IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

# Introduction to Statistical Data Analysis

**Diksha Bhola[1], Rachit Yadav[2], Vimmi Malhotra[3]**
Students, Department of Information Technology[1,2]
Professor, Department of Computer Science Engineering[3]
Dronacharya College of Engineering, Gurugram, India

**Abstract:** *This research paper provides comprehensive exposure to statistical analysis, elucidating its foundational concepts, modalities, and techniques. Beginning with the discussion on the importance of statistical analysis in making decisions and conclusions, the paper delves into the analysis measures which are Descriptive Analysis and Inferential Analysis. Furthermore, it includes how the data is graphically represented. Through this exposition, readers will gain a solid foundation of statistical data analysis..*

**Keywords:** Statistical Analysis, Data Analysis, Descriptive Analysis, Inferential Analysis

## I. INTRODUCTION

Indeed, the world is rapidly embracing the importance of data in almost every discipline. Learning data-related domains such as data analysis, statistics, and data science has become challenging for the generation. Though it involves complex mathematical and statistical concepts that may be difficult to grasp, the opportunities for career advancement and continuous growth have motivated individuals to learn about data.

In the sphere of statistics, comprehensive methodologies put into the practice to summarize and understand arithmetic information, with the aim of providing the ability to make decisions and predictions. According to J.W.Tukey, Statistics is not just a mathematical theorem, but rather an appropriate approach to systemize arithmetic. Statistics comprises of various modalities that are used to analyze data. It is pervasive across all realm of science including the gathering, sorting, analyzing, interpreting, and organizing data. Thus, Statistics can be conceptualized as the scientific practice dedicated to deriving knowledge from data and to quantify uncertainty.

Overall, the statistical and/or logical approach to manipulate, describe, abridge, and evaluate data defines the concept of data analysis. Mainly two approaches are used in data analysis that are- ***Descriptive Statistics*** and ***Inferential Statistics.***

To determine the appropriate statistical approach, one need to understand condition and assumption of the theory, aim of the study, type and nature of data and whether the observation is paired or unpaired. ***Descriptive Statistics*** is an approach that focuses on describing and summarizing the observable characteristics of a dataset. Aside from Descriptive Analysis, Inferential ***Statistics*** analyzes sampling and focus on making predictions and drawing conclusions using statistical tests.

## II. STATISTICAL ANALYSIS

Statistical analysis is a structural framework which provides important and deeper insights of arithmetic data and is a foundation for a successful data analysis.This paper presents the various approaches and measures which are used to analyze and implement data.

There are two important components of a statistical study, that are- ***Population*** and ***Sample***

Population is an assemblage of all the individuals who have certain characteristics and are of interest in a study, though it is quite impossible or difficult to examine every individual of an entire population.

Let's consider an example to understand. It's not practically possible to examine all customers who have purchased products from a particular company. This includes every individual who has bought something from that company, however deciding to examine randomly selected 450 customers about their preferences is possible. These 450 customers represent Sample. This subset of entire population is called Sample.


DOI: 10.48175/IJARSCT-17830

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

## III. DESCRIPTIVE ANALYSIS

Descriptive Analysis is an essential step in data analysis. It is a statistical analysis measure used to characterize and summarize the aspect of a dataset. Descriptive Analysis provides insights of samples and the observation conducted in an efficient and straightforward way using metrics such as mean, median, variance, range, graphs and charts. It simplifies large data into sensible and simpler summary. Descriptive Analysis is helpful for appprehending data patterns and linkages.

Descriptive Analysis typically involves measures of central tendency, dispersion, and distribution shapes. Additionally, graphical representations through pie charts, graphs, box plots, and scatter plots are some common graphical techniques to visualize data in descriptive statistics.

### 3.1 Measure of Central Tendency

The central tendency is defined as a statistical measure which is used to analyse and summarize complete dataset or distribution with a single typical value. This statistical measure is sometimes known as measure of central location. Sometimes, these are classed as summary statistics. These measures represent the whole distribution of data.

Following sections represent all valid central tendency measures, their working and formulae and kinds in depth.

**What is Mean?**

Mean or simply average is the most popular measure of central tendency which can be used with both discrete and continuous data. The Mean is the assemblage of all the components in a group/collection divided by the number of components present in that group or collection.

Mean of a data collection is generally denoted as $\bar{x}$ which is pronounced as "X Bar".

$$\text{Mean, } \bar{x} = \Sigma x_i / n$$

Example: The given table shows the marks obtained by different students in a class. What is the mean of the given data?

| Name of Student | Alen | Ben | Carie | Daisy | Frank | Jack | Nick |
|---|---|---|---|---|---|---|---|
| Marks Obtained | 80 | 52 | 40 | 70 | 76 | 65 | 84 |

The mean is given by:

$\bar{x}$ = (80 + 52+ 40 + 70 + 76 + 65 + 84) / 7

= 467 / 7 = 66.71

So, the mean of the given data is 66.71

**What is Median?**

Median of a dataset is the middle-most observation or the arithmetic average of two middle values after arranging the data in ascending order, which is one of the measures of central tendency. Median may be used to calculate the median of grouped data as well as ungrouped data.

**Median of Ungrouped Data (n is odd): $[(n + 1)/2]^{th}$ term**

**Median of Ungrouped Data (n is even): $[(n / 2)^{th}$ term + $((n / 2) + 1)^{th}$ term$]/2$**

Example: The given table shows the marks obtained by different students in a class. What is the median of the given data?

| Name of Student | Alen | Ben | Carie | Daisy | Frank | Jack | Nick |
|---|---|---|---|---|---|---|---|
| Marks Obtained | 80 | 52 | 40 | 70 | 76 | 65 | 84 |

The median is given by:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

Arrange the given data in ascending order: 40, 52, 65, 70, 76, 80, 84
Median = $[(n + 1) / 2]^{th}$ term
$= [(7 + 1) / 2]^{th}$ term $= 4^{th}$ term
$= 70$
Thus, median for the provided data collection is 70.

**What is Mode?**
Mode is the most common value of the group or collection which states the value that appears the most frequently in the given data, i.e. the observation with the highest frequency is considered as the mode of the dataset.A group in which each data value occursthe same number of times has no mode.
Example: The given table shows the marks obtained by different students in a class. What is the mode of the given data?

| Name of Student | Alen | Ben | Carie | Daisy | Frank | Jack | Nick |
|---|---|---|---|---|---|---|---|
| Marks Obtained | 76 | 52 | 40 | 70 | 76 | 65 | 84 |

Mode = Most repeated observation in the dataset
= 76
Thus, the mode of the provided dataset is 76.

**3.2 Measures of Dispersion**
The Measures of Dispersion are foundational principles of Descriptive Statistics and are used to quantify the spread or variability of a dataset. These measures complement the measures of central tendency of the data describing the extent to which the data points deviate from the average or in simple words, they provide insights into how the independent data points are dispersed around the central tendency, such as mean or median.
Understanding these measures is important for analysis because they help to evaluate the consistency within the dataset, which consecutively helps in drawing precise conclusions or making decisions.
Following sections represents the variability measures in detail.

**What is Range?**
The range can be defined as the simplest measure which calculates the difference between the maximum value and the minimum value in the dataset.
Range = maximum value – minimum value
The range must have the exact units as those of the data values from which it is calculated.
In the example given above shows,
Minimum value = 40
Maximum value = 84
Thus, the range can be evaluated as 84 – 40 = 44.

**What is Variance and Standard Deviation?**
Variance and Standard Deviation provide a measure of variability of a variable that is thoroughly used far and wide. Variance can be defined as an average of squared differences from the mean whereas Standard Deviation is calculated using variance i.e. it a square root of the variance. Variance measures an average degree to which each value point is different from the mean whereas Standard Deviation indicates the expansion between numbers in a data set.
Variance can be represented with the symbol **$\sigma^2$and Standard Deviation can be represented with the symbol σ.**
Variance, $\sigma^2 = \Sigma (xi - \bar{x})^2/n$

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

Example: Find the variance and standard deviation of all the odd numbers less than 10.
Odd Numbers less than 10 are {1, 3, 5, 7, 9}
This data set has five values (n) = 5
Before finding the variance, we need to find the mean of the data set.
Mean, $\bar{x} = (1 + 3 + 5 + 7 + 9)/5 = 5$
We will put the value of data and mean in the formula,
$\boldsymbol{\sigma^2 = [(1-5)^2 + (3-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2]}$
$= (16 + 4 + 0 + 4 + 16)/5 = 40/5$
Variance, $\boldsymbol{\sigma^2 = 8}$
Now, Standard Deviation, $\boldsymbol{\sigma} = \sqrt{(\sigma^2)} = \sqrt{(8)} = 2.828$

**What is Percentile and Quartile?**
Percentiles and quartiles are statistical measures that help to understand the distribution and spread of data in a dataset. They divide the data into equal parts, allowing for comparisons and insights into the relative positions of individual data points. Here's an explanation of percentiles and quartiles along with examples:

**Percentiles** divide a dataset into hundred equal parts, each representing a percentage of the total observations.For instance, the 50th percentile (also known as the median) divides the data into two equal halves, with 50% of the observations falling below it and 50% above it.The 25th percentile corresponds to the lower quartile (Q1) and marks the point below which 25% of the observations fall.Similarly, the 75th percentile corresponds to the upper quartile (Q3) and marks the point below which 75% of the observations fall.
Example: Let's say you're conducting a study on household incomes in a city. You collect data from 100 households and calculate the 25th, 50th, and 75th percentiles of their incomes. This provides insights into the income distribution, such as how many households earn below a certain threshold and where the median income lies.

**Quartiles** divide a dataset into four equal parts, each containing a quarter of the observations.The first quartile (Q1) is the same as the 25th percentile, indicating the value below which 25% of the observations lie.The second quartile (Q2) is the same as the 50th percentile (the median), dividing the data into two halves.The third quartile (Q3) is the same as the 75th percentile, indicating the value below which 75% of the observations lie.

Example: Consider a study on student performance in a class. You gather test scores from 50 students and calculate the quartiles of their scores. This allows you to identify the range within which most students scored, the performance level of the middle 50% of students, and the spread of scores in the upper quartile.

Percentiles and Quartiles provides valuable insights into the distribution of your data, helping readers understand the central tendency and variability more comprehensively. Additionally, they offer a robust way to compare datasets and assess the relative positions of observations within them.

**3.3 Measures of Distribution Shapes**
In statistics, measures of distribution shape help characterize the form or pattern of a distribution of a dataset. Understanding the shape of a distribution is crucial for making inferences, selecting appropriate statistical methods, and interpreting results accurately. Here are some common measures of distribution shape:
In statistics, measures of distribution shape help characterize the form or pattern of a dataset's distribution. Understanding the shape of a distribution is crucial for making inferences, selecting appropriate statistical methods, and interpreting results accurately. Here are some common measures of distribution shape:

**What is Skewness**?
Skewness measures the asymmetry of a distribution. A distribution is considered:
**Symmetric,** if the skewness is close to zero.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

**Positively skewed (right-skewed),** if the tail on the right side of the distribution is longer or fatter than the left side. This means that the majority of the data points are concentrated on the left side, with a few extreme values on the right side.

**Negatively skewed (left-skewed),** if the tail on the left side of the distribution is longer or fatter than the right side. This indicates that most of the data points are concentrated on the right side, with a few extreme values on the left side.

**What is Kurtosis**?

Kurtosis measures the flatness of a central peak of a distribution. A distribution is considered:

**Mesokurtic**, if it has a similar shape to a normal distribution, with a kurtosis value close to zero.

**Leptokurtic**, if it has a sharper central peak and heavier tails compared to a normal distribution. This means that it has more extreme values than a normal distribution, resulting in a higher kurtosis value.

**Platykurtic,** if it has a flatter central peak and lighter tails compared to a normal distribution. This indicates fewer extreme values than a normal distribution, resulting in a lower kurtosis value.

**What is Modality?**

Modality refers to the number of peaks or modes in a distribution.A distribution can be:

**Unimodal**, if it has one clear peak.

**Bimodal**, if it has two distinct peaks.

**Multimodal**, if it has more than two peaks.

**What is Tail?**

The tails of a distribution refer to the ends of the distribution where the data become less dense.

**Heavy-tailed** distributions have tails that decline more slowly than those of a normal distribution.

**Light-tailed** distributions have tails that decline more rapidly than those of a normal distribution.

Understanding these measures helps to describe, compare, and analyse different datasets effectively. They provide valuable insights into the underlying characteristics of the data, allowingto make informed decisions about the appropriate statistical methods to use and the validity of their conclusions.

### 3.4 Graphical Representation

Graphical representation is a visual tool used in statistics to present data in a clear, concise, and understandable manner. Graphs help decision-makers interpret data, identify patterns, trends, and relationships, and communicate findings effectively. Certainly! Let's delve into more detail about each type of graphical representation:

**What are Histograms**?

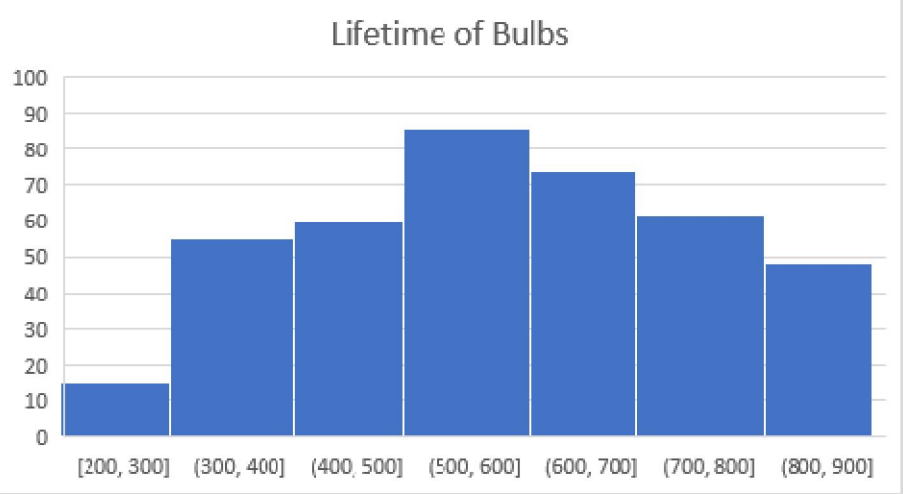
Histograms offer a visual summary of the distribution of continuous data by dividing it into intervals (bins) along the x-axis and displaying the frequency or relative frequency of observations within each interval through the height of the bars on the y-axis.They are particularly useful for revealing the shape, centre, spread, and any outliers present in the data distribution.Histograms are dynamic tools that can be adjusted by changing the number of bins, revealing different levels of detail in the data distribution.

Example: The following table gives the lifetime of 400 Lamps. Draw the histogram for the data below:

| Lifetime (in hours) | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 | 800-900 |
|---|---|---|---|---|---|---|---|
| No. of Lamps | 14 | 56 | 60 | 86 | 74 | 62 | 48 |

The histogram for the given data can be represented as:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024



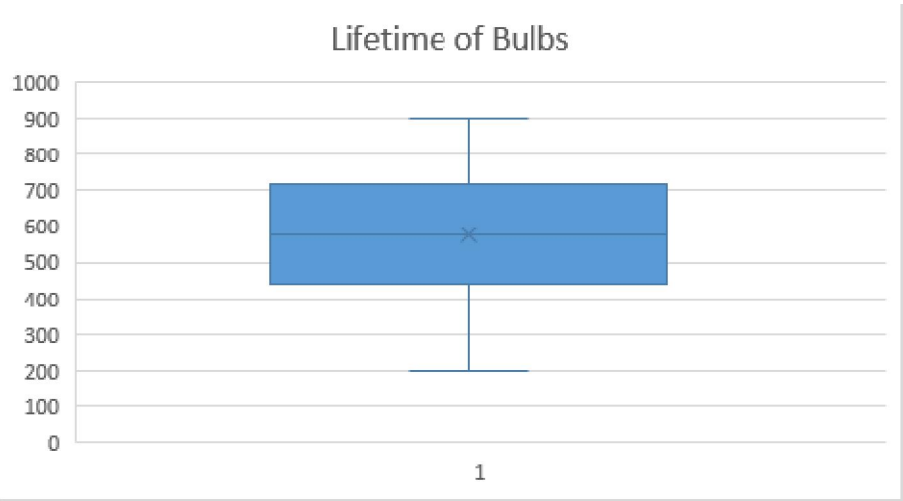

**What are Box Plots (Box-and-Whisker Plots)?**

Box plots provide a concise visual representation of the distribution of a dataset, displaying the median (line within the box), quartiles (edges of the box), and range or variability (whiskers).They are highly effective for comparing distributions between different groups or categories, identifying potential outliers, and gaining insights into the spread and central tendency of the data.Box plots can handle skewed or asymmetric data distributions, making them robust tools for exploratory data analysis.

Example: The following table gives the lifetime of 400 Lamps. Draw the Box Plot for the data below:

| Lifetime (in hours) | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 | 800-900 |
|---|---|---|---|---|---|---|---|
| No. of Lamps | 14 | 56 | 60 | 86 | 74 | 62 | 48 |

The Box Plot for the given data can be represented as:




**What are Scatter Plots?**

Scatter plots illustrate the relationship between two continuous variables by plotting individual data points on a Cartesian plane, with one variable on the x-axis and the other on the y-axis.They allow researchers to visualize patterns such as linear or nonlinear relationships, clusters, or trends in the data, providing insights into the strength

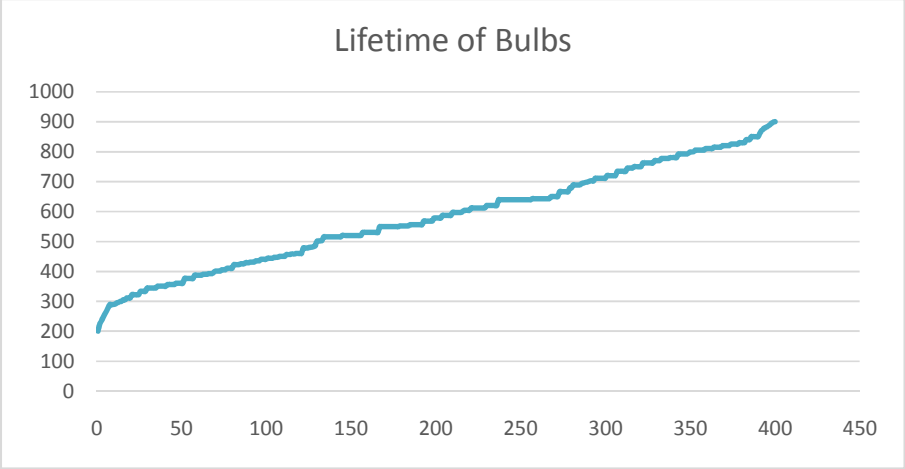
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

and direction of the relationship between the variables.Scatter plots are versatile tools that can be enhanced with additional features such as colour, size, or shape to represent additional variables or categories, enabling more complex data exploration.

Example: The following table gives the lifetime of 400 Lamps. Draw the Scatter Plot for the data below:

| Lifetime (in hours) | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 | 800-900 |
|---|---|---|---|---|---|---|---|
| No. of Lamps | 14 | 56 | 60 | 86 | 74 | 62 | 48 |

The Scatter Plot for the given data can be represented as:




**What are Bar Charts?**

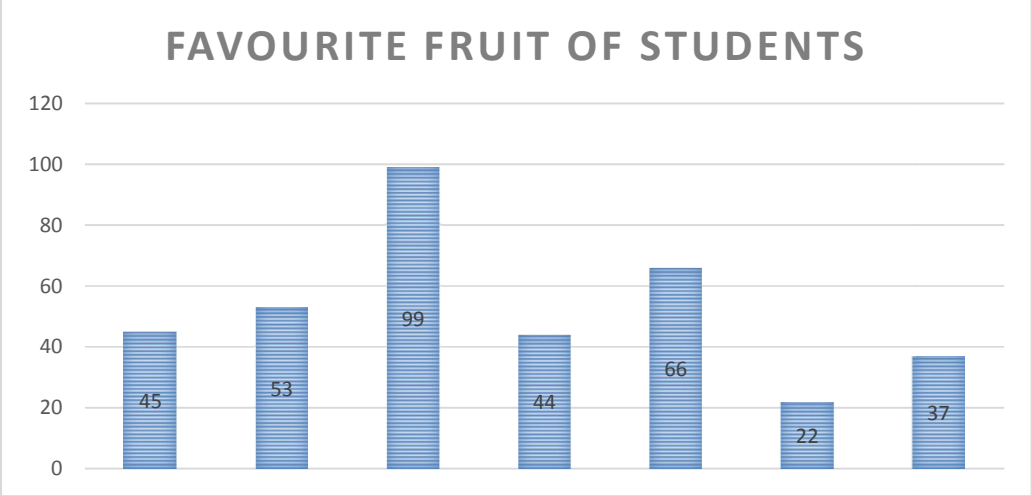
Bar charts are effective for comparing categorical data or discrete values by representing categories on the x-axis and the frequency, proportion, or count of data within each category through the height of the bars on the y-axis.They enable straightforward comparisons between different categories or groups, highlighting variations and trends within the data.Bar charts are commonly used in market research, social sciences, and business analytics to visualize survey results, market shares, and demographic distributions.

Example: A school conducted a survey to know the favourite fruit of the students. The table below shows the results of this survey. Draw a Bar Chart representing the given data.

| Name of the Fruit | Total Number of Students |
|---|---|
| Cherry | 45 |
| Strawberry | 53 |
| Mango | 99 |
| Banana | 44 |
| Apple | 66 |
| Kiwi | 22 |
| Pineapple | 37 |

The bar chart below depicts the following data:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024



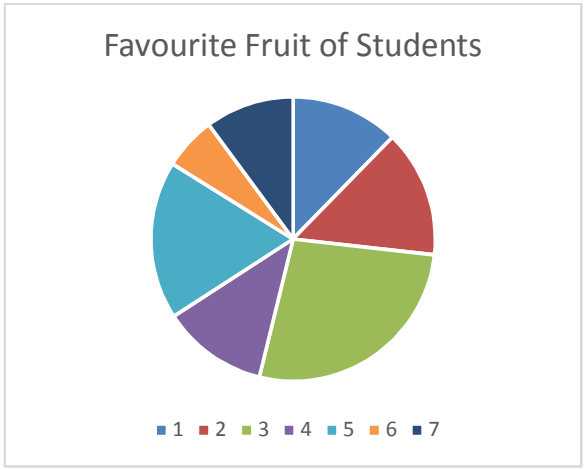

**What are Pie Charts?**

Pie charts depict the relative proportions or percentages of categorical data by dividing a circular chart into slices, with each slice representing a category and its size proportional to its share of the whole.They provide a clear visual representation of the distribution of data categories, making it easy to identify the most significant contributors and their relative importance. Pie charts are commonly used in presentations, reports, and dashboards to convey information about market segmentation, budget allocation, and survey responses.

Example: A school conducted a survey to know the favourite fruit of the students. The table below shows the results of this survey.Draw a Pie Chart representing the fruits and the total number of students.

| Name of the Fruit | Total Number of Students |
|---|---|
| Cherry | 45 |
| Strawberry | 53 |
| Mango | 99 |
| Banana | 44 |
| Apple | 66 |
| Kiwi | 22 |
| Pineapple | 37 |

The pie chart below depicts the following data:



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024

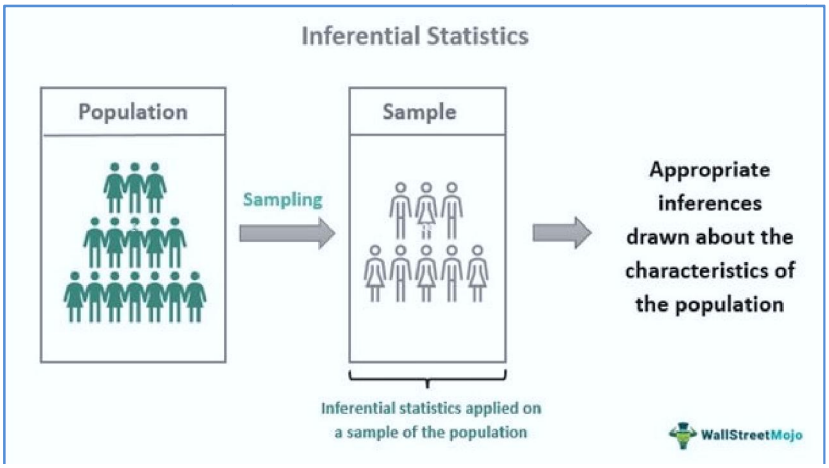
Each type of graphical representation offers unique advantages and insights, making them valuable tools for data exploration, analysis, and communication in variousfields and applications. By selecting the most appropriate graph type based on the data characteristics and objectives, one individual can effectively convey complex information and facilitate informed decision-making.

## IV. INFERENTIAL ANALYSIS

Inferential Statistical Analysis is the statistical analysis measure that involves using sample data to make inferences or draw conclusions about a larger population. There are many types of inferential statistics and each is appropriate for specific research design and sample characteristics. However, most inferential statistics are based on the principle that a test-statistic value is calculated on the basis of particular formula.

Inferential statistical analysis is widely used in various fields such as science, social sciences, business, and healthcare to make informed decisions, test hypotheses, and answer research questions. It allows individuals to generalize their findings beyond the sample studied and make predictions about the population of interest.

That value along with the degrees of freedom, a measure related to sample size, and rejection criteria are used to determine whether differences exist between the treatment group. Samples are used to generate the data, and inferential statistics are used to generalize that information to the population, a process in which error is inherent.




If we take samples from repeatedly from the same population, it is likely, when all the means are put together, their distribution will resemble the normal curve. The distribution referred to in this context is known as the sampling distribution, and its corresponding standard deviation is termed the standard error.

### 4.1 Analysis of Variance (ANOVA)

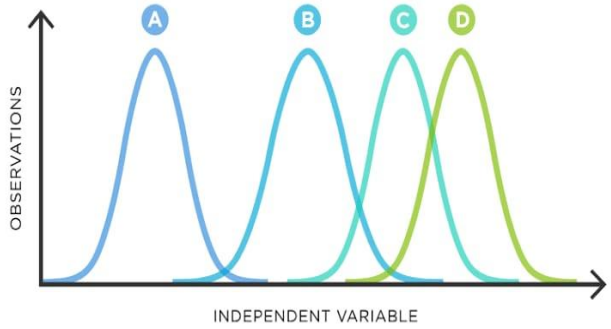
The analysis of variance, commonly known as ANOVA, is a statistical test utilized to ascertain if there exists a significant difference in the value of a single dependent variable across three or more levels of an independent variable.

The one-way ANOVA extends the comparison capability of the independent two-sample t-test. For instance, it can be applied to analyse the mean distances covered by a four-kilogram projectile across three tension levels of a catapult. Here, the independent variable consists of the different tension settings of the catapult, while the single dependent variable is the measurement of the projectile's travel distance.

The core concept behind ANOVA is that when the ratio of variance between groups to variance within groups is higher, it indicates a higher probability that the observed differences among the groups are genuine.

The objective of ANOVA is to reject the null hypothesis, which suggests no significant difference exists between the examined groups, and to accept the alternative hypothesis, which posits that the differences observed among the groups are genuine.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, April 2024




### 4.2 Statistically Significant

Statistical significance entails assessing whether the relationship between two or more variables is attributable to factors beyond random chance. It serves to furnish evidence regarding the validity of the null hypothesis, which suggests that the data is solely influenced by random chance.

The methodology for assessing the statistical significance of outcomes emerged in the early 20th century. It's crucial to note that "significance" in this context doesn't connote importance, and "statistical significance" shouldn't be confused with research significance, theoretical significance, or practical significance. For instance, "clinical significance" pertains to the practical relevance of a treatment's effect.

Mentioning its most common benefits, we can't overlook the importance of statistical significance, which includes:

- Enhanced confidence in decision-making by relying on data to make tough decisions.
- Minimizing the risk of false positives or false negatives clouding the results of A/B tests.
- Optimizing resource allocation for process improvement, leading to more significant outcomes.
- Improving reporting and communication efficiency, underpinned by robust research and experimentation, ensuring initiatives are well-supported.

Finally, one must always use measures of association along with tests for statistical significance.

## V. CONCLUSION

In the rapidly evolving digital landscape of 2024, data analytics has emerged as a critical discipline for businesses across industries. The IT industry commands the largest share (43%) of the data analytics market, with significant engagement observed among various companies in this sector. Following closely is the e-commerce sector, which accounted for an estimated US$ 50 billion in 2020 and is responsible for facilitating 1.2 million transactions daily, according to NASSCOM. By 2023, this sector is projected to surpass the US, becoming the second-largest retail market.

## REFERENCES

**[1].** McCune, S. (2009).Practice makes perfect statistics. McGraw Hill Professional.
**[2].** https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html#:~:text=Data%20Analysis%20is%20the%20process,and%20recap%2C%20and%20evaluate%20data
**[3].** Tukey, J.W.: Exploratory Data Analysis. Pearson, London (1977).
**[4].** Rumsey, D. J. (2010). Statistics essentials for dummies. John Wiley & Sons.
**[5].** RuiPortocarreroSarmento, Vera Costa (2019). An Overview of Statistical Analysis
**[6].** Claus Weihs, KatjaIckstadt (2018). Data Science: The impact of statistics
**[7].** Selection of Appropriate Statistical Methods for Data Analysis - PMC (nih.gov)
**[8].** Peers, I. (2006).Statistical analysis for education and psychology researchers: Tools for researchers in education and psychology. Routledge.