] = 20 + 25 = 12.5$$

This means the average value of the random variable X is 12.5.

2. Variance

The variance of the Uniform Distribution measures the spread of values around the mean and its formula is:

$$Var(X) = \frac{1}{12}(b-a)^2$$

Example: For $a=0$ and $b=25$, the variance is:

$$Var(X) = \frac{1}{12}(25-0)^2 = 12 \cdot 25 \approx 52.08$$

This tells us how much the values are expected to deviate from the mean.

3. Standard Deviation

The standard deviation is the square root of the variance and provides a measure of the dispersion of the distribution. The formula for the standard deviation is:

$$\sigma = \sqrt{Var(X)} = \sqrt{\frac{1}{12}(b-a)^2}$$

Example: For $a=0$ and $b=25$:

$$\sigma = \sqrt{12 \cdot 25} \approx 7.21$$

Example: Uniform Distribution in Copper Wire

Let's apply the Uniform Distribution to a real-world scenario. Suppose the current measured in a piece of copper wire is uniformly distributed over the interval $[0, 25]$. We can calculate the PDF, mean, variance, standard deviation and CDF.

1. PDF: $f(x) = \frac{1}{25-0} = 0.04, 0 \leq x \leq 25$

2. Expected Value: $E[X] = 20 + 25 = 12.5$

3. Variance: $Var(X) = \frac{1}{12}(25-0)^2 = 12 \cdot 25 \approx 52.08$

4. Standard Deviation: $\sigma = \sqrt{12 \cdot 25} \approx 7.21$

5. CDF: $F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 0.25x, & \text{if } 0 \leq x \leq 25 \\ 1, & \text{if } x > 25 \end{cases}$

Python Implementation for Uniform Distribution

Here we implement the Uniform Distribution in Python and we will be using [NumPy](#) and [Matplotlib](#) libraries to and visualize random samples from a uniform distribution.

```
import numpy as np
import matplotlib.pyplot as plt
```

```

a = 0
b = 25
samples = np.random.uniform(a, b, 1000)

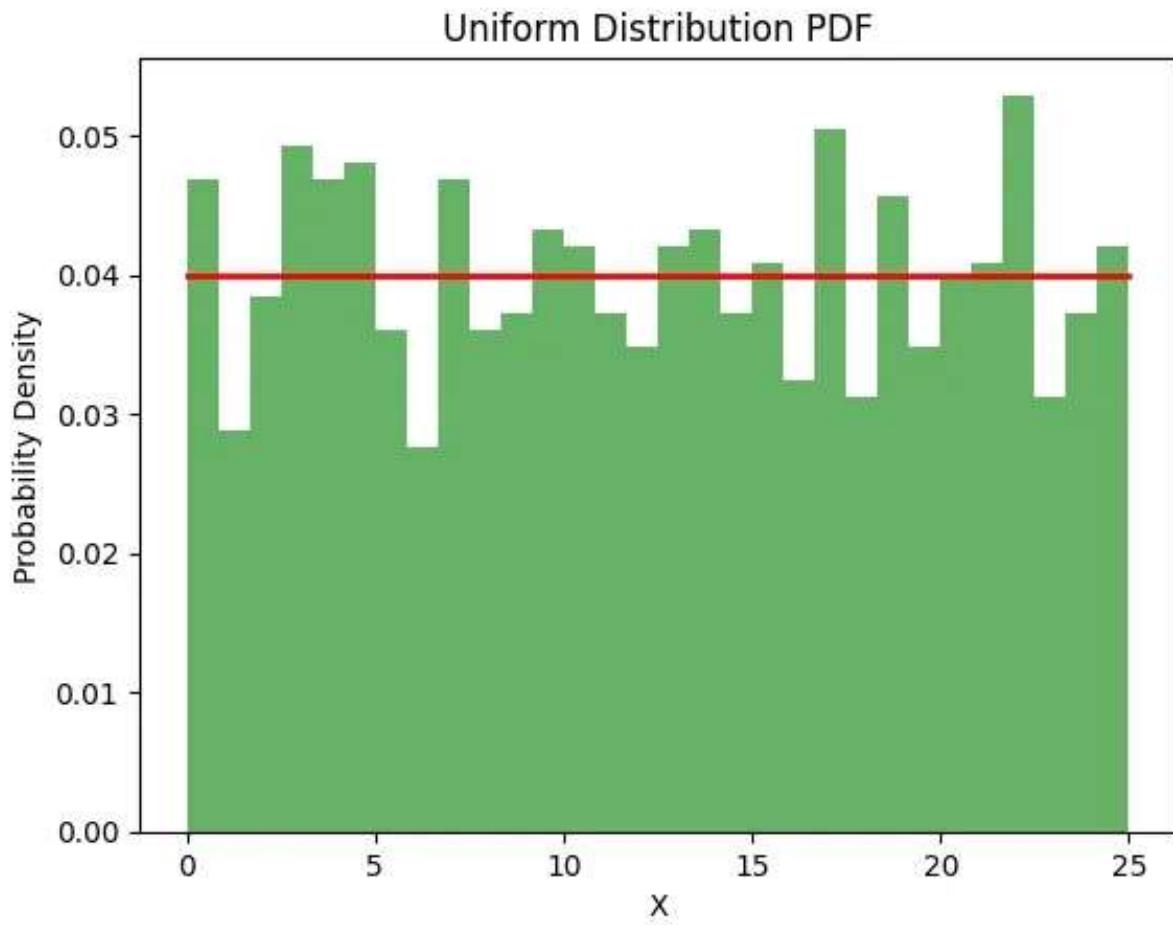
plt.hist(samples, bins=30, density=True, alpha=0.6, color='g')

x = np.linspace(a, b, 1000)
plt.plot(x, np.ones_like(x) / (b - a), 'r-', lw=2)

plt.title('Uniform Distribution PDF')
plt.xlabel('X')
plt.ylabel('Probability Density')
plt.show()

```

Output:



Result

Applications of the Uniform Distribution

Uniform Distribution is used in various real-world scenarios where all outcomes within a specific interval are equally likely, some common applications include:

1. **Random Sampling:** In simulations, random numbers are drawn from a Uniform Distribution to simulate random behavior.

2. **Quality Control:** In manufacturing, they model variations in product measurements when there's no systematic bias.
3. **Lottery and Gaming:** The distribution is used to model random number selection in lottery games or shuffling cards.
4. **Random Time Intervals:** If an event is equally likely to occur at any moment within a given time frame this can model the time of occurrence.

Exponential Distribution

The Exponential Distribution is one of the most commonly used probability distributions in statistics and data science. It is widely used to model the time or space between events in a Poisson process. In simple terms, it describes how long you have to wait before something happens, like a bus arriving or a customer calling a help center.

For example, if buses arrive at a bus stop every 15 minutes on average, the time you wait for the next bus can be modeled using an exponential distribution.

Probability Density Function (PDF)

The probability density function of the exponential distribution is:

$$f(x; \lambda) = \lambda e^{-\lambda x}, x \geq 0$$

Where:

- $\lambda > 0$ is the **rate parameter** (how often events occur)
- x is the time or distance until the next event

Cumulative Distribution Function (CDF)

The cumulative distribution function gives the probability that the event occurs within time :

$$F(x; \lambda) = 1 - e^{-\lambda x}, x \geq 0$$

Mean and Variance

- **Mean (Expected Value):** $E[X] = \lambda^{-1}$
- **Variance:** $\text{Var}(X) = \lambda^{-2}$

Memoryless Property

The exponential distribution is **memoryless**, which means:

$$P(X > s + t | X > s) = P(X > t)$$

This property tells us that the probability of waiting longer does not depend on how long you've already waited. This is unique to the exponential distribution.

Example

Suppose calls come into a customer support center at an average rate of 2 per minute. What is the probability that you wait more than 30 seconds for the next call?

Solution:

1. **Understand the Rate (λ):** Since 2 calls come in per minute, that means the average rate is:
 $\lambda=2$ calls per minute
2. **Convert Time:** Find the probability of waiting **more than 30 seconds**. But since the rate is in **minutes**, convert 30 seconds to minutes: 30 seconds=0.5 minutes
3. **Use the Exponential Distribution formula:**

$$P(X>0.5)=e^{-\lambda x}=e^{-2 \cdot 0.5}=e^{-1} \approx 0.3679$$

So, there is about a 36.79% chance that the next call comes **after** 30 seconds.

Example in Python

Before its implementation we should have some basic knowledge about numpy, matplotlib and seaborn.

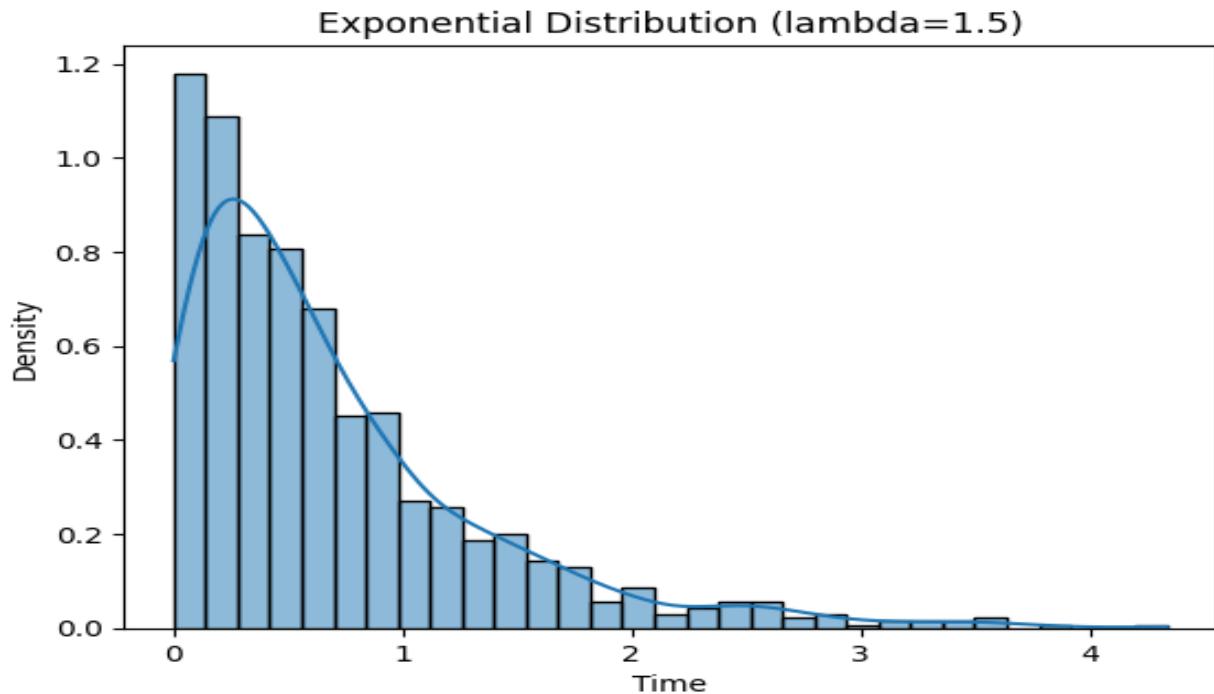
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Set the rate parameter lambda
lam = 1.5

# Generate exponential data
data = np.random.exponential(1/lam, 1000)

# Plot the KDE and histogram
sns.histplot(data, kde=True, stat="density")
plt.title("Exponential Distribution (lambda=1.5)")
plt.xlabel("Time")
plt.ylabel("Density")
plt.show()
```

Output:



Real-World Applications

1. **Call Centers:** Time between incoming calls
2. **Bank Queues:** Time between customer arrivals
3. **Web Servers:** Time between requests
4. **Manufacturing:** Time until a machine part fails
5. **Transport:** Time between buses or trains

Poisson Distribution in Data Science

Poisson Distribution is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space given a constant average rate of occurrence. Unlike the Binomial Distribution which is used when the number of trials is fixed, the Poisson Distribution is used for events that occur continuously or randomly over time or space. This makes it suitable for modeling rare events like accidents, phone calls or website hits. The distribution is defined by its mean λ which represents the expected number of events in the given interval.

Key Concepts of Poisson Distribution

1. Events: Poisson Distribution models the occurrence of events within a given time frame or spatial area. These events must occur independently which means the occurrence of one event doesn't affect the occurrence of others. Additionally, the events should happen at a constant average rate over the interval.

2. Average Rate (λ): The average rate λ also known as the rate parameter which represents the average number of occurrences of an event in the given time period or spatial area. This value remains constant throughout the observed interval. The parameter λ is central to the Poisson Distribution and finds the shape of the distribution.

3. Time or Space Interval: The interval during which we observe the occurrences of events is important in the Poisson Distribution. This interval can be defined in terms of time (e.g hours, days), space (e.g square miles) or any other metric where occurrences are spread out randomly and independently.

Poisson Distribution Formula

The Poisson Distribution calculates the probability of observing exactly x events in a fixed interval. The formula for the Poisson Probability Mass Function (PMF) is:

$$P(X=x) = \frac{x!}{x!} e^{-\lambda} \lambda^x$$

Where:

- $P(X=x)$ is the probability of observing exactly x events in the interval.
- λ is the average rate of occurrences (mean) in the interval.
- x is the number of events for which we are calculating the probability.
- e is Euler's number which is approximately equal to 2.718.
- This formula allows us to calculate the likelihood of a specific number of events occurring in the given time or space interval assuming that the events occur independently and at a constant rate.

Probability Mass Function (PMF)

The Poisson PMF is used to calculate the probability of exactly x events occurring in a fixed interval. The formula gives us the likelihood of observing x events given the average rate λ .

Example: Call Center

Let's us consider a call center which receives on average 3 calls per hour ($\lambda = 3$) and we want to know the probability of receiving exactly 4 calls in one hour ($x=4$).

We use the Poisson PMF formula:

$$P(X=4) = \frac{4!}{4!} e^{-3} 3^4 = 24e^{-3} 81 \approx 0.168$$

This means that the probability of receiving exactly 4 calls in one hour is approximately 0.168 or 16.8%. By calculating different values of x we can understand the distribution of events for various outcomes.

Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of the Poisson Distribution gives the probability of observing at most x events within a fixed interval. It's the sum of the probabilities from $P(X=0)$ to $P(X=x)$ which provides the cumulative probability.

The CDF is defined as:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} P(X=k)$$

Example: If we want to know the probability of receiving 3 or fewer calls in one hour we would calculate the CDF as:

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

This sum gives us the probability of receiving 0, 1, 2 or 3 calls in an hour which is helpful in scenarios where the exact number of events is not important but the total number of events up to a certain point is.

Expected Value of the Poisson Distribution

The expected value (mean) of a Poisson Distribution represents the average number of events we expect to occur in the given time or space interval. For the Poisson Distribution, the expected value is simply:

$$E[X] = \lambda$$

For example, if the average number of calls received by a call center is 4 per hour ($\lambda=4$), the expected number of calls in one hour is: $E[X]=4$

This means we expect to receive 4 calls on average every hour.

Variance and Standard Deviation

1. Variance: The variance of the Poisson Distribution is equal to λ , the average rate of events in the interval. The variance tells us how much the actual number of events deviates from the expected number of events.

$$Var[X] = \lambda$$

2. Standard Deviation: The standard deviation is the square root of the variance which gives us a measure of how spread out the number of events is from the expected value:

$$\sigma = \sqrt{\lambda}$$

For example if $\lambda=4$, the standard deviation would be: $\sigma=\sqrt{4}=2$

Example: Traffic Accidents

Let's apply the Poisson Distribution in a real-life scenario. Suppose that traffic accidents occur on a certain road at an average rate of 2 accidents per month ($\lambda=2$). We can use the Poisson Distribution to calculate the probability of having exactly 3 accidents in a given month. Using the Poisson PMF formula, we get:

$$P(X=3) = \frac{e^{-\lambda} \lambda^3}{3!} = \frac{e^{-2} 2^3}{3!} = \frac{6e^{-2}}{6} \approx 0.180$$

Thus the probability of having exactly 3 accidents in one month is 0.180 or 18%.

Python Implementation for Poisson Distribution

Now let's implement the Poisson Distribution in Python. Here we will be using Numpy, Matplotlib and Scipy libraries for this.

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import poisson

lambda_val = 3
```

```

k = np.arange(0, 10)
pmf = poisson.pmf(k, lambda_val)

plt.figure(figsize=(8, 6))
plt.bar(k, pmf, color='lightgreen', edgecolor='black')
plt.title('Poisson Distribution PMF ( $\lambda=3$ )', fontsize=14)
plt.xlabel('Number of events (k)', fontsize=12)
plt.ylabel('Probability', fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

```

```
cdf = poisson.cdf(k, lambda_val)
```

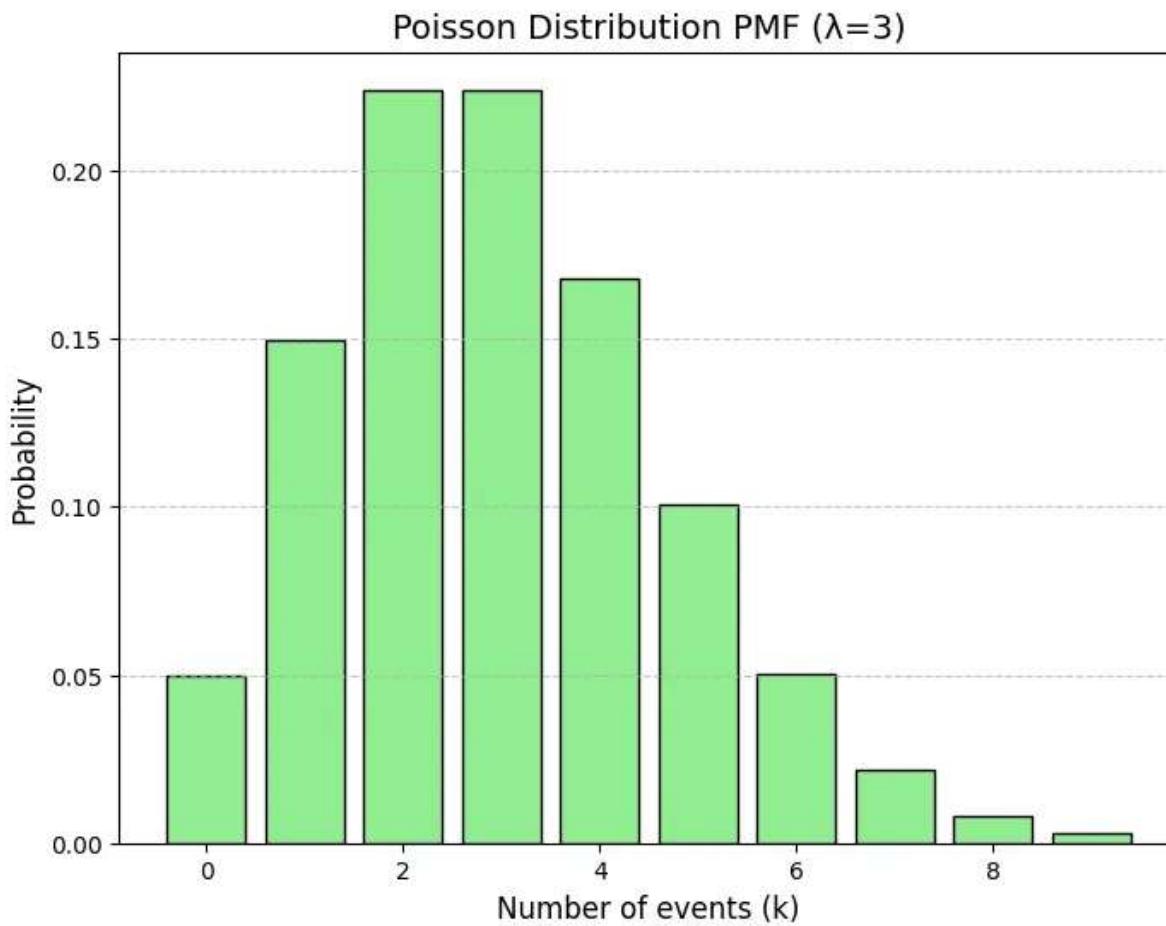
```

plt.figure(figsize=(8, 6))
plt.plot(k, cdf, color='purple', marker='o', linestyle='-', linewidth=2)
plt.title('Poisson Distribution CDF ( $\lambda=3$ )', fontsize=14)
plt.xlabel('Number of events (k)', fontsize=12)
plt.ylabel('Cumulative Probability', fontsize=12)
plt.grid(True)
plt.show()

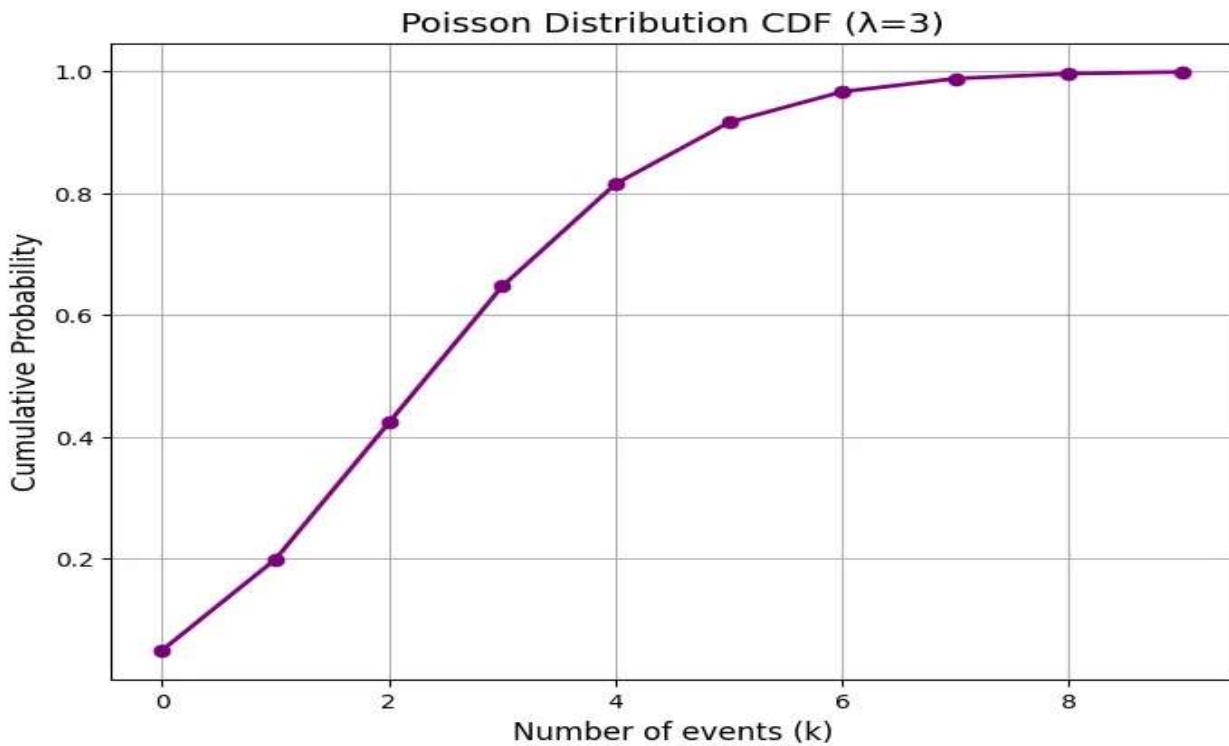
```

```
probability_4_events = poisson.pmf(4, lambda_val)
print(f'Probability of exactly 4 events: {probability_4_events:.4f}')
```

Output:



Output



Output

Probability of exactly 4 events: 0.1680

Relation between Poisson and Exponential Distributions

Poisson Distribution and Exponential Distribution are closely related probability distributions that describe different aspects of the same random process known as the **Poisson process**. In a Poisson process, events occur randomly and independently at a constant average rate over time or space. These two distributions are conceptually different but share a fundamental connection:

- **Poisson Distribution:** Models the number of events occurring in a **fixed** interval of time or space.
- **Exponential Distribution:** Models the time between consecutive events in the same process.

Both distributions are defined by the same rate parameter λ which represents the average number of events per unit of time or space. The relationship between the Poisson and Exponential distributions can be described as follows:

1. Poisson Distribution is used to calculate the probability of observing a certain number of events (k) in a fixed interval and its formula is:

$$P(X=k) = k! e^{-\lambda} \lambda^k, k=0,1,2,\dots$$

2. Exponential Distribution describes the waiting time between two consecutive events in a Poisson process. Its formula is:

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

Where:

- λ is the rate parameter, the average rate of events per unit of time.
- x is the waiting time between two consecutive events.

Skewness - Measures and Interpretation

Skewness is a key statistical measure that shows how data is spread out in a dataset. It tells us if the data points are skewed to the left (negative skew) or to the right (positive skew) in relation to the mean. It is important because it helps us to understand the shape of the data distribution which is important for accurate data analysis and helps in identifying outliers and finding the best statistical methods to use for analysis. In this article, we will see skewness, different types of skewness and its core concepts.

Skewness

Types of Skewness

Skewness describes the direction and degree of asymmetry in a dataset's distribution. Various types are as follows:

1. Positive Skewness (Right Skew)

In a positively skewed distribution, the right tail is longer than the left which means most data points are on the left with a few large values pulling the distribution to the right.

Relationship:

Mean > Median > Mode

Examples: Income distribution, exam scores and stock market returns.

2. Negative Skewness (Left Skew)

In a negatively skewed distribution, the left tail is longer which means most data points are on the right with a few smaller values pulling the distribution to the left.

Relationship:

Mean < Median < Mode

Examples: Test scores on easy exams, age at retirement and gestational age at birth.

3. Zero Skewness (Symmetrical Distribution)

Zero skewness shows a perfectly symmetrical distribution where the mean, median and mode are equal. In a symmetrical distribution, the data points are evenly distributed around the central point.

Relationship:

Mean = Median = Mode

Example: A perfectly balanced dataset with equal frequencies of all values.

Tests of Skewness

There are several ways to find the skewness of a dataset which can help to find whether the data is positively skewed, negatively skewed or roughly symmetric. Below are some common methods used to measure skewness:

1. Visual Inspection

This is the simplest and quickest method for assessing skewness by creating a histogram or a density plot of the given data.

- If the plot has a long tail on the right, the data is positively skewed (right-skewed).
- If the plot has a long tail on the left, the data is negatively skewed (left-skewed).
- If the plot is roughly symmetric, the data has no skewness (zero skew).

2. Skewness Coefficient (Pearson's First Coefficient of Skewness)

This is a numerical measure of skewness based on the relationship between the mean and mode. It helps us to find if the data is skewed when the mean and mode are not equal.

Formula : $\text{Skewness} = \text{Mean} - \text{Mode}$

- Positive Skew: If the mean is greater than the mode, the skewness is positive.
- Negative Skew: If the mean is smaller than the mode, the skewness is negative.
- Zero Skew: If the mean is equal to the mode, the skewness is zero which indicates a symmetric distribution.

3. Skewness Based on Quartiles

This method checks the distances between the quartiles to find skewness. If the quartiles are not equidistant, it suggests skewness:

- The third quartile (Q3) minus the median (Me) should ideally be equal to the median (Me) minus the first quartile (Q1) in a symmetric distribution.
- If this condition is not met, it shows either a positive or negative skew which depends on which side is longer.

Measurement of Skewness

Skewness is measured using different techniques to quantify the degree of asymmetry in a dataset's distribution. Below are three common methods to measure skewness:

1. Karl Pearson's Measure

Karl Pearson's Measure uses the mean, median and standard deviation of the given data to measure the asymmetry of the distribution. It provides a dimensionless number that helps to quantify how skewed the data is.

Formula:

1. *With respect to Mean and Median: $Sk = \sigma 3 \times (X^- - M)$*
2. *With respect to Mean and Mode: $Sk = \sigma (X^- - Mode)$*

Where:

- Sk is Karl Pearson's skewness coefficient
- X^- = Mean of the dataset
- M = Median of the dataset
- σ = Standard deviation of the dataset

Interpretation:

- **Skewness = 0:** The distribution is symmetric means the mean, median and mode are equal.
- **Skewness > 0:** The distribution is positively skewed (right-skewed) with the tail on the right side longer than the left.
- **Skewness < 0:** The distribution is negatively skewed (left-skewed) with the tail on the left side longer than the right.

Example: Calculate Pearson's skewness coefficient for a dataset of exam scores: 85, 88, 92, 94, 96, 98, 100, 100, 100, 100.

Solution:

Step 1: Calculation of Mean

$$\text{Mean}(X^-) = \frac{85+88+92+94+96+98+100+100+100+100}{10} = 95.3$$

Step 2: Calculation of Median

Since there are 10 data points, the median is the average of the 5th and 6th values when sorted in ascending order:

$$\text{Median} = \frac{96+98}{2} = 97$$

Step 3: Calculation of standard deviation.

$$\sigma^2 = N \sum (x_i - \mu)^2 = 10(85-95.3)^2 + \dots + (100-95.3)^2 = 10268.1 = 26.81$$

$$\text{Thus } \sigma = \sqrt{26.81}$$

$$\sigma = 5.$$

Step 4: Calculation of mode

It is clear from the data set that 100 is the most frequently occurring value in the data. Hence mode of given data is 100.

Step 5: Substitute the values in the formulae

1. With respect to Mean and Median

$$Sk = \sigma^3 (X^- - M) = 5(3 \times (95.3 - 97)) = 5 - 5.1$$

$$Sk = -1.02$$

2. With respect to Mean and Mode

$$Sk = \sigma(X^- - \text{Mode}) = 5(95.3 - 100)$$

$$Sk = -0.94$$

Since the skewness coefficient (Sk) is negative which shows a slight negative skewness in the distribution of exam scores. This means that the tail of the distribution is slightly longer on the left side and most of the scores are concentrated on the right side of the mean.

2. Bowley's Measure

Bowley's Skewness Coefficient is another method for calculating skewness based on quartiles (Q1, Q2, Q3). Unlike Karl Pearson's measure it does not rely on the mean or standard deviation which makes it useful for data that might not follow a normal distribution. It's calculated using the first quartile (Q1), the second quartile (Q2 or median) and the third quartile (Q3).

Formula:

$$B = \frac{Q_3 - Q_1}{Q_3 + Q_1 - 2Q_2}$$

Where:

- Q_1 = First quartile (25th percentile)
- Q_2 = Second quartile (50th percentile or median)
- Q_3 = Third quartile (75th percentile)

Interpretation:

- $B = 0$: The distribution is perfectly symmetric (no skewness).
- $B < 0$: The distribution is negatively skewed (left-skewed) with the tail on the left side longer.
- $B > 0$: The distribution is positively skewed (right-skewed) with the tail on the right side longer.

Example: Calculate Bowley's Measure of Skewness for the following dataset representing the ages of a group of people in a sample: 20, 24, 28, 32, 35, 40, 42, 45, 50.

Solution:

Step 1: Calculate the median (Q_2)

$$Q_2 = 35 \text{ (the middle value)}$$

Step 2: Calculate the first quartile (Q_1)

To find Q_1 lets consider the values to the left of the median: 20, 24, 28, 32

$$Q_1 = \frac{20 + 24 + 28}{3} = 24$$

Step 3: Calculate the third quartile (Q_3)

To find Q_3 lets consider the values to the right of the median: 40, 42, 45, 50.

$$Q_3 = \frac{40 + 42 + 45 + 50}{4} = 43.5$$

Step 4: Substitute the above values in the formula

$$B = \frac{Q_3 - Q_1}{Q_3 + Q_1 - 2Q_2} = \frac{43.5 - 24}{43.5 + 24 - 2 \times 35} = \frac{19.5}{2.5} = -7.8$$

$$B = -0.02$$

Since $B < 0$, this shows a negatively skewed (left-skewed) distribution means the tail is longer on the left side.

3. Kelly's Measure

Kelly's Skewness Measure calculates skewness by comparing certain percentiles in the data which typically the 10th, 50th (median) and 90th percentiles. This measure is useful when dealing with datasets that are not normally distributed or when other skewness measures may not be as effective.

Formula:

$$\text{Skewness} = P90 - P10P90 + P10 - 2P50$$

Where:

- $P90$ = 90th percentile
- $P50$ = 50th percentile (Median)
- $P10$ = 10th percentile

Interpretation:

- $\text{SKL} > 0$: Positive skew means the right tail is longer or heavier.
- $\text{SKL} < 0$: Negative skew means the left tail is longer or heavier.
- $\text{SKL} \approx 0$: The distribution is symmetric shows little or no skewness.

Example: Calculate Kelly's Coefficient of Skewness for the following data: 5, 7, 8, 9, 10, 12, 15, 16, 18, 20.

Solution:

Step 1: Find the 10th Percentile

To find the 10th percentile, we need to rank the data in ascending order and find the value below which 10% of the data falls. In this dataset, the 10th percentile corresponds to the value at position 1 since 10% of 10 data points is 1. So, the 10th percentile is 5.

$$P10=5$$

Step 2: Find the 50th Percentile (Median)

Since there are 10 data points, the median is the average of the 5th and 6th values when sorted in ascending order

$$\text{Median} = 210 + 12 = 222 = 11$$

$$P50=11$$

Step 3: Find the 90th Percentile

To find the 90th percentile we need to identify the value below which 90% of the data falls. In this dataset, the 90th percentile corresponds to the value at position 9 since 90% of 10 data points is 9. So the 90th percentile is 18.

$$P90=18$$

Step 4: Substitute the values in the formula.

$$\text{SKL} = 18 - 518 + 5 - 2 \times 11$$

$SKL=0.07$

Since $SKL > 0$, this shows a slight positive skew (right-skewed) means the distribution has a longer tail on the right side.

Interpretation of Skewness

Interpreting skewness involves understanding both the direction (left or right) and the magnitude (degree of skew) of the data distribution.

Direction of Skewness

1. Negative Skewness (Left Skewed): If the skewness is negative, it shows that the distribution is skewed to the left. In a left-skewed distribution:

- The tail on the left side (the smaller values) is longer and contains outliers.
- The majority of data points are concentrated on the right side.
- The mean is less than the median.

2. Positive Skewness (Right Skewed): A positive skewness shows that the distribution is skewed to the right. In a right-skewed distribution:

- The tail on the right side (the larger values) is longer and may contain outliers.
- Most data points are concentrated on the left side.
- The mean is greater than the median.

3. Zero Skewness (Symmetric): A skewness value close to zero suggests a symmetric distribution where the data is evenly distributed on both sides of the mean. This means there is no skewness.

Magnitude of Skewness

The magnitude of skewness gives us information about how extreme the skewness is:

- **Skewness close to 0 (between -0.5 and 0.5):** The distribution is approximately symmetric.
- **Skewness below -1:** Strong left skewness (negative skew) with a long tail on the left side.
- **Skewness above 1:** Strong right skewness (positive skew) with a long tail on the right side.

Handling Skewness in Data

When working with skewed data, it's important to understand how to handle skewness effectively. Skewed data can impact the accuracy of statistical analyses and predictions. There are various methods to handle skewness depending on the nature of the data and the analysis we want to perform. Let's see how we can handle skewness:

1. Data Transformation

- **Log Transformation:** It is useful for right-skewed data, compressing high values to create a more symmetric distribution.
- **Square Root/Cube Root:** It helps reduce positive skew, especially for count data.
- **Box-Cox Transformation:** A flexible method for handling both positive and negative skew.

2. Removing Outliers

Outliers can cause skewness, so removing them may improve symmetry:

- **Z-score:** It identify and remove data points with z-scores beyond ± 3 .
- **IQR Method:** It remove data points beyond 1.5 times the interquartile range.

3. Non-Parametric Tests

When transformations aren't effective, consider non-parametric tests like the Mann-Whitney U Test or Kruskal-Wallis Test which do not assume normal distribution and focus on medians rather than means.

4. Machine Learning Models

Some models handle skewed data better:

- **Tree-based Models:** Decision trees and random forests are less sensitive to skewness.
- **Generalized Linear Models (GLM):** Use appropriate link functions to model skewed data effectively.

Difference between Dispersion and Skewness

While dispersion and skewness may seem similar but they measure different aspects of data distribution. Dispersion refers to the extent to which data points are spread out from the central value (mean or median). It gives us an understanding of how varied the data is.

Now let's see a tabular differences for better understanding:

Dispersion	Skewness
Measures the spread of data around the central value (mean, median).	Measures the shape of the distribution and direction (left or right).
Variance, standard deviation, range, interquartile range (IQR).	Pearson's coefficient of skewness, moment skewness, Q-Q plots.
Dispersion affects the mean's interpretation but is not directly related.	Skewness shows the relationship between the mean and median.
High dispersion means data points are spread out widely.	Positive skew: Right tail longer. Negative skew: Left tail longer. Zero skew: Symmetric.
Helps understand the variability of data.	Helps identify the shape and asymmetry of data.
Test scores spread, stock price variability, age range.	Income distribution (right-skewed), exam scores (left/right-skewed).

The Central Limit theorem says that if we take many random samples from any population and calculate their averages, those averages will form a bell-shaped (normal) curve even if the original data is not normally distributed as long as the sample size is large enough. This helps us make predictions about the whole population using just sample data.

By calculating sample means these averages will tend to form a normal distribution. This normality holds true as long as the sample size is sufficiently large, typically $n \geq 30$ providing the foundation for making inferences about populations even when we don't have access to all the data.

Central Limit Theorem Formula

You have a population where the data follows some random variable X and this population has:

- **Mean** μ the average of the population
- **Standard deviation** σ

let's say we take a **sample** of size n from this population and calculate its **mean** \bar{X} then the Z-Score is given below:

Central Limit Theorem Formula

As the sample size increases the distribution of sample means becomes more concentrated around μ and resembles a normal distribution.

Key Assumptions for Central Limit Theorem

For the Central Limit Theorem (CLT) to work properly, a few conditions must be met:

- **Random Sampling:** The sample must be chosen randomly to fairly represent the whole population.
- **Independence:** Each data point should be independent one should not influence another.
- **Large Enough Sample Size:** A sample size of at least 30 is usually enough for the sample mean to follow a normal distribution.
- **Finite Mean and Variance:** The population should have a defined average and variation extreme or unlimited values can make CLT unreliable.

By ensuring these assumptions are met. The theorem can be used to draw conclusions about the population.

*While working with CLT we often need to work with skewed data, to learn more about skewed data refer to:
Skewness*

How CLT works in Data Science

You are a data analyst at a tech company. Users around the world have different web page load times, usually being biased based on network speed and location. You need to estimate the mean load time but it is impractical to verify every user.

Let's solve this problem step-by-step:

Step 1: Problem Identification

Instead of analyzing all user , you take a small sample (e.g., 50 users) to estimate the average load time. But since the data isn't normally distributed, can you trust this average? This is where the Central Limit Theorem comes into play.

Step 2: Data Sampling Process

To use the Central Limit Theorem (CLT):

- Take 50 random users and calculate their mean load time.
- Do it 1,000 times to obtain 1,000 sample means.
- When you graph these means, the outcome is an approximately normal distribution even though the original data is skewed.

Step 3: How to Implement the CLT

Now that we understand the scenario let us walk through the steps of how to implement the Central Limit Theorem using Python. Before its implementation we should have some basic knowledge about numpy and matplotlib.

We will generate fake web load times using an exponential distribution (to represent skewed data), take many random samples, and plot their means to observe how they form a normal distribution.

```
import numpy as np
import matplotlib.pyplot as plt

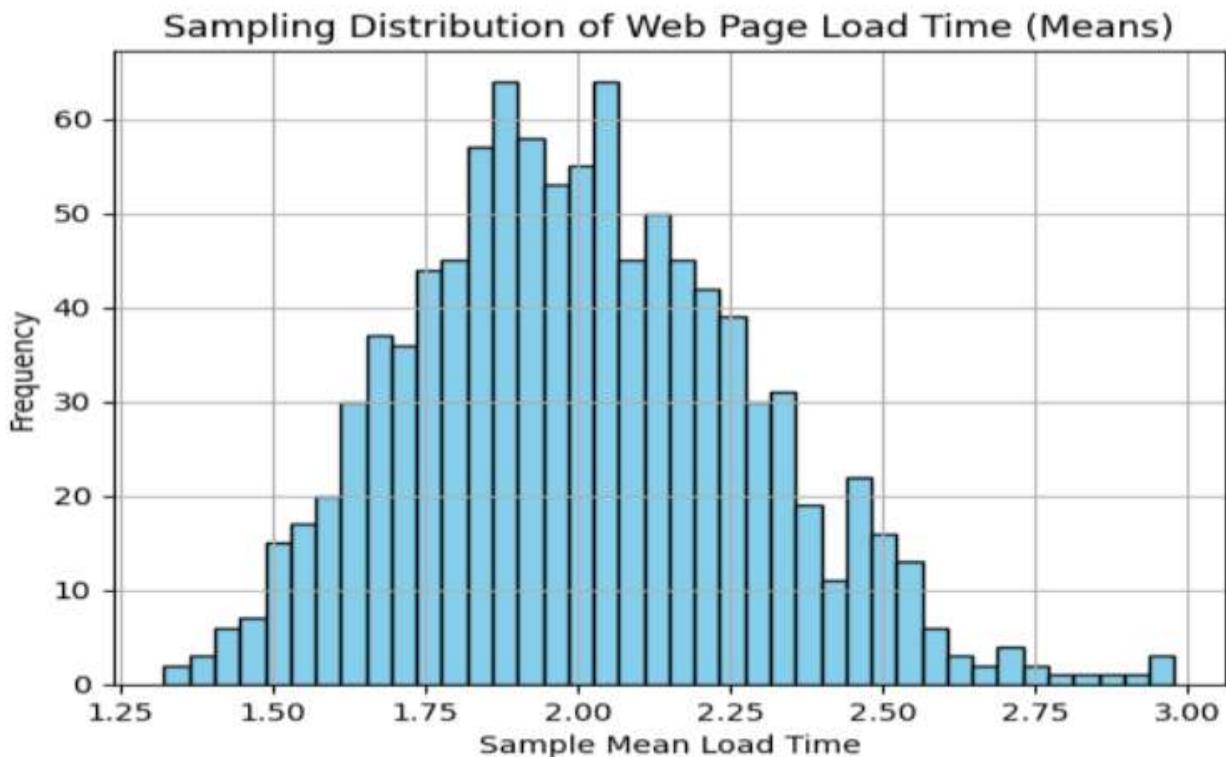
# Simulate skewed load time data
np.random.seed(0)
population = np.random.exponential(scale=2.0, size=100000)

# Parameters
sample_size = 50
num_samples = 1000
sample_means = []

# Take samples and compute means
for _ in range(num_samples):
    sample = np.random.choice(population, size=sample_size)
    sample_means.append(np.mean(sample))

# Plot the sample means
plt.hist(sample_means, bins=40, color='skyblue', edgecolor='black')
plt.title('Sampling Distribution of Web Page Load Time (Means)')
plt.xlabel('Sample Mean Load Time')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()

Output:
```



Although the original load time data is skewed, the histogram of sample means shows a normal curve. This confirms the Central Limit Theorem even non-normal data can produce a normal sampling distribution when you take enough samples.

Practical Applications of the Central Limit Theorem

The Central Limit Theorem (CLT) is widely used in machine learning and data analysis:

- **Model Evaluation and Confidence Intervals:** CLT helps build confidence intervals around model predictions, showing how reliable they are more data leads to tighter intervals and more trust in results.
- **A/B Testing:** A/B Testing is used in product development, CLT ensures that average outcomes from repeated experiments become normally distributed, even with skewed data.
- **Error and Uncertainty Estimation:** CLT allows us to estimate prediction errors and standard errors, helping assess model uncertainty on new data.
- **Bootstrapping:** By resampling data, CLT supports reliable estimation of metrics like MSE and confidence intervals for model parameters.
- **Feature Importance:** CLT helps check if feature rankings remain stable across samples, ensuring the most consistent and reliable features are chosen.

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.

Hypothesis Testing

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."

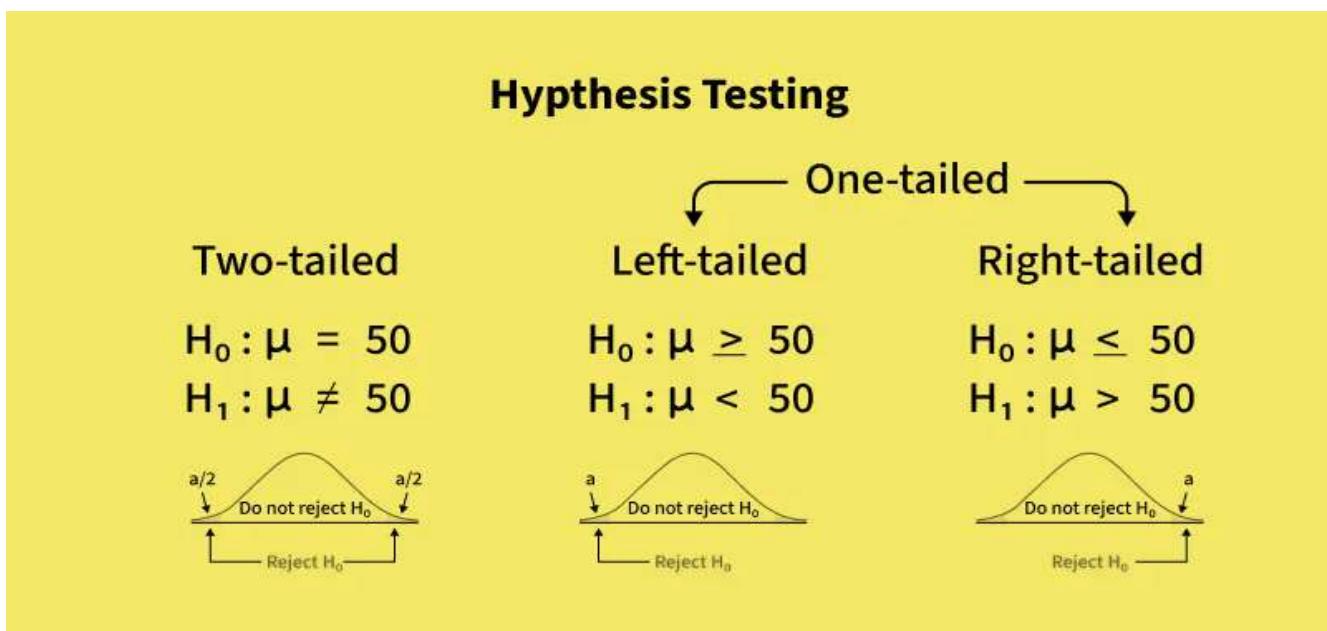
Key Terms of Hypothesis Testing

To understand the Hypothesis testing firstly we need to understand the key terms which are given below:

- **Significance Level (α):** How sure we want to be before saying the claim is false. Usually, we choose 0.05 (5%).
- **p-value:** The chance of seeing the data if the null hypothesis is true. If this is less than α , we say the claim is probably false.
- **Test Statistic:** A number that helps us decide if the data supports or rejects the claim.
- **Critical Value:** The cutoff point to compare with the test statistic.
- **Degrees of freedom:** A number that depends on the data size and helps find the critical value.

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.

There are two types of one-tailed test:

- **Left-Tailed (Left-Sided) Test:** Checks if the value is less than expected. Example: $H_0: \mu \geq 50$ and $H_1: \mu < 50$

- **Right-Tailed (Right-Sided) Test:** Checks if the value is greater than expected. Example: $H_0: \mu \leq 50$ and $H_1: \mu > 50$

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

Example: $H_0: \mu = 50$ and $H_1: \mu \neq 50$

To go deeper into differences into both types of test: [Refer to link](#)

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

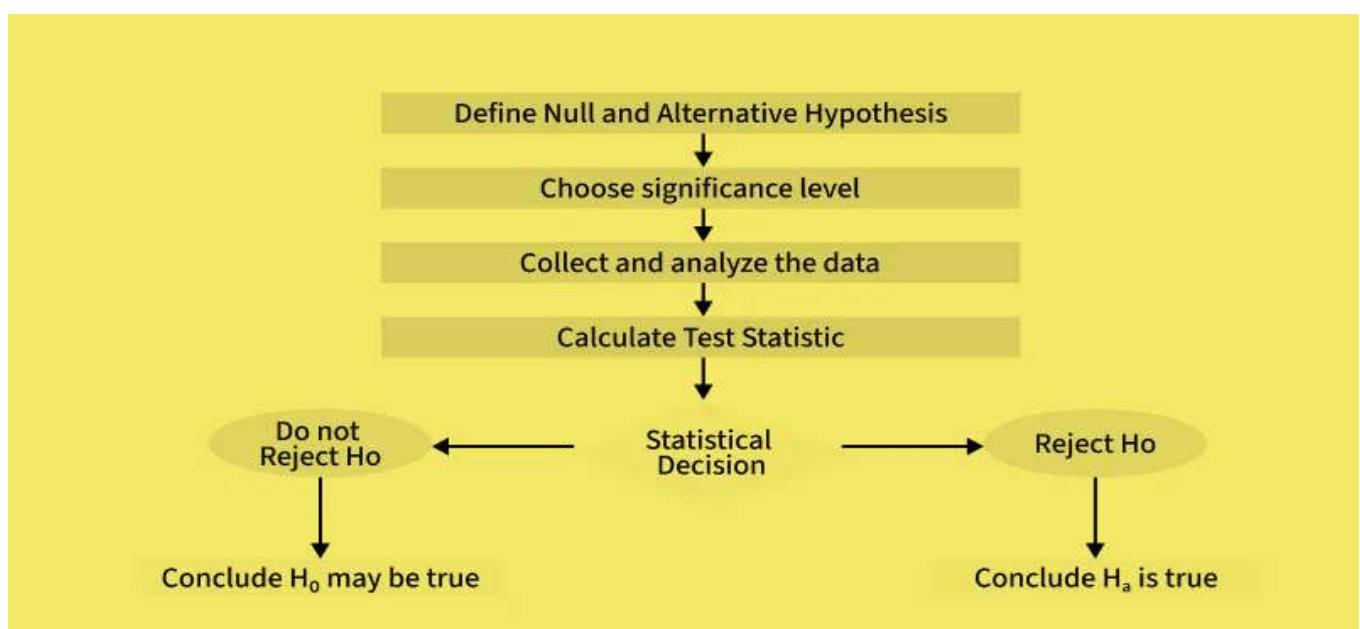
In hypothesis testing Type I and Type II errors are two possible errors that can happen when we are finding conclusions about a population based on a sample of data. These errors are associated with the decisions we made regarding the null hypothesis and the alternative hypothesis.

- **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).

	Null Hypothesis is True	Null Hypothesis is False
Null Hypothesis is True (Accept)	Correct Decision	Type II Error (False Negative)
Alternative Hypothesis is True (Reject)	Type I Error (False Positive)	Correct Decision

How does Hypothesis Testing work?

Working of Hypothesis testing involves various steps:



Steps of Hypothesis Testing

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**.
- **Example:** We collect data on user engagement before and after implementing the algorithm. We can also find the mean engagement scores for each group.

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 H_0 .
- If test statistic ≤ critical value → fail to reject H_0 .

2. Using P-value:

- If p-value ≤ α → reject H_0 .
- If p-value > α → fail to reject H_0 .

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.

Real life Examples of Hypothesis Testing

A pharmaceutical company tests a new drug to see if it lowers blood pressure in patients.

Data:

- Before Treatment: 120, 122, 118, 130, 125, 128, 115, 121, 123, 119
- After Treatment: 115, 120, 112, 128, 122, 125, 110, 117, 119, 114

Step 1: Define the Hypothesis

- **Null Hypothesis:** (H_0)The new drug has no effect on blood pressure.
- **Alternate Hypothesis:** (H_1)The new drug has an effect on blood pressure.

Step 2: Define the Significance level

Usually 0.05, meaning less than 5% chance results are by random chance.

Step 3: Compute the test statistic

Using paired T-test analyze the data to obtain a test statistic and a p-value. The test statistic is calculated based on the differences between blood pressure measurements before and after treatment.

$$t = m/(s/\sqrt{n})$$

Where:

- **m** = mean of the difference i.e $X_{\text{after}} - X_{\text{before}}$
- **s** = standard deviation of the difference (d) $d_i = X_{\text{after},i} - X_{\text{before},i}$
- **n** = sample size

then $m = -3.9$, $s = 1.37$ and $n = 10$. we calculate the T-statistic = -9 based on the formula for paired t test

Step 4: Find the p-value

With degrees of freedom = 9, p-value ≈ 0.0000085 (very small).

Step 5: Result

Since the p-value (8.538051223166285e-06) is less than the significance level (0.05) the researchers reject the null hypothesis. There is statistically significant evidence that the average blood pressure before and after treatment with the new drug is different.

Python Implementation of Case A

Now we will implement this using paired T-test with the help of `scipy.stats`. Scipy is a mathematical library in Python that is mostly used for mathematical equations and computations . Here we use the Numpy Library for storing the data in arrays.

```
import numpy as np
from scipy import stats
```

```

b = np.array([120, 122, 118, 130, 125, 128, 115, 121, 123, 119])
a = np.array([115, 120, 112, 128, 122, 125, 110, 117, 119, 114])

alpha = 0.05

t_stat, p_val = stats.ttest_rel(a, b)

m = np.mean(a - b)
s = np.std(a - b, ddof=1)
n = len(b)
t_manual = m / (s / np.sqrt(n))

decision = "Reject" if p_val <= alpha else "Fail reject"
concl = "Significant difference." if decision == "Reject" else "No significant difference."

print("T:", t_stat)
print("P:", p_val)
print("T manual:", t_manual)
print(f"Decision: {decision} H0 at α={alpha}")
print("Conclusion:", concl)

```

Output:

T:		-9.0		
P:		$8.538051223166285e-06$		
T	manual:	-9.0		
Decision:	Reject	H0	at	$\alpha=0.05$
Conclusion: Significant difference.				

The T-statistic of about -9 and a very small p-value provide strong evidence to reject the null hypothesis at the 0.05 level. This means the new drug significantly lowers blood pressure. The negative T-statistic shows the average blood pressure after treatment is lower than before.

Limitations of Hypothesis Testing

Although hypothesis testing is a useful technique but it have some limitations as well:

- **Limited Scope:** Hypothesis testing focuses on specific questions or assumptions and not capture the complexity of the problem being studied.

- **Data Quality Dependence:** The accuracy of the results depends on the quality of the data. Poor-quality or inaccurate data can lead to incorrect conclusions.
- **Missed Patterns:** By focusing only on testing specific hypotheses important patterns or relationships in the data might be missed.
- **Context Limitations:** It doesn't always consider the bigger picture which can oversimplify results and lead to incomplete insights.
- **Need for Additional Methods:** To get a better understanding of the data hypothesis testing should be combined with other analytical methods such as data visualization or machine learning techniques which we study later in upcoming articles.

Use of Skewness, Central Limit Theorem

In data science, we often work with samples instead of entire populations. The Central Limit Theorem (CLT) helps us assume normality in sample means, while Hypothesis Testing helps us make decisions based on those samples. These concepts are essential for tasks like A/B testing, model evaluation, and user behavior analysis.

Skewness in Data Science

Skewness tells us whether the data is symmetric or tilted to one side. Understanding skewness is important before applying statistical methods that assume normality.

Example: Income Distribution

Suppose a company's employee salaries are: [25k, 28k, 30k, 32k, 35k, 36k, 100k]

Most salaries are around 25k–36k, but one outlier (100k) pulls the average to the right.

This is a positively skewed distribution. The mean will be higher than the median due to the high-income outlier.

Central Limit Theorem (CLT) in Data Science

The Central Limit Theorem says that if we take many large samples and find their averages, those averages form a normal distribution even if the original data is skewed. This allows data scientists to use normal distribution formulas to calculate confidence intervals and error margins, even when the raw data is not normal.

Example: Estimating Average Daily Orders

A delivery company wants to estimate the average daily orders without checking the full month's data.

- Data (10 days): [100, 200, ..., 700], Population Mean = 360
- Sample Means: 262.5, 412.5, 425
- Avg. of Sample Means = 366.7 (close to 360)

This shows how the Central Limit Theorem helps small samples give a reliable estimate of the overall mean.

Use cases

- Estimate delivery time from few days
- Analyze average spending from user samples
- Approximate server time without full logs

Hypothesis Testing in Data Science

Hypothesis testing checks if a sample result is statistically different from a known or expected value. It helps in deciding whether to reject the null hypothesis based on a threshold called the significance level (commonly 0.05).

Example: Has the Complaint Rate Decreased?

A company claims its new packaging reduced complaints. Earlier, 10% of orders had complaints. In a recent sample of 1000 orders, only 70 complaints were reported.

- H_0 : Complaint rate = 10%
- H_1 : Complaint rate < 10%
- Sample rate = 0.07
- $z \approx -3.16$, Critical $z = -1.645$

Since $-3.16 < -1.645$, we reject H_0 there's strong evidence the complaint rate has dropped.

Applications

- Test if a website change improves clicks
- Compare two ML models
- Check if discounts increase sales
- Evaluate impact of process updates

Z-test : Formula, Types

A **Z-test** is a type of hypothesis test that compares the sample's average to the population's average and calculates the **Z-score** and tells us how much the sample average is different from the population average by looking at **how much the data normally varies**. It is particularly useful when the sample size is large >30 . This Z-Score is also known as Z-Statistics formula is:

$$Z\text{-Score} = \frac{\bar{x} - \mu}{\sigma}$$

where,

- \bar{x} : mean of the sample.
- μ : mean of the population.
- σ : Standard deviation of the population.

Let's understand with the help of example The average family annual income in India is 200k with a standard deviation of 5k and the average family annual income in Delhi is 300k. Then Z-Score for Delhi will be.

$$Z\text{-Score} = \frac{300 - 200}{5} = 20$$

This indicates that the average family's annual income in Delhi is 20 standard deviations above the mean of the population (India).

For a z-test to provide reliable results these assumptions must be met:

1. **Normal Distribution:** The population from which the sample is drawn should be approximately normally distributed.
2. **Equal Variance:** The samples being compared should have the same variance.
3. **Independence:** All data points should be independent of one another.

Steps to perform Z-test

1. First we identify the null and alternate hypotheses.
2. Then we determine the level of significance (α).
3. Next we find the critical value of Z in the z-test.
4. Then we calculate the z-test statistics using the formula :

$$Z = (\sigma/n)(x - \mu)$$

Where:

- x^- : mean of the sample.
- μ : mean of the population.
- σ : Standard deviation of the population.
- n : sample size.

5. Now we compare with the hypothesis and decide whether to reject or not reject the null hypothesis.

Type of Z-test

There are mainly two types of Z-tests. Let's understand them one by one:

1. One Sample Z test

A one-sample Z-test is used to determine if the mean of a single sample is significantly different from a known population mean. **When to Use:**

- The population standard deviation is known.
- The sample size is large (usually $n > 30$).
- The data is approximately normally distributed.

Suppose a company claims that their new smartphone has an average battery life of 12 hours. A consumer group tests 100 phones and finds an average battery life of 11.8 hours with a known population standard deviation of 0.5 hours.

Step 1: Hypotheses:

- $H_0: \mu = 12$:
- $H_1: \mu \neq 12$

Step2: Calculate the Z-Score:

We can calculate Z-score using the formula:

$$z = n\sigma x - \mu$$

Where: $x = 11.8$, $\mu = 12$, $\sigma = 0.5$ and $n = 100$

After putting the value we get:

$$z = 1000.511.8 - 12 = -4$$

Step3: Decision

Since $|Z| = 4 > 1.96$ (critical value for $\alpha = 0.05$) we reject H_0 indicate significant evidence against the company's claim.

Now let's implement this in Python using the Statsmodels and Numpy Library:

```
import numpy as np
from statsmodels.stats.weightstats import ztest

data = [11.8] * 100
population_mean = 12
population_std_dev = 0.5

z_statistic, p_value = ztest(data, value=population_mean)

print(f"Z-Statistic: {z_statistic:.4f}")
print(f"P-Value: {p_value:.4f}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The average battery life is different from 12 hours.")
else:
    print("Fail to reject the null hypothesis: The average battery life is not significantly different from 12 hours.")

Output:
```

Z-Statistic: -560128131373970.2500

P-Value: 0.0000

Reject the null hypothesis: The average battery life is different from 12 hour

2. Two-sampled z-test

In this test we have provided 2 normally distributed and independent populations and we have drawn samples at random from both populations. Here we consider μ_1 and μ_2 to be the population mean and X_1 and X_2 to be the observed sample mean. Here our null hypothesis could be like this:

- $H_0: \mu_1 - \mu_2 = 0$ and alternative hypothesis
- $H_1: \mu_1 - \mu_2 \neq 0$

and the formula for calculating the z-test score:

$$Z = n_1 \sigma_1^2 + n_2 \sigma_2^2 (X_1 - X_2) - (\mu_1 - \mu_2)$$

where σ_1 and σ_2 are the standard deviation and n_1 and n_2 are the sample size of population corresponding to μ_1 and μ_2 . Let's look at the example to understand:

Example: There are two groups of students preparing for a competition: Group A and Group B. Group A has studied offline classes, while Group B has studied online classes. After the examination the score of each student comes. Now we want to determine whether the online or offline classes are better.

- **Group A:** Sample size = 50, Sample mean = 75, Sample standard deviation = 10
- **Group B:** Sample size = 60, Sample mean = 80, Sample standard deviation = 12

Assuming a 5% significance level perform a two-sample z-test to determine if there is a significant difference between the online and offline classes.

Solution:

Step 1: Null & Alternate Hypothesis

- **Null Hypothesis:** There is no significant difference between the mean score between the online and offline classes
 $\mu_1 - \mu_2 = 0$
- **Alternate Hypothesis:** There is a significant difference in the mean scores between the online and offline classes.
 $\mu_1 - \mu_2 \neq 0$

Step 2: Significance Level

- Significance Level: 5%
 $\alpha = 0.05$

Step 3: Z-Score

$$Z\text{-score} = n_1 \sigma_1^2 + n_2 \sigma_2^2 (x_1 - x_2) - (\mu_1 - \mu_2) = 50(102 + 60(122)(75 - 80) - 0 = 2 + 2.4 - 5 = 2.0976 - 5 = -2.384$$

Step 4: Check to Critical Z-Score value in the Z-Table for alpha/2 = 0.025

- Critical Z-Score = 1.96

Step 5: Compare with the absolute Z-Score value

- $\text{absolute}(Z\text{-Score}) > \text{Critical Z-Score}$
- So we reject the null hypothesis and there is a significant difference between the online and offline classes.

Now we will implement the two sampled z-test using numpy and scipy.

```
import numpy as np
import scipy.stats as stats

n1 = 50
```

```

x1 = 75
s1 = 10

n2 = 60
x2 = 80
s2 = 12

D = 0
alpha = 0.05

z_score = ((x1 - x2) - D) / np.sqrt((s1**2 / n1) + (s2**2 / n2))
print('Z-Score:', np.abs(z_score))

z_critical = stats.norm.ppf(1 - alpha/2)
print('Critical Z-Score:', z_critical)

if np.abs(z_score) > z_critical:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")

```

Output:

Z-Score:		2.3836564731139807
Critical	Z-Score:	1.959963984540054
Reject the null hypothesis.		

So, There is a significant difference between the online and offline classes.

The Z-Table

You must calculate your z-statistic using the formula then compare it at the difference significance levels

If you have a two tailed example you compare your z value to the correct significance level on $z_{1-\alpha/2}$. For example: if your two tailed z statistic was 1.98, for the 5% level you would compare it to 1.96 as it is greater this is significant!

Significance level	10%	5%	1%
$z_{1-\alpha}$	1.28	1.645	2.33
$z_{1-\alpha/2}$	1.645	1.96	2.58

These are the most common significance levels, although there are others!

If you have a one tailed example you compare your z value to the correct significance level on $z_{1-\alpha}$

T-test

The t-test is used to compare the averages of two groups to see if they are significantly different from each other. Suppose you want to compare the test scores of two groups of students:

- **Group 1:** 30 students who studied with Method A.
- **Group 2:** 30 students who studied with Method B.

We use a **t-test** to check if there is a significant difference in the average test scores between the two.

T test

The t-test is part of **Hypothesis testing** where we start with an assumption the null hypothesis that the two group means are the same. Then the test helps you decide if there's enough evidence to reject that assumption and conclude that the groups are different.

Assumptions in T-test

- **Independence:** The observations within each group must be independent of each other means that the value of one observation should not influence the value of another observation.
- **Normality:** The data within each group should be approximately normally distributed i.e., the data within each group being compared should resemble a normal bell-shaped distribution.
- **Homogeneity of Variances:** The variances of the two groups being compared should be equal. This assumption ensures that the groups have a similar spread of values.
- **Absence of Outliers:** There should be no outliers in the data as outliers can influence the results especially when sample sizes are small.

Types of T-tests

There are three types of t-tests and they are categorized as dependent and independent t-tests.

Types of t-test

1. One sample T-test

One sample t-test is used for comparison of the sample mean of the data to a particularly given value. We can use this when the sample size is small. (under 30) data is collected randomly and it is approximately normally distributed. It can be calculated as:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Mean of the sample \bar{x} Reference Value
t = Standard deviation
 Number of cases

One sample t-test

Example: The weights of 25 obese people were taken before enrolling them into the nutrition camp. The population mean weight is found to be 45 kg before starting the camp. After finishing the camp for the same 25 people the sample mean was found to be 75 with a standard deviation of 25. Did the fitness camp work?

Python Implementation

Before its implementation we should have some basic knowledge about [numpy](#) and [scipy](#).

```
import scipy.stats as stats
import numpy as np

popu_mean = 45
s_mean = 75
s_std = 25
s_size = 25

t_stat = (s_mean - popu_mean) / (s_std / np.sqrt(s_size))
df = s_size - 1
```

```

alpha = 0.05

cr_t = stats.t.ppf(1 - alpha, df)

p_v = 1 - stats.t.cdf(t_stat, df)

print("T-Statistic:", t_stat)
print("Critical t-value:", cr_t)
print("P-Value:", p_v)

print('With T-value :')

if t_stat > cr_t:
    print("Significant difference. Camp had effect.")

else:
    print("No significant difference. Camp had no effect.")

print('With P-value :')

if p_v > alpha:
    print("Significant difference. Camp had effect.")

else:
    print("No significant difference. Camp had no effect.")

```

Output:

<i>T-Statistic:</i>	6.0			
<i>Critical</i>	<i>t-value:</i>	1.7108820799094275		
<i>P-Value:</i>	1.703654035845048e-06			
<i>With</i>	<i>T-value</i>	:		
<i>Significant</i>	<i>difference.</i>	<i>Camp</i>	<i>had</i>	<i>effect.</i>
<i>With</i>	<i>P-value</i>	:		
<i>No significant difference. Camp had no effect.</i>				

Interpretation

- T-value (6.0) is much greater than the critical t-value (1.71), so we reject the null hypothesis.
- The P-value (0.0000017) is much smaller than 0.05, also indicating a significant result.

The fitness camp had a significant effect on participants weights, causing a measurable change.

2. Independent sample T-test

An Independent sample t-test commonly known as an unpaired sample t-test is used to find out if the differences found between two groups is actually significant or just a random occurrence. We can use this when:

- the population mean or standard deviation is unknown. (information about the population is unknown)
- the two samples are separate/independent. For i.e. boys and girls (the two are independent of each other)

It can be calculated using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Mean of the sample 1 Mean of the sample 2
 \bar{X}_1 \bar{X}_2
 Standard deviation sample 1 and 2
 s_1^2 s_2^2
 Number of cases sample 1 and 2 n_1 n_2

Let's Take a example to understand

Researchers want to see if two teaching methods, A and B, produce different exam scores. Samples for both methods are collected independently.

```
Sample A (Teaching Method A): 78,84,92,88,75,80,85,90,87,7978,84,92,88,75,80,85,90,87,79
Sample B (Teaching Method B): 82,88,75,90,78,85,88,77,92,8082,88,75,90,78,85,88,77,92,80
```

Python Implementation

```
from scipy import stats
import numpy as np

A
np.array([78, 84, 92, 88, 75, 80, 85, 90, 87, 7978, 84, 92, 88, 75, 80, 85, 90, 87, 79]) =
B
np.array([82, 88, 75, 90, 78, 85, 88, 77, 92, 8082, 88, 75, 90, 78, 85, 88, 77, 92, 80]) =
t_val, p_val = stats.ttest_ind(A, B)
```

```

alpha = 0.05

df = len(A)+len(B)-2

crit_t = stats.t.ppf(1 - alpha/2, df)

print("T-value:", t_val)
print("P-Value:", p_val)
print("Critical t-value:", crit_t)

print('T-test Result:')

if np.abs(t_val) > crit_t:
    print('Significant difference found.')
else:
    print('No significant difference.')

print('P-test Result:')

if p_val > alpha:
    print('Fail to reject H0. No strong evidence of difference.')
else:
    print('Reject H0. Significant difference found.')

```

Output:

<i>T-value:</i>	-0.008275847896130646
<i>P-Value:</i>	0.9934425963209128
<i>Critical</i>	<i>t-value:</i> 2.0280940009804502
<i>T-test</i>	<i>Result:</i> difference.
<i>No</i>	<i>significant</i>
<i>P-test</i>	<i>Result:</i>

Fail to reject H0. No strong evidence of difference.

Interpretation

- T-value (0.989) is less than the critical t-value (2.1009), so we fail to reject the null hypothesis.
- P-value (0.336) is greater than 0.05, meaning no significant difference.

There is no statistically significant difference between exam scores of Teaching Method A and Teaching Method B.

3. Paired Two-sample T-test

Paired sample t-test also known as dependent sample t-test is used to find out if the difference in the mean of two samples is 0. The test is done on dependent samples usually focusing on a particular group of people or things. In this each entity is measured twice resulting in a pair of observations.

We can use this when:

- Two similar samples are given. [i.e. Scores obtained in English and Math (both subjects)]
- The dependent variable data is continuous.
- The observations are independent of one another.
- The dependent variable is approximately normally distributed.

It can be calculated using

paired t-test

Example Problem

Consider the following example. Scores (out of 25) of the subjects Math1 and Math2 are taken for a sample of 10 students. We have to perform the paired sample t-test for this data.

```
Math1:      4,      4,      7,     16,     20,     11,     13,      9,     11,     15  
Math2:15, 16, 14, 14, 22, 22, 23, 18, 18, 19
```

Python Implementation

```
from scipy import stats  
import numpy as np  
  
A = np.array([4, 4, 7, 16, 20, 11, 13, 9, 11, 15])  
B = np.array([15, 16, 14, 14, 22, 22, 23, 18, 18, 19])  
  
t_val, p_val = stats.ttest_rel(A, B)  
  
alpha = 0.05  
df = len(A)-1  
  
c_t = stats.t.ppf(1 - alpha/2, df)  
  
print("T-value:", t_val)  
print("P-Value:", p_val)  
print("Critical t-value:", c_t)  
  
print('T-test:')
```

```

if np.abs(t_val) >c_t:
    print('Significant difference found.')
else:
    print('No significant difference.')

print('P-test:')
if p_val >alpha:
    print('Reject H0')
else:
    print('Fail to reject H0')

```

Output:

<i>T-value:</i>	-4.953488372093023	
<i>P-Value:</i>	0.0007875235561560145	
<i>Critical</i>	<i>t-value:</i>	2.2621571628540993
<i>T-test:</i>		
<i>Significant</i>	<i>difference</i>	<i>found.</i>
<i>P-test:</i>		
<i>Fail to reject H0</i>		

Interpretation

- T-value (-4.95) is less than the negative critical t-value (-2.2622), so we reject the null hypothesis.
- P-value (0.00079) is less than 0.05, indicating significance.

Difference and Use of T-Test and Z-Test

Z-Test Vs T-Test

Some of the common difference between Z-test and T-test are:

Aspect	T-Test	Z-Test
Purpose	Compare means of small samples ($n < 30$)	Compare means of large samples ($n \geq 30$)
Assumptions	Normally distributed data, approximate normality	Normally distributed data, known population standard deviation

Population Standard Deviation	Unknown	Known
Sample Size	Small ($n < 30$)	Large ($n \geq 30$)
Test Statistic	T-distribution	Standard normal distribution (Z-distribution)
Degrees of Freedom	$n_1 + n_2 - 2$	Not applicable
Use Case	Small sample analysis, comparing means between groups	Large sample analysis, population mean comparisons
One-Sample vs. Two-Sample	Both	Usually two-sample
Data Requirement	Raw data	Raw data
Complexity	Relatively more complex	Relatively simpler

Z-Test in Data Science

In data science, we often deal with real-world questions like: "Did a website update improve load time?" or "Is the change in user conversion meaningful?" To answer such questions statistically we use z-tests and t-tests.

Example: Delivery Time Analysis

Suppose the company claims average delivery takes 2 days. You check a sample of 50 deliveries that averaged 1.8 days, with a known population standard deviation of 0.5.

- Null hypothesis (H_0): Mean = 2
- $z = (1.8 - 2) / (0.5 / \sqrt{50}) = -2.83$

Since the z-score is quite far from zero, we reject H_0 . It means delivery time has likely improved.

Such tests are useful when analyzing large product usage data or detecting performance drops across thousands of users.

T-Test in Data Science

T-test is widely used for:

- A/B testing, where we compare two versions of a webpage or product feature
- Evaluating user behavior patterns from small experimental groups
- Testing model performance differences

Example: Website Load Time

You want to test if a new version of a website loads faster. You test it on 10 users and get:

- Sample mean = 2.1s
- Old mean = 2.5s
- Sample SD = 0.3s

Using a t-test: $t = (2.1 - 2.5) / (0.3 / \sqrt{10}) = -4.21$

Chi-square test in Data Science

Chi-Square test helps us determine **if there is a significant relationship between two categorical variables** and the target variable. It is a non-parametric statistical test meaning it doesn't follow normal distribution.

Example of Chi-square test

The Chi-square test compares the observed frequencies (actual data) to the expected frequencies (what we would expect if there was no relationship). This helps identify which features are important for predicting the target variable in machine learning models.

Formula for Chi-square test

Chi-square statistic is calculated as:

$$\chi^2 = \sum E_i (O_i - E_i)^2 \dots \text{eq(1)}$$

where,

- c is degree of freedom
- O_i is the observed frequency in cell i
- E_i is the expected frequency in cell i

Often used with **non-normally distributed data**. Before we jump into calculations. let's understand some important terms:

- **Observed Values (O):** Actual counts from the data.
- **Expected Values (E):** Counts expected if variables are independent.
- **Contingency Table:** A table showing counts of two categorical variables.
- **Degrees of Freedom (df):** Number of independent values, helps find critical values.

Types of Chi-Square test

The two main types are the chi-square test for independence and the chi-square goodness-of-fit test.

Types of chi-square tests

1. Chi-Square Test for Independence: This test is used whether there is a significant relationship between two categorical variables.

- This test is applied when we have counts of values for two nominal or categorical variables.
- To conduct this test two requirements must be met: independence of observations and a relatively large sample size.

- We test if shopping preference (Electronics, Clothing, Books) is related to payment method (Credit Card, Debit Card, PayPal). The **null hypothesis** assumes no relationship between them.

2. Chi-Square Goodness-of-Fit Test: The Chi-Square Goodness-of-Fit test is used to check if a variable follows a specific expected pattern or distribution.

- This test is used with counts of categorical data to see if the observed values match what we expect based on a hypothesis. It helps determine if the data represents the whole population well.
- For example, when testing if a six-sided die is fair, the null hypothesis assumes each face has an equal chance of landing face up meaning the die is unbiased and all sides occur equally often.

Steps to perform Chi-square test

Step 1: Define Your Hypotheses

- **Null Hypothesis (H_0):** The two variables are **independent** (no relationship).
- **Alternative Hypothesis (H_1):** The two variables are **related** (there is a relationship).

Step 2: Create a Contingency Table: This is simply a table that displays the frequency distribution of the two categorical variables.

Step 3: Calculate Expected Values: To find the expected value for each cell use this formula:

$$E_i = \frac{\text{Grand Total} \times \text{Row Total}}{\text{Column Total}}$$

Step 4: Compute the Chi-Square Statistic: Now use the Chi-Square formula:

$$\chi^2 = \sum E_i (O_i - E_i)^2$$

where:

- O_i = Observed value
- E_i = Expected value

If the observed and expected values **are very different** the **Chi-Square value will be high** which indicate a **strong relationship**.

Step 5: Compare with the Critical Value:

- If $\chi^2 > \text{critical value}$ → Reject H_0 (There is a relationship).
- If $\chi^2 < \text{critical value}$ → Fail to reject H_0 (No relationship).

Why do we use the Chi-Square Test?

The Chi-Square Test helps us find relationships or differences between categories. Its main uses are:

1. **Feature Selection in Machine Learning:** It helps decide if a categorical feature (like color or product type) is important for predicting the target (like sales or satisfaction), improving model performance.
2. **Testing Independence:** It checks if two categorical variables are related or independent. For example, whether age or gender affects product preferences.
3. **Assessing Model Fit:** It helps check if a model's predicted categories match the actual data, which is useful to improve classification models.

Example: Income Level vs Subscription Status

Let us examine a dataset with features including "income level" (low, medium, high) and "subscription status" (subscribed, not subscribed) indicate whether a customer subscribed to a service. The goal is to determine if this feature is relevant for predicting subscription status.

Step 1: Make Hypothesis

- Null hypothesis: No significant association between features
- Alternate Hypothesis: There is a significant association between features.

Step 2: Contingency table

Income Level	Subscribed	Not subscribed	Row Total
Low	20	30	50
Medium	40	25	65
High	10	15	25
Column Total	70	70	140

Step 3: Now calculate the expected frequencies: For example the expected frequency for "Low Income" and "Subscribed" would be:

- As Total count for each row R_i is 70 and each column C_j is 70 and Total number of observations are 140.
- Low Income, subscribed = $(50 \times 70) / 140 = 25$

Similarly we can find expected frequencies for other aspects as well:

	Subscribed	Not Subscribed
Low Income	25	25
Medium Income	35	30
High Income	10	15

Step 4: Calculate the Chi-Square Statistic: Let's summarize the observed and expected values into a table and calculate the Chi-Square value:

	Subscribed (O)	Not Subscribed (O)	Subscribed (E)	Not Subscribed (E)
Low Income	20	30	25	25
Medium Income	40	25	35	30
High Income	10	15	10	15

Now using the formula specified in equation 1 we can get our chi-square statistic values in the following manner:

$$\chi^2 = 25(20-25)^2 + 25(30-25)^2 + 35(40-35)^2 + 30(25-30)^2 + 10(10-10)^2 + 15(15-15)^2$$

$$= 1 + 1.2 + 0.714 + 0.833 + 0 + 0 = 3.747$$

Step 5: Degrees of Freedom

$$\text{Degrees of Freedom (df)} = (3-1) \times (2-1) = 2$$

Step 6: Interpretations

Now compare the calculated χ^2 value (3.747) with the critical value for 2 degrees of freedom. If χ^2 is greater than the critical value, reject the null hypothesis. This means "income level" is significantly related to "subscription status" and is an important feature. Before its implementation we should have some basic knowledge about numpy, matplotlib and scipy.

```
import numpy as np  
import matplotlib.pyplot as plt  
import scipy.stats as stats
```

```
df = 2  
alpha = 0.05
```

```
critical_value = stats.chi2.ppf(1 - alpha, df)
```

```
critical_value
```

Output:

```
5.991464547107979
```

For $df = 2$ and significance level $\alpha=0.05$, the critical value is 5.991.

- Since $3.747 < 5.991$, we **fail to reject** the null hypothesis.
- **Conclusion:** No significant association between income level and subscription status.

Visualizing Chi-Square Distribution

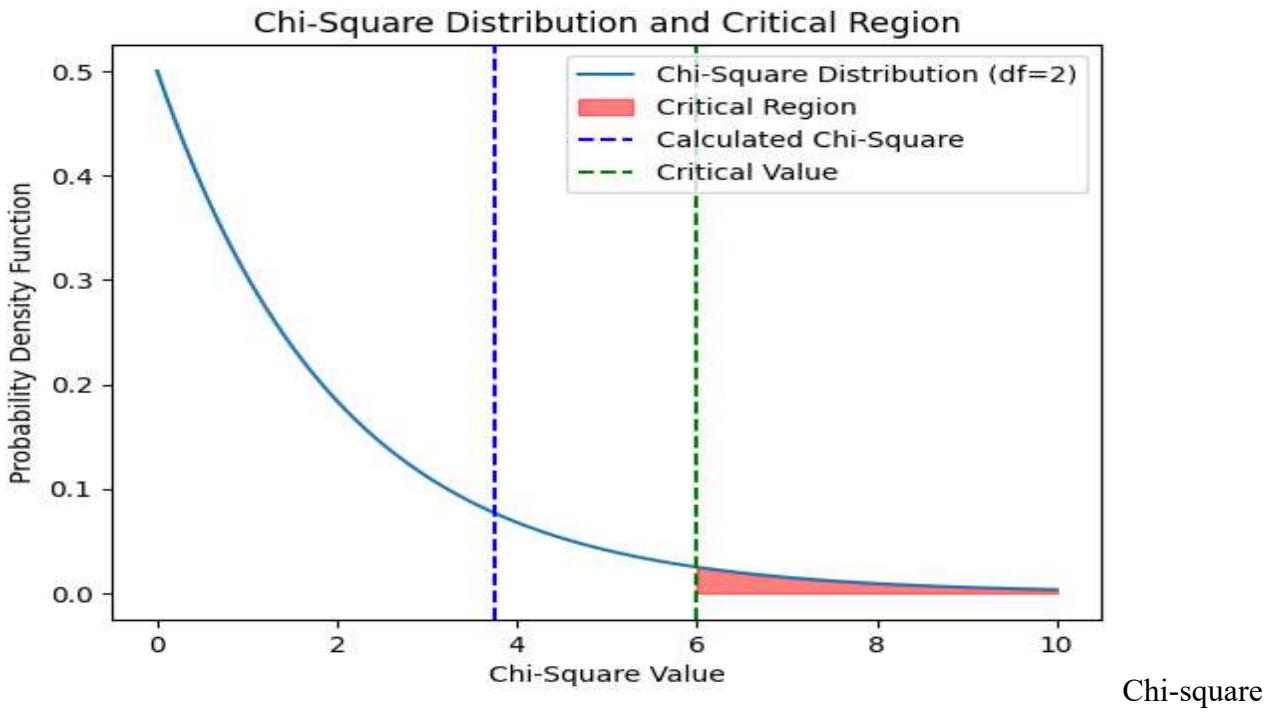
```
import numpy as np  
import matplotlib.pyplot as plt  
import scipy.stats as stats
```

```
df = 2  
alpha = 0.05  
c_val = stats.chi2.ppf(1 - alpha, df)  
cal_chi_s = 3.747
```

```
x = np.linspace(0, 10, 1000)  
y = stats.chi2.pdf(x, df)
```

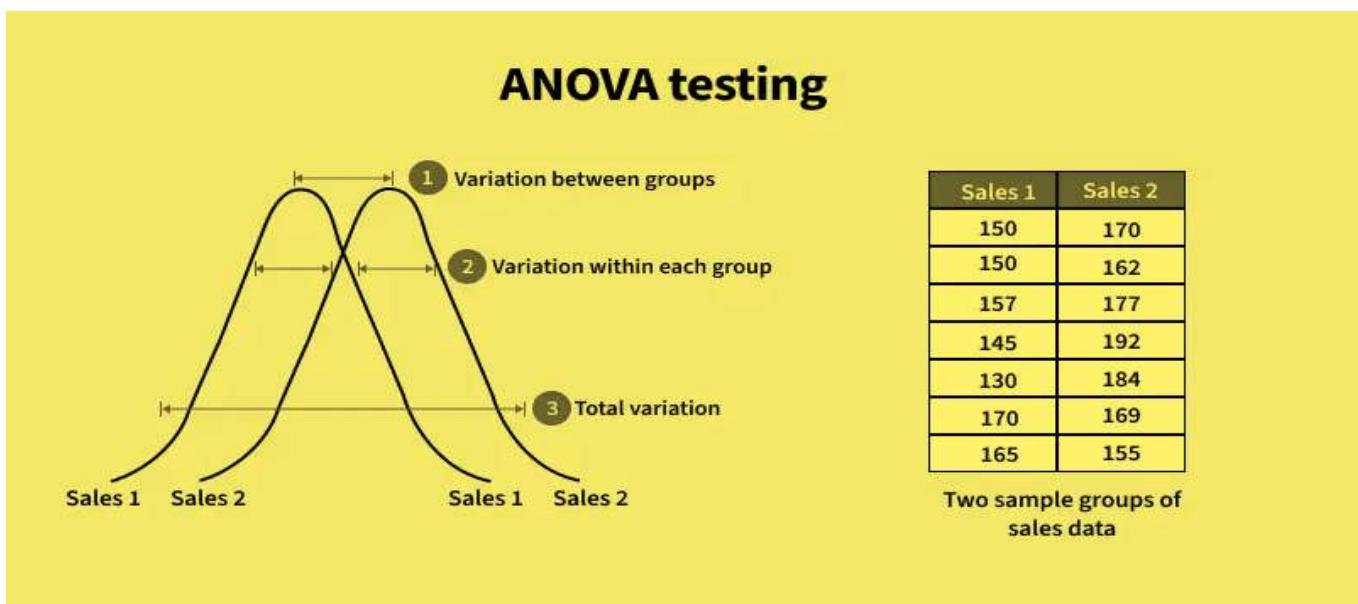
```
plt.plot(x, y, label='Chi-Square Distribution (df=2)')  
plt.fill_between(x, y, where=(x > c_val), color='red', alpha=0.5, label='Critical Region')  
plt.axvline(cal_chi_s, color='blue', linestyle='dashed', label='Calculated Chi-Square')  
plt.axvline(c_val, color='green', linestyle='dashed', label='Critical Value')  
plt.title('Chi-Square Distribution and Critical Region')  
plt.xlabel('Chi-Square Value')  
plt.ylabel('Probability Density Function')  
plt.legend()  
plt.show()
```

Output:



ANOVA for Data Science and Data Analysis

ANOVA is useful when we need to compare more than two groups and determine whether their means are significantly different. Suppose you're trying to understand which ingredients in a recipe affect its taste. Some ingredients, like spices might have a strong influence while others like a pinch of salt might not change much.



ANOVA testing

In machine learning, features act like these ingredients they contribute differently to the final prediction. Instead of guessing, we need a way to measure which features matter most. This is where ANOVA (Analysis of Variance) comes in. It helps us determine if differences in feature values lead to meaningful changes in the target variable, guiding us in selecting the most relevant features for our model.

Understanding ANOVA with a Real-World Example

Let's say we have three schools: **School A, School B and School C**. We collect test scores from students in each school and calculate the average score for each group. The key question is:

Do students from at least one school perform significantly differently from the others?

To answer this ANOVA uses hypothesis testing:

- **Null Hypothesis (H_0):** There is no significant difference between the mean scores of the three schools.
- **Alternative Hypothesis (H_1):** At least one school's mean score is significantly different from the others.

ANOVA does not tell us which group is different it only tells us a difference exists. If the p-value from the ANOVA test is less than 0.05 we reject the null hypothesis and conclude that at least one group has a significantly different mean score.

Key Assumptions of ANOVA

For ANOVA to work effectively three important assumptions must be met:

1. Independence of Observations:

- Each data point should be independent of others.
- In our example one student's test score should not influence another student's score.

2. Homogeneity of Variances (Equal Variance):

- The variation in scores across all groups should be roughly the same.
- If one school's scores vary widely while the others have similar scores ANOVA results may be unreliable.

3. Normal Distribution:

- The data within each group should follow a normal distribution.
- If the data is highly skewed it can not work well.

How ANOVA Test Works?

To understand how ANOVA works let's go through it step by step focusing on key concepts with the help of a example.

Step 1. Calculate Group Means

First we calculate the mean for each group. Let's say you are comparing smartphone prices from three brands: Brand A, Brand B and Brand C. Let's assume the following data for the smartphone prices:

- **Brand A:** [200, 210, 220, 230, 250]
- **Brand B:** [180, 190, 200, 210, 220]
- **Brand C:** [210, 220, 230, 240, 250]

Now we calculate the mean for each brand:

- Mean of Brand A = $(200 + 210 + 220 + 230 + 250) / 5 = 222$
- Mean of Brand B = $(180 + 190 + 200 + 210 + 220) / 5 = 200$
- Mean of Brand C = $(210 + 220 + 230 + 240 + 250) / 5 = 230$

Step 2. Calculate Overall Mean

Next we calculate the overall mean.

$$\text{Overall mean} = (200 + 210 + 220 + 230 + 250 + 180 + 190 + 200 + 210 + 220 + 210 + 220 + 230 + 240 + 250) / 15 = 215$$

Step 3. Calculate variances:

There are basically two methods to calculate the variance of the data:

1. Within-group variance: This measures how much the scores in a group differ from the group's average. If scores are close to the average, the variance is small. If scores are spread out, the variance is large. The formula for calculation is :

$$\text{Within-group variance} = n - 1 \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$$

Where:

- X_{ij} = individual prices
- \bar{X}_i = mean of the group
- n = number of prices in the group

For Brand A: Prices: [200, 210, 220, 230, 250] and Mean: $\bar{X}=222$

The squared differences are:

- $(210-230)^2 = (-20)^2 = 400$
- $(220-230)^2 = (-10)^2 = 100$
- $(230-230)^2 = (0)^2 = 0$
- $(240-230)^2 = (10)^2 = 100$
- $(250-230)^2 = (20)^2 = 400$

Sum of squared differences = $484 + 144 + 4 + 64 + 784 = 1480$

Now calculate the variance for Brand A:

- Variance for Brand A = $5 - 1480 = 41480 = 370$

similarly we will calculate for both Brand B and Brand C and we get:

- Variance for Brand B = $5 - 11000 = 41000 = 250$
- Variance for Brand C = $5 - 11000 = 41000 = 250$

2. Between-group variance: It measures how much the **group means** differ from the overall mean. If the group means are far apart then the variance will be large. If the group means are close to each other the variance will be small. To calculate this we use the formula:

$$\text{Between-group variance} = k - 1 \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

Where:

- n_i is the number of data points in each group (5 in each group),

- $\bar{X_i}$ is the mean of each group,
- \bar{X} is the overall mean.

Step-by-step Calculation:

- For Brand A: $(\bar{X_A} - \bar{X})^2 = (222 - 215)^2 = (7)^2 = 49$

Contribution to between-group variance: $5 \times 49 = 245$

- For Brand B: $(\bar{X_B} - \bar{X})^2 = (200 - 215)^2 = (-15)^2 = 225$

Contribution to between-group variance: $5 \times 225 = 1125$

- For Brand C: $(\bar{X_C} - \bar{X})^2 = (230 - 215)^2 = (15)^2 = 225$

Contribution to between-group variance: $5 \times 225 = 1125$

Sum of contributions: Between-group variance = $3 - 1245 + 1125 + 1125 = 22495 = 1247.5$

Step 4. F-Ratio Calculation

Once we have the within-group and between-group variances we calculate the **F-ratio** by dividing the between-group variance by the within-group variance:

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} = \frac{1247.5}{2901} = 4.3$$

- **A high F-ratio** suggests that the between-group variance is much larger than the within-group variance. This means that the groups are significantly different from each other.
- **A low F-ratio** indicates that the groups are not very different from each other.

Step 5. Interpreting the F-Ratio

To understand the results of the F-ratio we compare it to a **critical value** from the F-distribution table.

- If the F-ratio is greater than the critical value it indicates that there is a **significant difference** between at least one group's mean and the others and we **reject the null hypothesis**.
- On the other hand if the F-ratio is small we **fail to reject the null hypothesis** means there is not enough evidence to say that the group means are different.

The **F-ratio** is 4.3 which we can compare to a critical value from the **F-distribution table** based on the degrees of freedom:

- Degrees of freedom for the numerator ($df_{between}$): $k-1 = 3-1 = 2$
- Degrees of freedom for the denominator (df_{within}): $n-k = 15-3 = 12$

If the calculated **F-ratio** is greater than the critical value from the table (which depends on the significance level usually 0.05), we **reject the null hypothesis** and conclude that there are significant differences between the group means

Types of ANOVA Tests

ANOVA has two main types: **one-way** and **two-way** depending on how many independent variables are involved.

1. One-Way ANOVA

This test is used when we have one independent variable with two or more groups. It helps check if at least one group is different from the others. Imagine we are comparing the average prices of smartphones from three brands: Brand A, Brand B, and Brand C and we have Independent variable: Brand (A, B, and C) and Dependent variable is Smartphone price.

Firstly We set up two hypotheses:

- **Null Hypothesis (H_0):** All brands have the same average price.
- **Alternative Hypothesis (H_1):** At least one brand has a different average price.

ANOVA helps determine if the price differences are due to real variation between brands or just random chance. However it only considers **one factor (brand)** at a time. If we want to check **multiple factors** we use two-way ANOVA.

2. Two-Way ANOVA

A **two-way ANOVA** is used when we have **two independent variables** which allow us to analyze their individual effects and their interaction.

Two way Anova

For example suppose we want to see how brand and storage capacity (64GB, 128GB, 256GB) affect smartphone prices.

- Factor 1: Brand (A, B, C)
- Factor 2: Storage capacity
- Dependent variable: Price

Using two-way ANOVA, we test:

- Does brand affect price?
- Does storage size affect price?
- Does the effect of storage size depend on the brand? (interaction effect)

If there's an interaction, it means one factor's effect changes depending on the other. For example, Brand A's prices rise with more storage, but Brand C's prices stay the same.

In machine learning, detecting interactions can help create new features (like brand \times storage) to improve predictions. This helps us understand how brand and storage together influence price.

ANOVA for Feature Selection in Machine Learning

ANOVA is also used in machine learning for feature selection. When building a model, not all features help predict the target. ANOVA helps find important numerical features when the target is categorical (like "Yes" or "No"). Feature selection makes the model simpler, faster, and more accurate.

For example, a teacher wants to know if study hours, assignments, or attendance impact student grades (A, B, C, D). The ANOVA F-test (like Scikit-learn's `f_classif`) checks if the average values of a feature differ across target groups.

How it works:

- The F-test checks if the feature's means differ across groups (e.g., study hours across grades).
- If there's a big difference, the feature is important; if not, it's less important.

The test gives an F-statistic and a p-value:

- **Low p-value (< 0.05)** = important feature
- **High p-value** = less important, can be removed

This helps pick the best features for the model.

Difference between One way Anova and Two way Anova

The difference between the Oneway Anova and Two way anova is given below:

Aspect	One way Anova	Two way Anova
Number of Independent Variables	It have only one independent Variable	It have two independent variable
Purpose	Tests if there's a significant difference in means across multiple groups based on one factor.	Tests if there's a significant difference in means based on two factors, and their interaction.
Usage	Used when selecting features where a single categorical factor affects a numerical feature like the effect of study hours on student grades .	Used when analyzing the effect of two categorical factors and their interaction on a numerical feature e.g., how both study hours and school type impact grades .
Example	It is used in comparing average sales across different types of advertising (TV, online, print).	Used in Comparing sales based on advertising type (TV, online, print) and sales region (East, West, North, South).
Complexity	It is a simple test.	It is more complex involves two factors and interaction terms.

Use of Chi-Square and ANOVA

In data science, we often deal with comparisons comparing categories or checking if multiple groups behave differently. Chi-Square and ANOVA (Analysis of Variance) are two key statistical tools used for this.

Chi-Square Test in Data Science

The Chi-Square Test is used when:

- Both variables are categorical (e.g., gender, product type).

- We want to see if there is a relationship between them.

Example: Customer Preference by Gender

Suppose a store wants to know if product choice depends on gender.

	Product A	Product B
Male	30	20
Female	10	40

Using the Chi-Square test, we check if this difference is by chance or if gender really affects choice.

In Data Science:

- Used in market basket analysis.
- Check if click rates differ by ad type or region.
- See if fraud is more likely in specific transaction types.

ANOVA in Data Science

ANOVA is used when:

- You compare means of 3 or more groups.
- You want to know if at least one group is significantly different.

Example: Ad Performance on 3 Platforms

You run an ad on Facebook, Instagram and YouTube. You measure click rates:

- Facebook: 2.5%
- Instagram: 2.8%
- YouTube: 3.6%

ANOVA helps check if the click rate difference is statistically significant or just random.

In Data Science:

- Compare performance of multiple ML models.
- Analyze user engagement across multiple regions.
- Test sales across different store locations.

Confidence Interval

A **Confidence Interval (CI)** is a range of values that contains the true value of something we are trying to measure like the average height of students or average income of a population.

Instead of saying: "The average height is 165 cm."

We can say: "We are 95% confident the average height is between 160 cm and 170 cm."

Before diving into confidence intervals you should be familiar with:

- t-test
- z-test

Interpreting Confidence Intervals

Let's say we take a sample of 50 students and calculate a 95% confidence interval for their average height which turns out to be 160–170 cm. This means If we repeatedly take similar samples 95% of those intervals would contain the true average height of all students in the population.

Confidence Interval

Confidence level tells us how sure we are that the true value is within a calculated range. If we have to repeat the sampling process many times we expect that a certain percentage of those intervals will include the true value.

Confidence Level	Meaning
90%	90 out of 100 intervals will include the true value
95%	95 out of 100 intervals will include the true value (most commonly used)
99%	99 out of 100 intervals will include the true value (more conservative)

Formula

$$\text{Confidence Level} = 1 - \alpha$$

Where α is the **significance level** (commonly 0.05 for 95% CI).

Why are Confidence Intervals Important in Data Science?

- They helps to **measure uncertainty** in predictions and estimates.
- Through this data scientists finds the reliable results instead of just giving a single number.
- They are widely used in A/B testing, machine learning, and survey analysis.

Steps for Constructing a Confidence Interval

To calculate a confidence interval follow these simple 4 steps:

Step 1: Identify the sample problem.

Define the population parameter you want to estimate e.g., mean height of students. Choose the right statistic such as the **sample mean**.

Step 2: Select a confidence level.

In this step we select the confidence level some common choices are **90%, 95% or 99%**. It represents how sure we are about our estimate.

Step 3: Find the margin of error.

To find the **Margin of Error**, you use the formula:

$$\text{Margin of Error} = \text{Critical Value} \times \text{Standard Error}$$

- **Critical Value:** Found using Z-tables (for large samples) or T-tables (for small samples).
- **Standard Error (SE):** Measures how much the sample mean varies.

$$SE = n \text{Standard Deviation}$$

Combine these to get your **Margin of Error** the amount you add/subtract from your estimate to create a range.

Step 4: Specify the confidence interval.

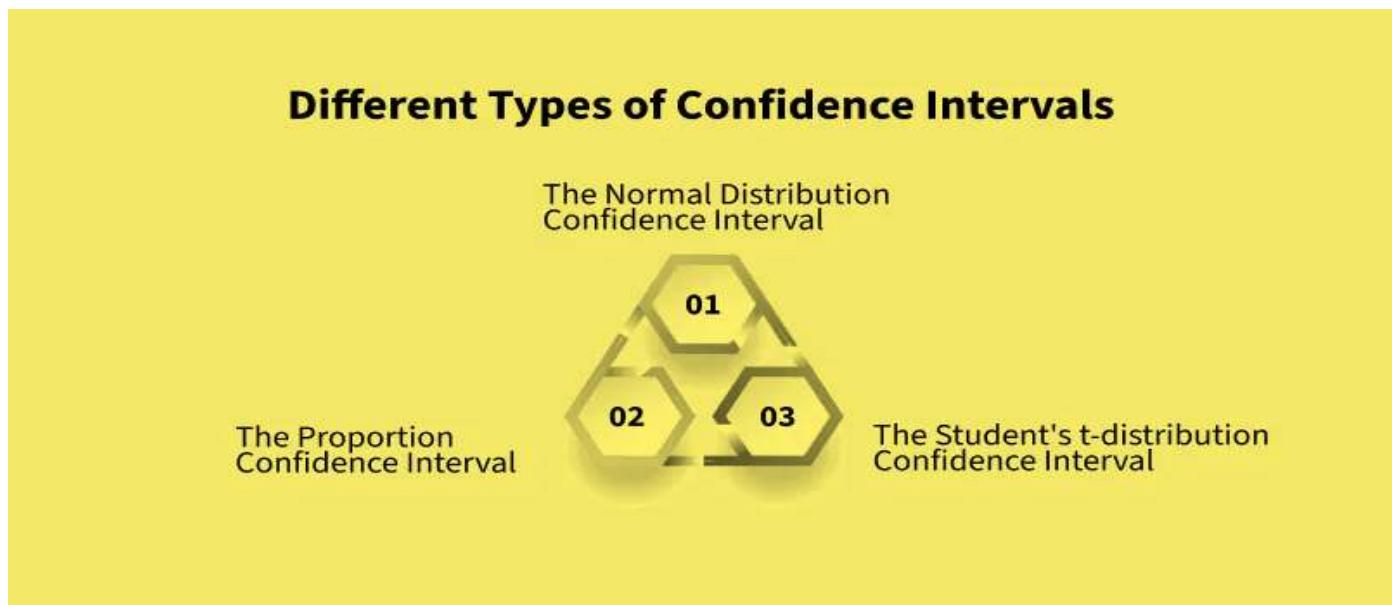
To find a **Confidence Interval**, we use this formula:

$$\text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin of Error}$$

- The Point Estimate is usually your sample mean.
- Adding and subtracting the margin of error gives the range where the true value is likely to be.

Types of Confidence Intervals

Some of the common types of Confidence Intervals are:



Types of confidence Interval

1. Confidence Interval for the Mean of Normally Distributed Data

When we want to find the mean of a population based on a sample we use this method.

- If the **sample size is small** (less than 30) we use the **T-distribution**.
- If the **sample size is large** (more than 30) then we use the **Z-distribution**.

2. Confidence Interval for Proportions

This type is used when estimating population proportions like the percentage of people who like a product. Here we use the **sample proportion**, the **standard error** and the **critical Z-value** to calculate the interval. It gives us the idea where the real value could fall based on sample data.

3. Confidence Interval for Non-Normally Distributed Data

If your data isn't normally distributed (doesn't follow a bell curve), use bootstrap methods:

- Randomly resample the data many times
- Calculate intervals from these resamples

This gives a good estimate even if the data is skewed or irregular.

For Calculating Confidence Interval

To calculate a confidence interval you need two key statistics:

- **Mean (μ):** Average of all sample values
- **Standard Deviation (σ):** Shows how much values vary from the mean

Once you have these you can calculate the confidence interval either using **t-distribution** or **z-distribution** depend on the sample size whether the population standard deviation is known.

A) Using t-distribution

Used when:

- Sample size is small
- Population standard deviation is unknown

Example:

Sample size = 10

Mean weight = 240 kg

Std deviation = 25 kg

Confidence Level = 95%

Step-by-Step Process:

- **Degrees of Freedom (df):** For t-distribution we first calculate the degrees of freedom: $df=n-1=10-1=9$
- **Significance Level (α):** The confidence level (CL) is 95% so the significance level is: $\alpha=1-CL=1-0.95=0.025$
- **Find t-value from t-distribution table:** From the t-table for $df = 9$ and $\alpha = 0.025$ the t-value is 2.262 which can be find using the below table.

(df)/(α)	0.1	0.05	0.025	..
∞	1.282	1.645	1.960	..
1	3.078	6.314	12.706	..
2	1.886	2.920	4.303	..
:	:	:	:	..
8	1.397	1.860	2.306	..
9	1.383	1.833	2.262	..

- **Apply t-value in the formula:**
The formula for the confidence interval is: $\mu \pm t(n\sigma)$ Using the values: $240 \pm 2.262 \times (10 \times 25)$
- The **confidence interval** becomes: (222.117, 257.883)

Therefore we are **95% confident** that the true mean weight of UFC fighters is between **222.117 kg and 257.883 kg**.

This can be calculated using Python's `scipy` and `math` library to find the t-value and perform the necessary calculations. The `stats` module provides various statistical functions, probability distributions, and statistical tests.

```
import scipy.stats as stats
import math

mean = 240
std = 25
n = 10
df = n - 1
alpha = 0.025
t = stats.t.ppf(1 - alpha, df)

moe = t * (std / math.sqrt(n))

lower = mean - moe
upper = mean + moe

print(f"Confidence Interval: ({lower:.2f}, {upper:.2f})")
```

Output:

```
Confidence Interval: (222.1160773511857, 257.8839226488143)
```

B) Using Z-distribution

Used when:

- Sample size is large
- Population standard deviation is known

Consider the following example. A random sample of 50 adult females was taken and their RBC count is measured. The sample mean is 4.63 and the standard deviation of RBC count is 0.54. Construct a 95% confidence interval estimate for the true mean RBC count in adult females.

Step-by-Step Process:

1. **Find the mean and standard deviation** given in the problem.
2. **Find the z-value for the confidence level:**
For a 95% confidence interval the z-value is 1.960.
3. **Apply z-value in the formula:** $\mu \pm z(n\sigma)$

Using the values: some common values in the table given below:

Confidence Interval	z-value
90%	1.645
95%	1.960

99%

2.576

The confidence interval becomes: (4.480, 4.780)

Therefore we are **95% confident** that the true mean RBC count for adult females is between **4.480** and **4.780**.

Now let's do the implementation of it using Python. But before its implementation we should have some basic knowledge about numpy and scipy.

```
from scipy import stats  
import numpy as np
```

```
mean = 4.63
```

```
std_dev = 0.54
```

```
n = 50
```

```
z = 1.960
```

```
se = std_dev / np.sqrt(n)
```

```
moe = z * se
```

```
lower = mean - moe
```

```
upper = mean + moe
```

```
print(f"Confidence Interval: ({lower:.3f}, {upper:.3f})")
```

Output:

Confidence Interval: (4.480, 4.780)

A/B Testing using Python

A/B testing is a way to compare two versions of something to find out which one works better. In this you divide people into two groups, show them different versions and then measure which version performs better based on a specific goal. Suppose you're sending out two different email subject lines to people and you want to see which one gets more people to open the email.

- **Group A:** Gets an email with the subject "50% Off This Weekend!"
- **Group B:** Gets an email with the subject "Special Deal Just for You!"

A/B testing example

After sending you count how many people open each email. If more people open the email from Group B you can decide that the second subject line is better. As you can see, Option B works better than Option A because more people responded to it (25% compared to 17%).

When to Use A/B Test

- **Not Getting Good Results:** If something in your campaign isn't working well, try out different versions with A/B testing to find what needs fixing.
- **Starting Something New:** Before launching a new page or message, test a couple of versions to see which one works better.

Key terminologies used in A/B Testing

To understand more about **A/B testing** first you have to learn these concepts:

1. Hypothesis Testing

Before you start any A/B test you need to come up with a **hypothesis**. Think of it as a smart guess about what you believe will happen in the experiment. For example if you're testing two versions of a website button then your hypothesis would be: "I think changing the color of the button from blue to green will make more people click it. A clear hypothesis gives your test direction.

2. Randomization

Next we need to make sure the users are split into two groups: **the control group and the experimental group**. This is where randomization comes in. The control group (A) will see the original version of what you're testing. The experimental group (B) will see the new or changed version. It is done to avoid bias in test results.

3. Sample Size

Sample size means how many people you need to include in your test. You want enough people to get reliable results but not too many that it wastes resources. The more people you test the more accurate your results will be.

4. Performance Metrics

Now that your test is set up you need to decide what you're measuring. These are called **performance metrics** or **KPIs (Key Performance Indicators)**. These are the things you'll look at to see if your changes worked. Some common performance metrics include:

- **Conversion Rate:** It is like how many people took the action you wanted like buying a product, signing up for a newsletter.
- **Average Order Value:** How much on average people spend during a transaction.
- **User Retention:** It shows how many people come back to use your product after their first visit. They help you measure success. Without them you wouldn't know if the change you made actually improved anything.

5. Statistical Analysis

Finally once you've collected the data from your test you need to analyze it to see if the changes you made were really effective. This is done using **statistical methods**.

Types of A/B Tests

1. One-Sample A/B Test (Single Model Comparison)

- One-Sample A/B test is used to compare a new model (test) against a baseline model (control). You can test whether the performance of the new model is significantly better than the old one.

- **Example:** A company wants to test a new version of its customer churn prediction model against the existing model.

2. Two-Sample A/B Test (Comparing Two Models)

- In this test you compare the performance of two different models to see if one outperforms the other.
- **Example:** You might test two recommendation algorithms (A and B) to see which one produces better user engagement or conversion rates.

Steps to Conduct an A/B Test

Let's take a real world example to understand the A/B Testing. Suppose you are working for an e-commerce company wants to improve user engagement by testing a new machine learning-based recommendation system against their current rule-based system. They need to determine if the new model actually improves engagement before fully implementing it.

Step 1: Define Your Hypothesis

Before running the A/B test you must clearly define what you are testing and how success will be measured.

- **Good Hypothesis:** "The new ML-based recommendation system will increase the click-through rate (CTR) by at least 15% compared to the existing rule-based system."
- **Bad Hypothesis:** "The new recommendation system might work better."

Step 2: Set Up Control and Test Groups

A/B testing requires splitting users into two groups randomly:

- **Control Group:** Users who see recommendations from the existing rule-based system.
- **Test Group:** Users who see recommendations from the new ML-based system.

*The groups must be of **similar size** to ensure statistical validity.*

Step 3: Collect Data

Once the A/B test is live we need to track key performance indicators (KPIs) that help us measure the impact of the change. Common KPIs include:

- **Click-Through Rate (CTR):** It measures user engagement by calculating the percentage of users who click on recommended items.

The formula of CTR is:

$$CTR = \frac{\text{Number of Clicks}}{\text{Number of Impressions}}$$

Example: If 1,000 users see recommendations and 150 click on them then CTR is:

$$\frac{150}{1000} = 15\%$$

- **Conversion Rate (CR):** Measures how many users make a purchase after clicking a recommendation.

$$CR = \frac{\text{Number of Purchases}}{\text{Number of Clicks}}$$

- **Bounce Rate:** The percentage of users who leave without interacting.
Data collection should run long enough to **capture a representative sample** of user behavior.

Step 4: Analyze the Results Using Python

Once we have collected sufficient data we need to analyze whether the observed differences between the control and test groups are statistically significant.

Key statistical measures used:

- **Average Performance:** It compare CTR between groups.
- **Confidence Interval (CI):** Confidence Interval indicates the range within which the true effect likely falls.
- **Statistical Significance (p-value):** Statistical Significance determines if the difference is due to chance.

Before its implementation we should have some basic knowledge about numpy and scipy.

```
import numpy as np

import scipy.stats as stats


cc = 1200 # control clicks
ci = 10000 # control impressions

tc = 1500 # test clicks
ti = 10000 # test impressions

ctr_c = cc / ci
ctr_t = tc / ti

table = np.array([[cc, ci - cc],
                  [tc, ti - tc]])

chi2, p, _, _ = stats.chi2_contingency(table)

print(f"Control CTR: {ctr_c:.2%}")
print(f"Test CTR: {ctr_t:.2%}")
print(f"Chi-Square Test p-value: {p:.5f}")

if p < 0.05:
```

```

    print("The difference is statistically significant. Implement the new
recommendation system.")

else:
    print("No significant difference. Further testing needed.")

```

Output:

Control CTR: 12.00%

Test CTR: 15.00%

Chi-Square Test p-value: 0.00000

The difference is statistically significant. Implement the new recommendation system.

Step 5: Make a Decision

After analyzing the results there are two possible outcomes:

1. If the test group performs significantly better ($p < 0.05$):

- The ML-based recommendation system should replace the existing rule-based system.
- Deploy the new model for all users.

2. If results are inconclusive ($p > 0.05$):

- The observed difference may be due to randomness.
- Further testing or model improvements may be needed.

Tools for A/B Testing

Several tools make running A/B tests easier and more effective:

- **Google Optimize:** A free tool that integrates with Google Analytics. Great for basic A/B testing and audience targeting within the Google ecosystem.
- **Optimizely:** A premium platform offering advanced features like multivariate testing, cross-channel experiments, and real-time results. Ideal for large-scale, complex testing needs.
- **VWO (Visual Website Optimizer):** Another paid tool combining A/B testing with extras like heatmaps and session recordings for deeper user insights and improved conversions.

Differentiation Formulas

Differentiation is the mathematical process of determining the derivative of a function, representing the rate at which the function's value changes with respect to its independent variable. The derivative, denoted as $\frac{dx}{df(x)}$, provides a precise measure of the function's immediate rate of change. This operation forms the basis of differential calculus, with specific formulas and rules applicable to algebraic, trigonometric, exponential, logarithmic and inverse trigonometric functions.

Derivative Formulas

Derivative Formulas

Power Formula

$$\frac{d}{dx} (x^n) = nx^{n-1}$$

Sum Formula

$$\frac{d}{dx} [f(x)+g(x)] = f'(x) + g'(x)$$

Difference Formula

$$\frac{d}{dx} [f(x)-g(x)] = f'(x) - g'(x)$$

Chain Formula

$$\frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$$

Product Formula

$$\frac{d}{dx} [f(x)g(x)] = f(x)g'(x) + g(x)f'(x)$$

Constant Multiple Formula

$$\frac{d}{dx} [cf(x)] = cf'(x)$$

Quotient Formula

$$\frac{d}{dx} \left[\frac{f(x)}{g(x)} \right] = \left[\frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2} \right]$$

Table of Content

- [Basic Differentiation Formulas](#)
- [Differentiation of Trigonometric Functions](#)
- [Differentiation of Inverse Trigonometric Functions](#)
- [Differentiation of Hyperbolic Functions](#)
- [Differentiation Rules](#)
- [Differentiation of Special Functions](#)
- [Implicit Differentiation](#)
- [Solved Examples of Differentiation Formulas](#)
- [Practice Problems on Differentiation Formulas](#)

The derivative of $f(x)$ at x is given by the limit as h approaches 0:

$$dx dy = \lim_{h \rightarrow 0} (x+h) - xf(x+h) - f(x)$$

Mathematically,

$$dx dy = f'(x) = \lim_{h \rightarrow 0} h f(x+h) - f(x)$$

This limit represents the instantaneous rate of change of y with respect to x or the slope of the tangent line to the curve $y = f(x)$ at the point $(x, f(x))$.

Differentiation formulas are used to find the differentiation of the various functions. The first principal formula states that, for any function $f(x)$ its derivative with respect to x is,

$$f'(x) = \lim_{h \rightarrow 0} h f(x+h) - f(x)$$

Basic Differentiation Formulas

The differentiation formulas for some elementary functions are:

Function (y)	Differentiation Formula (dy/dx)
--------------	-------------------------------------

c (constant)	0
x^n (power)	nx^{n-1}
$\ln x$ (logarithmic)	$1/x$
e^x(exponent)	e^x
a^x (exponent)	$a^x \log a$

Differentiation of Trigonometric Functions

Derivatives of the trigonometric functions are:

Function (y)	Derivative (dy/dx)
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\tan x$	$\sec^2 x$
$\sec x$	$\sec x \cdot \tan x$
$\operatorname{cosec} x$	$-\operatorname{cosec} x \cdot \cot x$
$\cot x$	$-\operatorname{cosec}^2 x$

Differentiation of Inverse Trigonometric Functions

The differentiation formulas for the Inverse trigonometric functions are:

Function (y)	Differentiation Formula (dy/dx)
$\sin^{-1} x$	$1/\sqrt{1-x^2}$
$\cos^{-1} x$	$-1/\sqrt{1-x^2}$
$\tan^{-1} x$	$1/(1+x^2)$
$\sec^{-1} x$	$1/(x \cdot\sqrt{x^2-1})$
$\operatorname{cosec}^{-1} x$	$-1/(x \cdot\sqrt{x^2-1})$
$\cot^{-1} x$	$-1/(1+x^2)$

Differentiation of Hyperbolic Functions

Let's discuss the Differentials of Hyperbolic functions.

Function (y)	Differentiation Formula (dy/dx)
$\sinh x$	$\cosh x$
$\cosh x$	$\sinh x$
$\tanh x$	$\operatorname{sech}^2 x$
$\operatorname{sech} x$	$-\operatorname{sech} x \cdot \tanh x$
$\operatorname{cosech} x$	$-\operatorname{cosech} x \cdot \coth x$
$\coth x$	$-\operatorname{cosech}^2 x$

Differentiation Rules

Various rules of finding the derivative of functions have been given below:

Rules	Function Form (y)	Differentiation Formula (dy/dx)	Formula

Sum Rule	$u(x) \pm v(x)$	$du/dx \pm dv/dx$
Product Rule	$u(x) \times v(x)$	$u dv/dx + v du/dx$
Quotient Rule	$u(x) \div v(x)$	$(v du/dx - u dv/dx) / v^2$
Chain Rule	$f(g(x))$	$f'[g(x)] g'(x)$
Constant Rule	$k f(x), k \neq 0$	$k d/dx f(x)$

Differentiation of Special Functions

If we have two parametric functions $x = f(t)$, $y = g(t)$, where t is the parameter, then the differentiation of parametric functions is as follows,

As $dy/dt = g'(t)$ and $dx/dt = f'(t)$ then dy/dx is given by:

$$dx dy = dx/dt dy/dt = f'(t)g'(t)$$

Implicit Differentiation

If y is related to x but can not conveniently expressed in the form $y = f(x)$ but can be expressed in the form $f(x,y) = 0$, then we say that y is an implicit function of x . In the case of implicit function dy/dx can be found by following steps.

- Differentiate each term of $f(x, y) = 0$ with respect to x .
- Collect the terms containing dy/dx on one side and the terms not involving dy/dx on the other side.
- Express dy/dx as a function of x or y or both.

Example: Find the differentiation of $x^2 + y^2 + 4xy = 0$

Solution:

$$x^2 + y^2 + 4xy = 0$$

Differentiating with respect to x ,

$$\begin{array}{ccccccccc} 2x & + & 2ydy/dx & + & 4(xdy/dx) & + & y) & = & 0 \\ \boxed{2} & 2x & + & 4y & + & 2dy/dx(y & + & 2x) & = & 0 \\ \boxed{2} & x & + & 2y & + & dy/dx(y & + & 2x) & = & 0 \\ \boxed{2} & dy/dx(y & + & 2x) & = & -(x & + & 2y) & & \\ \boxed{2} dy/dx = -(x + 2y)/(y + 2x) & & & & & & & & \end{array}$$

Higher Order Differentiation

Higher order differentiation is nothing, but the differentiation of a function more than one time suppose we have a function $y = f(x)$ then its differential in higher order is calculated as,

$$\text{First Derivative} = dy/dx = f'(x)$$

$$\text{Second Derivative} =$$

$$\text{Third Derivative} = dx^3 d^3 y = f'''(x)$$

....
....
....

n^{th} Derivative = $\frac{d}{dx} \cdots \frac{d}{dx} f(x) = f(n)(x)$

This can be understood using the example added below,

Example: Find the second-order derivative of $f(x) = 4x^4 + 3x^3 + 2x^2 + x + 1$

Solution:

$$f(x) = 4x^4 + 3x^3 + 2x^2 + x + 1$$

Differentiating with respect to x ,

$$\begin{aligned} f'(x) &= 4(4x^3) + 3(3x^2) + 2(2x) + 1 + 0 \\ f'(x) &= 16x^3 + 9x^2 + 4x + 1 \end{aligned}$$

For second-order derivative differentiating with respect to x ,

$$\begin{aligned} f''(x) &= 16(3x^2) + 9(2x) + 4 + 0 \\ f''(x) &= 48x^2 + 18x + 4 \end{aligned}$$

This is the required second-order derivative.

Solved Examples of Differentiation Formulas

Let's solve some example problems on the rules of derivative.

Example 1: Find the differentiation of $y = 4x^3 + 7x^2 + 11x + 12$

Solution:

$$\text{Given, } y = 4x^3 + 7x^2 + 11x + 12$$

Differentiating with respect to x ,

$$\frac{dy}{dx} = 4(3x^2) + 7(2x) + 11(1) + 0$$

$$f'(x) = 12x^2 + 14x + 11$$

This is the required differentiation

Example 2: Find the differentiation of $y = \cos(\log x)$

Solution:

$$\text{Given, } y = \cos(\log x)$$

Differentiating with respect to x ,

$$\frac{dy}{dx} = \frac{d}{dx} \{ \cos(\log x) \}$$

$$\begin{aligned} \frac{dy}{dx} &= \sin(\log x) \cdot \frac{1}{x} \\ dy/dx &= \sin(\log x) \cdot (1/x) \end{aligned}$$

This is the required differentiation

Gradient

The gradient is a fundamental concept in calculus that extends the idea of a derivative to multiple dimensions. It plays a crucial role in vector calculus, optimization, machine learning, and physics. The gradient of a function provides the direction of the steepest ascent, making it essential in areas such as gradient descent in machine learning and optimization problems.

Mathematical Definition

Given a scalar function $f(x_1, x_2, \dots, x_n)$ of multiple variables, the gradient is defined as a vector of its partial derivatives:

$$\nabla f = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n})$$

where each component $\frac{\partial f}{\partial x_i}$ represents the rate of change of f with respect to x_i .

Gradient in Two and Three Dimensions

For a function $f(x, y)$, the gradient is:

$$\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$$

For a function $f(x, y, z)$, the gradient is:

$$\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z})$$

Geometric Interpretation

1. Direction of Steepest Ascent:

The gradient vector points in the direction where the function increases most rapidly.

2. Magnitude Represents Rate of Change:

The length $\|\nabla f\|$ indicates how steep the function is in that direction.

3. Gradient Perpendicular to Level Curves:

If $f(x, y)$ defines a surface, its gradient at a point is perpendicular to the level curves $f(x, y) = c$, where c is a constant.

Numerical Example

Consider the function:

$$f(x, y) = x^2 + 3y^2$$

Step 1: Compute the Gradient

The partial derivatives are:

$$\partial_x \partial f = 2x, \partial_y \partial f = 6y$$

Thus, the gradient is:

$$\nabla f = (2x, 6y)$$

Step 2: Evaluate at a Point

At $(x,y) = (1,2)$:

$$\nabla f(1,2) = (2(1), 6(2)) = (2, 12)$$

This means the function increases most rapidly in the direction $(2, 12)$.

Python Implementation

We can compute the gradient using [SymPy](#) (for symbolic differentiation) and [NumPy](#) (for numerical computation).

1. Computing the Gradient Symbolically

```
import sympy as sp
```

```
# Define variables
x, y = sp.symbols('x y')
```

```
# Define function
f = x**2 + 3*y**2
```

```
# Compute gradient
grad_f = [sp.diff(f, var) for var in (x, y)]
print("Gradient:", grad_f)
```

Output:

```
Gradient: [2*x, 6*y]
```

2. Evaluating the Gradient at (1,2)

```
# Convert symbolic expressions to functions
grad_f_func = [sp.lambdify((x, y), expr) for expr in grad_f]
```

```
# Evaluate at (1,2)
grad_value = [func(1, 2) for func in grad_f_func]
print("Gradient at (1,2):", grad_value)
```

Output:

```
Gradient at (1,2) : [2, 12]
```

3. Visualizing the Gradient Field

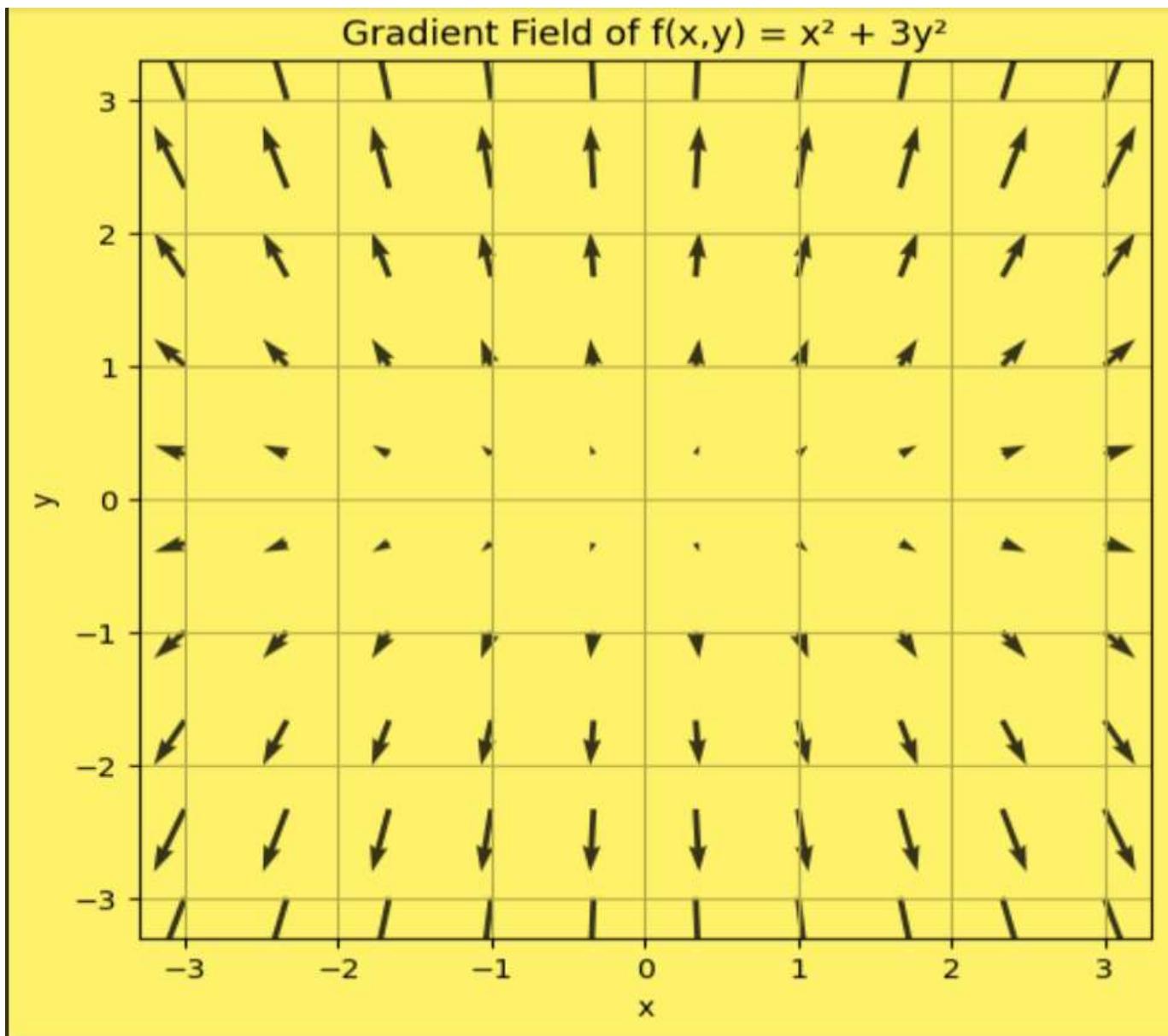
```
import numpy as np
import matplotlib.pyplot as plt

# Generate grid
X, Y = np.meshgrid(np.linspace(-3, 3, 10), np.linspace(-3, 3, 10))

# Compute gradients
U = 2 * X # ∂f/∂x
V = 6 * Y # ∂f/∂y

# Plot gradient field
plt.figure(figsize=(6,6))
plt.quiver(X, Y, U, V, color='r', angles='xy')
plt.xlabel('x')
plt.ylabel('y')
plt.title('Gradient Field of f(x,y) = x2 + 3y2')
plt.grid()
plt.show()
```

Output:



Visualizing Gradient Field

Applications of the Gradient

1. Optimization (Gradient Descent)

In machine learning, the gradient guides gradient descent, an optimization algorithm used to minimize loss functions. The update rule is:

$$\theta \leftarrow \theta - \alpha \nabla f(\theta)$$

where:

- θ are the parameters
- α is the learning rate

2. Physics (Electric and Gravitational Fields)

In electromagnetism and gravity, the gradient of a potential function gives the field direction:

$$E = -\nabla V$$

$$F = -\nabla U$$

where V is electric potential and U is gravitational potential.

3. Computer Vision and Image Processing

The gradient is used for edge detection in images. Operators like Sobel filters compute image gradients to highlight edges.

4. Robotics and Navigation

Robots use gradient-based path planning to navigate toward a goal while avoiding obstacles.

Use of Calculus in Data Science

In data science, calculus plays a big role in optimization and probability. It helps in model training, error minimization and probability distribution analysis. Concepts like differentiation, gradients and Jacobians are used regularly in machine learning and AI.

1. Differentiation in Data Science

Differentiation measures how one value changes with respect to another. In data science, it is used to understand how a small change in input affects the output.

Example: In machine learning, the error between predicted and actual values is minimized by adjusting parameters. Differentiation helps to find the direction in which error decreases fastest.

Use cases

- Optimize error in regression models
- Improve prediction accuracy in ML algorithms
- Sensitivity analysis of input features

2. Partial Derivatives in Data Science

Partial derivatives help when functions depend on multiple variables. They show how the function changes with respect to one variable while keeping others fixed.

Example: In a neural network, each weight and bias influences the output. Partial derivatives show how much each parameter contributes to the error, guiding weight updates.

Use cases

- Training deep learning models
- Feature importance evaluation
- Multi-variable optimization problems

3. Gradient in Data Science

The gradient is a collection of partial derivatives. It points in the direction where the function increases the fastest and its opposite direction helps minimize errors.

Example: When training a model, the gradient indicates how to adjust weights to minimize the loss. Taking steps against the gradient reduces error over time.

Use cases

- Gradient descent optimization
- Backpropagation in neural networks
- Feature optimization in ML pipelines

4. Chain Rule in Data Science

The chain rule helps compute derivatives of functions built from multiple layers. In machine learning this is essential since models often stack functions together.

Example: Neural networks apply multiple functions one after another (layers). The chain rule helps calculate how a change in one layer affects the final output error.

Use cases

- Backpropagation in deep learning
- Complex transformations in feature engineering
- Layer-wise optimization in models

5. Jacobian and Hessian Matrices in Data Science

Jacobian is a matrix of first-order derivatives, useful in multivariate transformations. Hessian is a matrix of second-order derivatives, useful in optimization problems.

Example: In optimization problems like logistic regression or Newton's method, the Jacobian helps transform features while the Hessian improves convergence speed by analyzing curvature.

Use cases

- Multivariate optimization of ML models
- Dimensionality reduction techniques
- Faster convergence in training through second-order methods

Applications of Calculus in Data Science

- Training ML models with gradient descent and backpropagation
- Probability and distribution modeling
- Loss function minimization and optimization
- Hyperparameter tuning in algorithms
- Feature transformations and scaling

MANOVA Test in Data Science

Multivariate Analysis of Variance (MANOVA) is a statistical test used to determine if there are significant differences between multiple groups on multiple dependent variables.

MANOVA Test Example

For example, here three groups taking different medications (shown by the pills). We measure their weight and cholesterol levels to see if the groups differ. MANOVA tests if there are significant differences across these health measures together, helping us understand the overall effect of the treatments.

Unlike ANOVA, which deals with a single dependent variable, MANOVA considers multiple dependent variables simultaneously, thus preserving the relationships between them. It helps in reducing Type I errors, which occur when multiple ANOVA tests are performed separately.

When to Use MANOVA?

MANOVA is appropriate when:

- There are two or more dependent variables that are correlated.
- There are one or more categorical independent variables.
- The data meets the assumptions of normality, homogeneity of variance, and independence of observations.
- The researcher wants to examine the interaction effects between independent variables on multiple dependent variables.

Assumptions of MANOVA

Before conducting a MANOVA test, certain assumptions must be met:

1. **Multivariate Normality:** The dependent variables should be normally distributed for each group.
2. **Homogeneity of Variance-Covariance Matrices:** The variance-covariance matrices of dependent variables should be equal across groups.
3. **Linearity:** The relationships between dependent variables should be linear.
4. **Absence of Multicollinearity:** The dependent variables should not be highly correlated.
5. **Sufficient Sample Size:** The sample size should be adequate to ensure reliable statistical results.

MANOVA Test for Feature Selection

In this implementation, 150 samples with six independent variables and four dependent variables (computed with some noise) are analyzed using MANOVA. The goal is to identify significant relationships between independent and dependent variables, enabling feature selection by retaining key predictors, reducing dimensionality, and enhancing model efficiency.

Before its implementation we should have some basic knowledge about numpy, pandas and statsmodels.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.multivariate.manova import MANOVA
```

```

np.random.seed(42)
num_samples = 150

# Independent Features
X = pd.DataFrame({
    'feature1': np.random.normal(10, 2, num_samples),
    'feature2': np.random.normal(20, 5, num_samples),
    'feature3': np.random.normal(30, 10, num_samples),
    'feature4': np.random.normal(40, 15, num_samples),
    'feature5': np.random.normal(50, 20, num_samples),
    'feature6': np.random.normal(15, 20, num_samples)
})

# Dependent variables
Y = pd.DataFrame({
    'target1': 0.5 * X['feature1'] + 0.2 * X['feature2'] + np.random.normal(0, 1, num_samples),
    'target2': 0.3 * X['feature3'] + 0.1 * X['feature4'] + np.random.normal(0, 1, num_samples),
    'target3': 0.4 * X['feature5'] + 0.3 * X['feature1'] + np.random.normal(0, 1, num_samples),
    'target4': 0.9 * X['feature6'] + 0.3 * X['feature4'] + np.random.normal(0, 1, num_samples)
})

# Combine X and Y into a single DataFrame for MANOVA
data = pd.concat([X, Y], axis=1)

The MANOVA test is performed using MANOVA.from_formula(), where the dependent variables are target1, target2, and target3, and the independent variables are feature1 to feature6.

# MANOVA model
formula = 'target1 + target2 + target3 ~ feature1 + feature2 + feature3 + feature4 + feature5 + feature6'
manova = MANOVA.from_formula(formula, data=data)
results = manova.mv_test()

Features with p-values < 0.05 are selected as significant predictors. This helps in reducing the dimensionality of the dataset while maintaining significant variables for further modeling.

# Extract p-values
p_values = results.results['feature1']['stat']['Pr > F'] # Extracting p-values for feature1
all_features = X.columns.tolist()

# Select significant features (p-value < 0.05)
selected_features = [feature for feature, p in zip(all_features, p_values) if p < 0.05]
print("\nMANOVA Results:\n", results)
print("\nSelected Features Based on MANOVA (p < 0.05):", selected_features)
Output:

```

MANOVA Results:

Multivariate linear model

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9966	3.0000	141.0000	0.1626	0.9214
Pillai's trace	0.0034	3.0000	141.0000	0.1626	0.9214
Hotelling-Lawley trace	0.0035	3.0000	141.0000	0.1626	0.9214
Roy's greatest root	0.0035	3.0000	141.0000	0.1626	0.9214

feature1	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.3775	3.0000	141.0000	77.5035	0.0000
Pillai's trace	0.6225	3.0000	141.0000	77.5035	0.0000
Hotelling-Lawley trace	1.6490	3.0000	141.0000	77.5035	0.0000
Roy's greatest root	1.6490	3.0000	141.0000	77.5035	0.0000

feature2	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.4935	3.0000	141.0000	48.2294	0.0000
Pillai's trace	0.5065	3.0000	141.0000	48.2294	0.0000
Hotelling-Lawley trace	1.0262	3.0000	141.0000	48.2294	0.0000
Roy's greatest root	1.0262	3.0000	141.0000	48.2294	0.0000

manova

result

1

feature3	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.1265	3.0000	141.0000	324.4203	0.0000
Pillai's trace	0.8735	3.0000	141.0000	324.4203	0.0000
Hotelling-Lawley trace	6.9026	3.0000	141.0000	324.4203	0.0000
Roy's greatest root	6.9026	3.0000	141.0000	324.4203	0.0000

feature4	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.3509	3.0000	141.0000	86.9441	0.0000
Pillai's trace	0.6491	3.0000	141.0000	86.9441	0.0000
Hotelling-Lawley trace	1.8499	3.0000	141.0000	86.9441	0.0000
Roy's greatest root	1.8499	3.0000	141.0000	86.9441	0.0000

feature5	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0123	3.0000	141.0000	3784.1680	0.0000
Pillai's trace	0.9877	3.0000	141.0000	3784.1680	0.0000
Hotelling-Lawley trace	80.5142	3.0000	141.0000	3784.1680	0.0000
Roy's greatest root	80.5142	3.0000	141.0000	3784.1680	0.0000

Manova

result

2

feature6	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9859	3.0000	141.0000	0.6704	0.5716
Pillai's trace	0.0141	3.0000	141.0000	0.6704	0.5716
Hotelling-Lawley trace	0.0143	3.0000	141.0000	0.6704	0.5716
Roy's greatest root	0.0143	3.0000	141.0000	0.6704	0.5716

Selected Features Based on MANOVA ($p < 0.05$): ['feature1', 'feature2', 'feature3', 'feature4']

Manova result 3

After running the above script, the output will include:

- **Pillai's Trace**
- **Wilks' Lambda (to determine significance)**
- **Hotelling's Trace**
- **Roy's Largest Root**

These statistics indicate whether there is a statistically significant difference in the dependent variables across different groups. A low p-value (typically < 0.05) suggests that the group differences are significant.

Types of MANOVA

- 1. One-way MANOVA:** Compares means of multiple outcome variables across three or more groups based on one independent variable. Example: Effect of different drug treatments (Pill A, B, or C) on weight and cholesterol.
- 2. Two-way MANOVA:** Compares means across multiple outcome variables using two independent variables. Example: Effects of drug treatment and blood group on weight and cholesterol.
- 3. Repeated Measures MANOVA:** Used when the same individuals are measured multiple times, accounting for repeated data points.

Comparison of One-way and Two-way Manova

The figure illustrates the difference between one-way and two-way MANOVA. One-way MANOVA uses a single independent variable, like drug treatment, to compare multiple outcome variables such as weight and cholesterol levels. In contrast, two-way MANOVA includes two independent variables, for example, drug treatment and blood group, to assess their combined effects on the same outcomes. This allows for a more detailed analysis of how different factors interact. The diagram highlights how explanatory variables connect to multiple dependent variables in each method.

MANOVA vs ANOVA

Here is the difference between **ANOVA** and **MANOVA**:

Feature	ANOVA	MANOVA
Outcome Variables	One	Two or more
Measures	Mean difference in a single variable	Combined mean difference across multiple variables
Use Case	Test score only	Test score + satisfaction rating

Bayesian Statistics & Probability

Bayesian statistics sees unknown values as things that can change and updates what we believe about them whenever we get new information. It uses Bayes' Theorem to combine what we already know with new data to get better estimates. In simple words, it means changing our initial guesses based on the evidence we find. This ongoing update helps us deal with uncertainty and make smarter decisions as more information comes in.

For example, when flipping a coin, usual statistics say there's a 50% chance of heads. But if you already know the coin might be heavier on one side, Bayesian statistics lets you use that knowledge to adjust the chance of heads.

Before understanding Bayes' Theorem, let us first understand conditional probability.

Conditional Probability

Conditional probability is the probability of an event occurring given that another event has already occurred. It is denoted by $P(A|B)$ read as "the probability of event A given event B"

$$P(\theta|X) = P(X)\cdot P(\theta, X)$$

Here:

- $P(\theta|X)$: Probability of parameter θ given observed data X
- $P(\theta, X)$: Joint probability of θ and X
- $P(X)$: Probability of observed data X (marginal)

Bayes' Theorem

Bayes' Theorem is a mathematical formula that describes how to update the probability of a hypothesis based on new evidence. In simple terms it allows us to calculate the **posterior probability** (updated belief) by combining the **prior probability** (prior belief) and the likelihood of observing the evidence.

Mathematically Bayes' Theorem is expressed as:

$$P(\theta|X) = P(X)\cdot P(X|\theta)\cdot P(\theta)$$

Where:

- $P(\theta|X)$ is the **posterior probability** the updated belief after observing the data.
- $P(X|\theta)$ is the **likelihood** the probability of observing the data given the hypothesis.
- $P(\theta)$ is the **prior probability**, our initial belief about the hypothesis before observing the data.
- $P(X)$ is the **marginal likelihood** a normalizing constant that ensures the posterior probability sums to 1.

Bayesian Statistics Components

Bayesian statistics uses three key parts: the likelihood function, prior belief, and posterior belief. These help handle yes/no outcomes and let us update our beliefs as we get new information. Let us understand them one by one:

1. Likelihood Function

The **Bernoulli likelihood function** is used for binary outcomes like **success** or **failure**. Like if we are studying the probability of a customer clicking on an ad (success) or not clicking (failure) this function helps us identify how likely it is to observe specific data given the probability of success.

Mathematically the Bernoulli likelihood function is represented as:

$$P(X|\theta) = \theta^x(1-\theta)^{1-x}$$

Where:

- X represents the observed data (0 for failure and 1 for success).
- θ is the probability of success (e.g., click rate).
- x is the observed outcome (0 for failure, 1 for success).

2. Prior Distribution

Before we observe any data we have some **prior beliefs** about the parameters that we are estimating. For example we might have an initial belief that the probability of a customer clicking on an ad is around 0.3. The prior belief distribution reflects this knowledge. A commonly used probability parameter is the **Beta distribution** which is used as the prior distribution for parameters like θ .

The prior belief distribution is mathematically expressed as:

$$P(\theta) = B(\alpha, \beta) \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Where:

- θ represents the probability of success.
- α and β are parameters that control the shape of the Beta distribution.
- $B(\alpha, \beta)$ is the Beta function which ensures the distribution integrates to 1.

3. Posterior Distribution

Once new data is available we use **Bayes' Theorem** to update our beliefs. The updated belief is represented by the **posterior belief distribution** which combines the prior belief and the new evidence.

$$P(\theta|X) = P(X|\theta) \times P(\theta)$$

The posterior distribution shows the updated probability of success or failure after we observe the data. As we receive new data our beliefs about the parameter will change accordingly

This graph explains how Bayesian statistics update our understanding of relative risk by combining prior beliefs with new data.

- The **green curve** represents the **data** which suggests the possible values for the risk based on observations.
- The **red curve** is the **prior** which **shows** our belief about the risk before seeing the data.
- The **blue curve** is the **posterior** which is the updated belief after combining both.
- A **steeper posterior** means the data has a stronger influence while a **flatter posterior** means the prior has more effect.

Example of Bayesian Statistics and Probability

Suppose a patient takes a test for a disease that affects 5% of the population (prior probability = 0.05).

The test results depend on:

- **Sensitivity:** 95% chance of a positive result if the patient has the disease.
- **False Negative Rate:** 5% chance of a negative result despite having the disease.
- **False Positive Rate:** 10% chance of a positive result without the disease.
- **Specificity:** 90% chance of a negative result if the patient is healthy.

The patient tests positive. Using Bayes' Theorem, we update our belief about the patient having the disease:

$$P(\text{Disease} \mid \text{Positive}) = P(\text{Positive})P(\text{Positive} \mid \text{Disease}) \times P(\text{Disease})$$

Where:

$$P(\text{Positive}) = P(\text{Positive} \mid \text{Disease}) \times P(\text{Disease}) + P(\text{Positive} \mid \text{No Disease}) \times P(\text{No Disease})$$

This calculation helps estimate the true chance the patient has the disease after the positive test.

Why Not Frequentist Approach?

The confusion between frequentist and Bayesian approaches has been constant for beginners. It's important to find the difference between these methods:

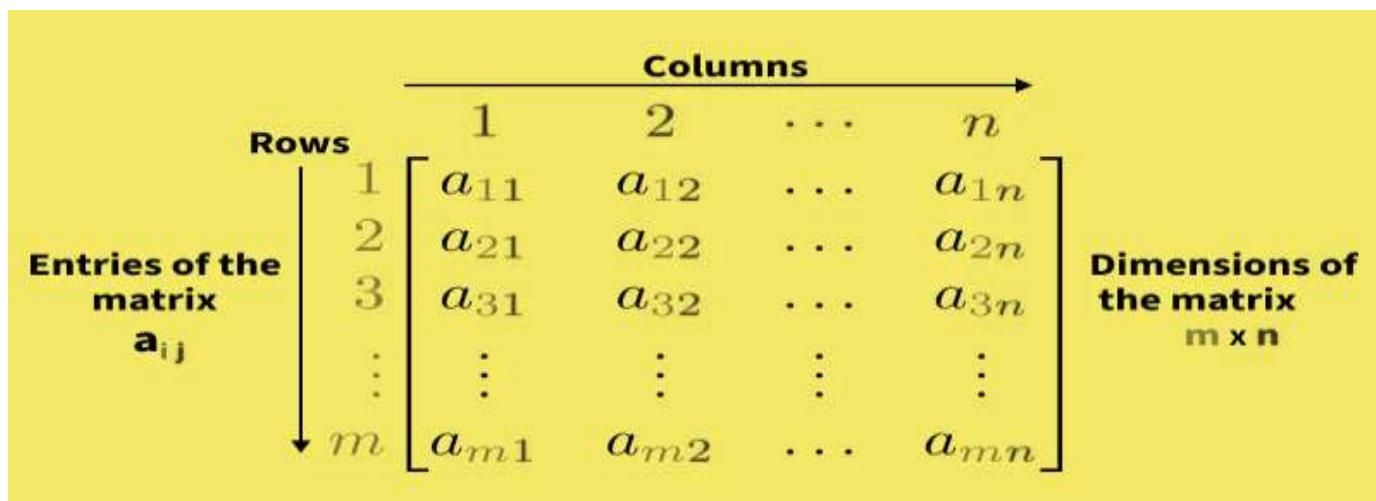
- **Frequentist statistics** relies solely on observed data and long-term frequencies, often ignoring prior knowledge. It uses point estimates and hypothesis testing with p-values, which can lead to rigid decisions.
- **Bayesian statistics** incorporates prior beliefs and updates them as data accumulates, offering more nuanced probability statements. This is especially useful for unique events or when data is limited.

Introduction to Matrices

Matrices are rectangular arrays of numbers, symbols, or characters where all of these elements are arranged in each row and column.

- A matrix is identified by its order, which is given in the form of rows \times and columns, and the location of each element is given by the row and column it belongs to.
- A matrix is represented as $([P]_{m \times n})$, where P is the matrix, m is the number of rows, and n is the number of columns.

Given below is a general example of a matrix:



In mathematics, matrices are mainly used to represent and solve systems of linear equations, perform linear transformations, and study concepts like eigenvalues, determinants, and vector spaces.

Some common examples of matrices are:

$$A = [1324] \text{ } 2 \times 2 \text{ and } B = [134-12-2265] \text{ } 3 \times 3$$

Here, A is a 2×2 matrix (2 rows and 2 columns), and B is a 3×3 matrix (3 rows and 3 columns).

Order of Matrix

The Order Of a Matrix tells about the number of rows and columns present in a matrix. The order of a matrix is represented as the number of rows times the number of columns. Let's say if a matrix has 4 rows and 5 columns, then the order of the matrix will be 4×5 . Always remember that the first number in the order signifies the number of rows present in the matrix, and the second number signifies the number of columns in the matrix.

Operation on Matrices

We can perform various mathematical operations on matrices, such as addition, subtraction, scalar multiplication, and multiplication. These operations are performed between the elements of two matrices to give an equivalent matrix that contains the elements that are obtained as a result of the operation between the elements of two matrices.

Addition of Matrices

In matrix addition or subtraction of matrices, the operation is performed between two matrices of the same order to yield a matrix that contains elements obtained by performing the operations on the elements of the two matrices.

The addition of matrices A and B:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{bmatrix}$$

Example:

Find the sum of [1425] and [2637]

Solution:

Here, we have $A = [1425]$ and $B = [2637]$

$$A + B = [1425] + [2637]$$

$$\therefore A + B = [1+24+62+35+7] = [310512]$$

Subtraction of Matrices

The subtraction of two matrices can be represented in terms of the addition of two matrices. Let's say we have to subtract matrix B from matrix A, then we can write $A - B$. We can also rewrite it as $A + (-B)$.

The subtraction of matrices A and B:

Figure 3: Subtracting two 2×2 matrices.

Example:

Subtract [1425]from [2637].

Solution:

Let us assume $A = [2637]$ and $B = [1425]$

$$A - B = [2637] - [1425]$$

$$\boxed{A - B = [2-16-43-27-5] = [1212]}$$

Scalar Multiplication of Matrices

Scalar Multiplication of matrices refers to the multiplication of each term of a matrix by a scalar. If a scalar let's 'k' is multiplied by a matrix, then the equivalent matrix will contain elements equal to the product of the scalar and the element of the original matrix. Let's see an example:

Figure 4: Multiplying a matrix by a scalar (k)

Example: Multiply

3

[1425].

Solution :

$$3[A] = [3 \times 13 \times 43 \times 23 \times 5]$$

$$\boxed{3[A] = [312615]}$$

Multiplication of Matrices

In the multiplication of matrices, two matrices are multiplied to yield a single equivalent matrix. The multiplication is performed in the manner that the elements of the row of the first matrix multiply with the elements of the columns of the second matrix and the product of elements is added to yield a single element of the equivalent matrix. If a matrix $[A]_{i \times j}$ is multiplied by matrix $[B]_{j \times k}$, then the product is given as $[AB]_{i \times k}$.

Note: Matrix multiplication between $[A]$ and $[B]$ is only possible if no of rows in $[A]$ is equal to number of rows of $[B]$.

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11}+a_{12}b_{21} & a_{11}b_{12}+a_{12}b_{22} \\ a_{21}b_{11}+a_{22}b_{21} & a_{21}b_{12}+a_{22}b_{22} \end{bmatrix}$$

Example:

Find

the

product

of [1425]and [2637]

Solution:

Let $A = [1425]$ and $B = [2637]$

$$\boxed{AB = [1425][2637]}$$

$$\boxed{AB = [1 \times 2 + 2 \times 64 \times 2 + 5 \times 61 \times 3 + 2 \times 74 \times 3 + 5 \times 7]}$$

$$\boxed{AB = [14381747]}$$

Transpose

Transpose of a Matrix is the rearrangement of row elements in columns and column elements in a row to yield an equivalent matrix. A matrix in which the elements of the row of the original matrix are arranged in columns or vice versa is called a Transpose Matrix. The transpose matrix is represented as A^T . If $A = [a_{ij}]_{m \times n}$, then $A^T = [b_{ij}]_{n \times m}$ where $b_{ij} = a_{ji}$.

Figure 6: Transposing a 2×3 matrix to a 3×2 matrix.

Example: The transpose

of $[18381747]$.

Solution:

$$\text{Let } A = [18381747]$$

$$\boxed{A^T = [18173847]}$$

Trace

Trace of a Matrix is the sum of the principal diagonal elements of a square matrix. The trace of a matrix is only found in the case of a square matrix because diagonal elements exist only in square matrices. Let's see an example.

Example: Find the trace of the matrix $\boxed{[147258369]}$

Solution: Let us assume $A = \boxed{\boxed{147258369}}$

$$Trace(A) = 1 + 5 + 9 = 15$$

Types of Matrices

Based on the number of rows and columns present and the special characteristics shown, type of matrices are classified into various types.

- **Row Matrix:** A Matrix that has only one row and one or more columns is called a Row Matrix.
- **Column Matrix:** A matrix that has only one column and one or more rows is called a Column Matrix.
- **Horizontal Matrix:** A Matrix in which the number of rows is less than the number of columns is called a Horizontal Matrix.
- **Vertical Matrix:** A Matrix in which the number of columns is less than the number of rows is called a Vertical Matrix.
- **Rectangular Matrix:** A Matrix in which the number of rows and columns is unequal is called a Rectangular Matrix.
- **Square Matrix:** A matrix in which the number of rows and columns is the same is called a Square Matrix.

- **Diagonal Matrix:** A square matrix in which the non-diagonal elements are zero is called a Diagonal Matrix.
- **Zero or Null Matrix:** A matrix whose all elements are zero is called a Zero Matrix. A zero matrix is also called as Null Matrix.
- **Unit or Identity Matrix:** A diagonal matrix whose all diagonal elements are 1 is called a Unit Matrix. A unit matrix is also called an Identity matrix. An identity matrix is represented by I.
- **Symmetric matrix:** A square matrix is said to be symmetric if the transpose of the original matrix is equal to its original matrix. i.e. $(A^T) = A$.
- **Skew-symmetric Matrix:** A skew-symmetric (or antisymmetric or antimetric[1]) matrix is a square matrix whose transpose equals its negative, i.e., $(A^T) = -A$.
- **Orthogonal Matrix:** A matrix is said to be orthogonal if $AA^T = A^TA = I$
- **Idempotent Matrix:** A matrix is said to be idempotent if $A^2 = A$
- **Involutory Matrix:** A matrix is said to be Involutory if $A^2 = I$.
- **Upper Triangular Matrix:** A square matrix in which all the elements below the diagonal are zero is known as the upper triangular matrix
- **Lower Triangular Matrix:** A square matrix in which all the elements above the diagonal are zero is known as the lower triangular matrix
- **Singular Matrix:** A square matrix is said to be a singular matrix if its determinant is zero, i.e., $|A|=0$
- **Non-singular Matrix:** A square matrix is said to be a non-singular matrix if its determinant is non-zero.

Note: Every Square Matrix can uniquely be expressed as the sum of a symmetric matrix and a skew-symmetric matrix. $A = 1/2 (A^T + A) + 1/2 (A - A^T)$.

Determinant of a Matrix

The Determinant of a matrix is a number associated with that square matrix. The determinant of a matrix can only be calculated for a square matrix. It is represented by $|A|$. The determinant of a matrix is calculated by adding the product of the elements of a matrix with their cofactors.

Figure 8: Determinant of a 3×3 matrix

Example 1: How to find the determinant of a 2×2 square matrix?
Solution :

Let say we have matrix $A = [acbd]$

Then, determinant is of A is $|A| = ad - bc$

Example 2: How to find the determinant of a 3×3 square matrix?
Solution :

Let's say we have a 3×3 matrix $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$

Then $|A| = a(-1)^{1+1} \cdot ehfi - b(-1)^{1+2} \cdot dgfi + c(-1)^{1+3} \cdot dgeh$

Minor of a Matrix

Minor of a matrix for an element is given by the determinant of a matrix obtained after deleting the row and column to which the particular element belongs. Minor of a Matrix is represented by M_{ij} . Let's see an example.

Example: Find the minor of the matrix $\begin{bmatrix} a & d & g & b & e & h & c & f \end{bmatrix}$ for the element 'a'.
Solution :

Minor of element 'a' is given as $M_{11} = \begin{bmatrix} e & h & f \end{bmatrix}$

Cofactor of a Matrix

The cofactor of a matrix is found by multiplying the minor of the matrix for a given element by $(-1)^{i+j}$. Cofactor of a Matrix is represented as C_{ij} . Hence, the relation between the minor and cofactor of a matrix is given as $C_{ij} = (-1)^{i+j} M_{ij}$. If we arrange all the cofactors obtained for an element, t, then we get a cofactor matrix given as $C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$

Adjoint of a Matrix

The adjoint is calculated for a square matrix. The adjoint of a matrix is the transpose of the cofactor of the matrix. The Adjoint of a Matrix is thus expressed as $\text{adj}(A) = C^T$, where C is the Cofactor Matrix.

Figure 7: Adjoint of a 2×2 matrix

Let's say, for example we have a matrix:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

then:

$$\text{adj}(A) = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad T \quad \text{adj}(A) = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}$$

where,

$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ is a cofactor of Matrix A.

Inverse of a Matrix

For a square matrix A of order n, its inverse A^{-1} can be defined as a matrix which, when multiplied by the original matrix, generates an identity matrix I of order n. i.e. $A \times A^{-1} = I$. The inverse is only calculated for a square matrix whose determinant is non-zero. The formula for the inverse of a matrix is given as:

$$A^{-1} = \text{adj}(A)/\det(A) = (1/|A|)(\text{Adj } A),$$

where $|A|$ should not be equal to zero, which means matrix A should be non-singular.

Elementary Operations on Matrices

Elementary Operations on Matrices are performed to solve the linear equation and to find the inverse of a matrix. Elementary operations are between rows and between columns. There are three types of elementary operations performed for rows and columns. These operations are mentioned below:

Elementary operations include:

- Interchanging two rows/columns
- Multiplying a row/column by a non-zero number
- Adding two rows/columns

Rank of a Matrix

The Rank of a Matrix is given by the maximum number of linearly independent rows or columns of a matrix. The rank of a matrix is always less than or equal to the total number of rows or columns present in a matrix. A square matrix has linearly independent rows or columns if the matrix is non-singular, i.e. determinant is not equal to zero. Since a zero matrix has no linearly independent rows or columns, its rank is zero. The rank of the matrix A is represented by $\rho(A)$.

Matrices Formulas

The basic matrices formulas has been discussed below:

- $A^{-1} = \text{adj}(A)/|A|$
- $A(\text{adj } A) = (\text{adj } A)A = I$, where I is an Identity Matrix
- $|\text{adj } A| = |A|^{n-1}$ where n is the order of matrix A
- $\text{adj}(\text{adj } A) = |A|^{n-2}A$ where n is the order of the matrix
- $|\text{adj}(\text{adj } A)| = |A|^{(n-1)^2}$
- $\text{adj}(AB) = (\text{adj } B)(\text{adj } A)$
- $\text{adj}(A^p) = (\text{adj } A)^p$
- $\text{adj}(kA) = k^{n-1}(\text{adj } A)$ where k is any real number
- $\text{adj}(I) = I$
- $\text{adj } 0 = 0$
- If A is symmetric, then $\text{adj}(A)$ is also symmetric
- If A is a diagonal Matrix, then $\text{adj}(A)$ is also a diagonal matrix
- If A is a triangular matrix, then $\text{adj}(A)$ is also a triangular matrix
- If A is a singular Matrix then $|\text{adj } A| = 0$
- $(AB)^{-1} = B^{-1}A^{-1}$

Why Matrices Matter in Data Science

- **Efficient Data Representation:** Tabular data (like datasets in CSV files or spreadsheets) can be easily stored as matrices.
- **Foundation for ML Models:** Algorithms like linear regression, neural networks and PCA use matrix operations.
- **Vectorized Computation:** Libraries like NumPy, TensorFlow and PyTorch use matrix operations to speed up calculations using hardware acceleration (CPU/GPU).
- **Multivariate Data:** Datasets with multiple features per observation are naturally represented as matrices.

Common Uses of Matrices in Data Science

1. Storing Datasets: Each row is an observation (e.g., a customer) and each column is a feature (e.g., age, income). For example:

2. Linear Algebra in Machine Learning:

Matrices are used for:

- Matrix multiplication in linear regression

- Gradient calculation in optimization
- Transformation and projection in dimensionality reduction (e.g., PCA)

3. Image Processing: Images are represented as matrices (grayscale) or tensors (color images with RGB channels), where each pixel is a value in the matrix.

4. Natural Language Processing (NLP): Matrices represent word embeddings or sentence vectors. For example, a word2vec model converts words into dense vectors and arranges them into a matrix.

5. Recommender Systems: A user-item matrix stores preferences, which can be used for collaborative filtering using matrix factorization.

Practice Problems Based on Introduction to Matrices

Question 1. Find the sum of the matrices $A=[3748]$ and $B=[1526]$

Question 2. Find the determinant of the matrix $A=[4231]$

Question 3. Find the product of the matrices $A=[1425]$ and $B=[2637]$

Question 4. Find the trace of the matrix $A=\begin{bmatrix} 2 & 1 & 7 & 4 & 3 & 8 & 6 & 5 & 9 \end{bmatrix}$

Question 5. Find the inverse of the matrix (if possible) $A=[1324]$

Question 6. Find the transpose of the matrix $A=\begin{bmatrix} 5 & 7 & 9 & 6 & 8 & 1 & 0 \end{bmatrix}$

Answer:

1. $[412614]$
2. -2
3. $[14381747]$
4. 14
5. $[-21.51-0.5]$
6. $[5678910]$

Use of MANOVA Test, Bayesian Statistics

In data science, we often deal with problems involving multiple variables or uncertain outcomes. MANOVA, Bayesian Statistics and Matrices are advanced statistical techniques used when traditional methods like a single t-test or basic probability are not enough. They provide better insights, especially in multivariate analysis or when incorporating prior knowledge into predictions.

1. MANOVA Test in Data Science

MANOVA (Multivariate Analysis of Variance) is used:

- To test differences across multiple outcomes at once
- Avoids increasing error rate due to running multiple ANOVAs
- Detects interaction between variables

Example: User Engagement Across Regions

Suppose you're testing if users from 3 regions differ in:

- Time spent on site
- Number of pages visited

Instead of running two separate ANOVA tests, you use MANOVA to test if both metrics differ together by region. If the result is significant, it means at least one group behaves differently across at least one metric.

Use cases:

- Evaluating effect of marketing campaigns on multiple KPIs
- Testing user behavior differences across segments
- Comparing model outputs across several performance measures

2. Bayesian Statistics in Data Science

Bayesian statistics updates probability estimates as new data becomes available. It helps:

- To incorporate domain knowledge or previous results
- Works well with small datasets
- Provides probability-based outputs instead of binary conclusions

Example: Spam Detection

Suppose you already believe 20% of emails are spam (prior belief). Then, based on the words in a new email, you calculate the likelihood it's spam. Bayesian methods combine both pieces to give a posterior probability e.g., a 90% chance the email is spam.

Use cases:

- Recommendation systems with prior user behavior
- Fraud detection with evolving patterns
- Updating model predictions as new data arrives

3. Matrices in Data Science

Matrices are foundational in data science for representing and manipulating structured data. They help in:

- Organize multivariate data for analysis
- Support fast computations using libraries like NumPy, PyTorch, etc
- Power core operations in machine learning, linear algebra and statistics

Example: Customer Dataset Representation

Suppose you have a dataset where each row represents a customer and each column a feature like age, income, purchase frequency, etc. This dataset is naturally stored as a matrix form and we can perform matrix operation on it for further analysis.

Use Cases:

- Storing datasets in tabular (2D) form
- Performing matrix multiplication in linear regression and neural networks
- Representing and transforming image data in computer vision
- Calculating eigenvalues/vectors for PCA and dimensionality reduction
- Encoding text data in NLP using word embeddings (as matrices or tensors)

Feature Selection using F-Anova

F-ANOVA (F-statistic-based Analysis of Variance) is a statistical method that finds how well a feature distinguishes between different classes by comparing variability between classes to the variability within each class. A higher F-statistic shows that the feature is effective in class separation while a lower value suggests it is less useful.

F-statistic formula is defined as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

where:

- **Variance between groups:** Average of squared differences between each class's mean and the overall mean.
- **Variance within groups:** Average of squared differences between data points in each class and their respective class means.

Steps to Calculate the F-statistic:

1. Calculate the mean of each class.
2. Calculate the overall mean of the target variable.
3. Compute the variance between classes.
4. Compute the variance within each class.
5. Calculate the F-statistic by taking the ratio of between-group variance to within-group variance.

How F-ANOVA is used for Feature Selection?

F-ANOVA helps in selecting the most important features for classification problems. The steps involved in feature selection using it are as follows:

1. **Calculate F-statistic for each feature:** For each feature in the dataset find the F-statistic with respect to the target variable. This will give a measure of how well the feature differentiates between the classes.
2. **Rank the features:** Sort the features based on their F-statistics with higher values showing more discriminative features.
3. **Select top features:** Choose the top features with the highest F-statistics which are considered the most important for the target variable.
4. **Model training:** Once the top features are selected they can be used to train machine learning models like Logistic Regression, Decision Trees, Support Vector Machines, etc. This reduces the dimensionality of the dataset and improves model performance.

Implementation of F-ANOVA for Feature Selection

In this example we will use the **f_classif** function from **sklearn.feature_selection** to find the F-statistic for each feature and select the top features based on the highest F-scores.

Step 1: Loading Iris Dataset from **sklearn.datasets**

We will be using Pandas and Scikit-learn libraries for its implementation.

- **train_test_split(X, y, test_size=0.3, random_state=42):** Splits the dataset X (features) and y (target) into training and testing sets with 30% of data allocated to testing and a random seed set for reproducibility.

```
import pandas as pd
from sklearn.datasets import load_iris
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

```
data = load_iris()
X = pd.DataFrame(data.data, columns=data.feature_names)
y = pd.Series(data.target)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Step 2: F-ANOVA Feature Selection

We use **SelectKBest** from **sklearn.feature_selection** with the scoring function **f_classif** which finds the F-statistic. Here we select the top 2 features (k=2) but we can modify k to select more or fewer features.

```
selector = SelectKBest(score_func=f_classif, k=2)
X_train_selected = selector.fit_transform(X_train, y_train)
X_test_selected = selector.transform(X_test)
```

```
selected_features = X.columns[selector.get_support()]
f_scores = selector.scores_[selector.get_support()]
print(f"Selected Features: {selected_features}")
print(f"F-Scores: {f_scores}")
```

Output:

```
Selected Features: Index(['petal length (cm)', 'petal width (cm)'], dtype='object')
F-Scores: [713.45534904 526.54162416]
```

Step 3: Model training and Evaluation

We train a Random Forest classifier using only the selected features.

```
model = RandomForestClassifier(random_state=42)
model.fit(X_train_selected, y_train)
```

```
y_pred = model.predict(X_test_selected)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy of the model with selected features: {accuracy:.4f}")
```

Output:

Accuracy of the model with selected features: 1.0000

We can also experiment it with other classifiers like SVM, Logistic Regression, etc for evaluation.

Limitations of F-ANOVA

While F-ANOVA is a useful tool for feature selection it has some limitations:

1. **Assumes Normality:** It assumes that the data within each class is normally distributed. If the data is not normally distributed, the F-statistic might not be reliable.
2. **Only for Classification:** It is generally used for classification problems and may not be suitable for regression tasks.
3. **Ignores Feature Interactions:** It evaluates each feature independently and may not capture complex interactions between features that contribute to the target variable.
4. **Sensitive to Outliers:** Like other statistical methods it can be sensitive to outliers which may distort the variance calculations.
5. **Multicollinearity:** It may not perform well in the presence of highly correlated features.

THANKYOU