International Journal of Innovation Scientific Research and Review
Vol. 07, Issue, 01, pp.7613-7621, January 2025
Available online at http://www.journalijisr.com
SJIF Impact Factor 2023: 6.599

Research Article

ISSN: 2582-6131

MASTERING STATISTICS: A JOURNEY FROM DATA SCIENCE TO DOCTORAL EXCELLENCE

[1,] * Muhammed Sameer Uddin, [2]Omaima Eltahir Babikir Mohamed, [3]Ziarat H. Khan, [4]John Ebert

[1]School of Business and Technology, Saint Mary's University of Minnesota, Winona, MN 55987, USA.
[2]Bank Rakyat School of Business & Entrepreneurship, Universiti Tun Abdul Razak (UNIRAZK), Kuala Lumpur, Malaysia.
[3]Faculty of Business Administration, American International University-Bangladesh (AIUB).
[4]Department of Resource Analysis, Saint Mary's University of Minnesota, Winona, MN 55987, USA.

Received 13th November 2024; Accepted 14th December 2024; Published online 31st January 2025

ABSTRACT

Statistics is an important tool in data science, business analytics, and doctoral research because it helps people find useful patterns in data. This paper looks at the basic ideas of statistics, like descriptive and inferential methods, regression analysis, multivariate analysis, and time series analysis, with a focus on how they can be used in real life. It also discusses the problems researchers face like how to learn difficult methods and combine knowledge from different fields. This study shows how powerful statistics can be in solving real-world problems and improving academic and professional performance by handling these problems and stressing the importance of cleaning and exploring data.

Keywords: Statistics, Data Science, Analytics, Research.

INTRODUCTION

The business is today, and twenty-five years or more back, were different. Organizations had to make decisions based on their experiences, intuitions, and qualitative data. Moreover, business decisions were guided by personal judgments, customers' feedback, historical data, or limited surveys based on case studies. The trial-and-error approaches were common as tools and data sets were limited for decision-making purposes. Statistics is a fundamental branch of mathematics. It simply states collecting, organizing, analyzing, and interpreting data for decision purposes. Statistics, nowadays, plays a critical role in everything from social science to rocket science. The metaphorical expression, "Data is GOD," describes the immense power and influence in today's technological world. So, questions arise about big-data metadata and what to use it for regarding decision criteria. Here, statistics play the ground-breaking role of shading the light through data.

This paper explores how statistics is an important part of giving students, researchers, and professionals the information and skills they need to do well in fields like data science, business analytics, and doctoral-level research. This journey aims to show "what" statistics can be used to solve problems in the real world and improve classroom performance by breaking down important statistical ideas and showing how they can be used.

LITERATURE REVIEW

This section focuses on setting the context and providing a conceptual foundation, drawing upon the Resource-Based View (RBV) and Strategic Choice Theory (SCT). It introduces the why and what of this paper (Uddin *et al.,* 2023a):

***Corresponding Author: Muhammed Sameer Uddin,**
School of Business and Technology, Saint Mary's University of Minnesota, Winona, MN 55987, USA.

A. Statistics in Data Science: Importance and Applications

Statistics are an important part of both data science and business analytics because they help us understand and make decisions based on data. Predictions and inferences can be made with tools like hypothesis testing, regression analysis, and probability distributions (Gastelum *et al*., 2023; Gupta & Tawar, 2020). Statistical ideas are often used in machine learning methods, and visualizing data helps people understand patterns and trends (Gupta & Tawar, 2020). It can be used in many fields to predict customer behavior, handle risks, boost quality, advance healthcare, make education better, and solve environmental problems (Balakrishnan, 2010; Ellis & Slade, 2023; Hsieh, 2023; Martha, 2024; Sarraf, 2023).

B. Challenges for Doctoral Researchers in Mastering Statistics

Doctoral researchers face significant challenges in mastering statistics. The complexity of statistical methods often leads to misinterpretation of concepts like p-values, even among well-trained individuals (Lytsy *et al*., 2022). Integrating statistical techniques with domain-specific knowledge poses another hurdle, requiring advanced methods like machine learning for complex phenomena (Hofmann *et al*., 2022; Penchev, 2021). Access to tools like statistical software and training is limited, which makes learning even harder. Time constraints make the problem even worse (Patil & Satagopan, 2022; Salloum *et al*., 2016). Taking these problems on is very important for improving the quality of study and people's ability to use statistics.

C. Integration of Statistical Knowledge Across Fields

Statistics make it easier for people from different fields to work together and solve problems. It makes algorithms like regression and probability models work in machine learning. It also makes applications like picture recognition and natural language processing possible (Priya *et al*., 2021). Liu (2023) says that econometrics uses statistical methods to predict trends and look at policies, which helps people make better economic decisions. For example, Arıkan *et al*.,

(2023) say that statistics can help improve logistics, job safety, and the use of resources. This cross-disciplinary collaboration shows how important statistics are for fixing complex problems and sparking new ideas across all fields.

THEORETICAL BACKGROUND

A. Introduction to Statistics

What is Statistics

Statistics is a branch of mathematics that involves collecting, organizing, analyzing, and interpreting data for decision purposes (Robert & John, 2022). We can extract information using data through statistics. For example:

A. Imagine you are the owner of a bookstore. Over a week, you record the daily customer numbers: 50, 60, 55, 65, 70, 80, 75. Now, you want to find out the average number of customers visiting each day to plan staffing.
B. Imagine you are managing a small online grocery store. You track the daily sales revenue for a week: $500, $600, $550, $650, $700, $800, $750. Now, you want to analyze the sales trend to determine which day of the week generates the highest revenue.
C. Imagine you are a school teacher. You collect the test scores of your students for a recent math quiz: 85, 90, 78, 92, 88, 95, 87. You want to evaluate the overall performance by identifying the average score and the score distribution in the class.

B. Types of Statistics

There are two main types of statistics:

1. Descriptive
2. Inferential

Descriptive – It defines as collecting, analyzing, and interpreting of data. The main purpose of descriptive statistics is to understand the data. Descriptive statistics can all so provide graphs, charts, tables etc. to describe the data sets (Robert & John, 2022; Thomas, 2021).

Inferential–Is all about drawing conclusions from the data. It uses methods like confidence intervals, estimation, and hypothesis testing to infer insights about the population (Robert & John, 2022; Thomas, 2021).

C. Types of data

In statistics, there are mainly two types of data,

1. Categorical
2. Numerical (Quantitative)

Categorical– the data that describes qualities or characteristics of an object and cannot be measured numerically(Robert & John, 2022; Thomas, 2021). Categorical data can be divided into two groups:

- **Nominal Data**- is labels or names without a specific order(Thomas, 2021). For example: Gender (Male/Female), Colors (Red, Blue), Nationality (Country), Blood-type (A, B), etc.
- **Ordinal Data** – is a categorical or qualitative data that groups variables or data into ordered categories or in other words, this is a dataset that has natura order or rank(Thomas, 2021). For example, Rankings (1st, 2nd, 3rd), Satisfaction Levels (Happy, Neutral, Sad); Olympic medals(Gold, silver, and bronze)

Numerical (Quantitative) – the data that represents quantities or numbers and can be measured or counted (Robert & John, 2022; Thomas, 2021). Numerical data can be grouped into four groups:

- **Discrete Data** – is data that consists of whole numbers or counts(Thomas, 2021). It cannot have decimals or fractions. For example: Number of students in a class (20, 25), Tickets sold (100, 150), Number of pets (1, 2, 3).
- **Continuous Data** – is data that can take any value within a range, including decimals or fractions(Thomas, 2021). For example: Height of a person (5.5 feet, 6.2 feet), Weight of an apple (1.5 kg, 2.3 kg), Temperature (37.5°C, 100.2°C).
- **Interval data** - is characterized by the presence of meaningful intervals between values, but it lacks a true zero point. This means that while differences between values can be quantified, ratios cannot be meaningfully calculated (Bonett & Price, 2020). For example, temperature measured in Celsius is interval data; while one can say that 30°C is 10 degrees warmer than 20°C, one cannot say that 30°C is "1.5 times as hot" as 20°C.
- **Ratio data** - possesses all the properties of interval data, but it also includes a true zero point, allowing for both differences and ratios to be calculated (Bonett & Price, 2020). An example of ratio data is height or weight, where a measurement of zero indicates the absence of the quantity being measured.

Apart from conventional data, like numerical and categorical, data can also be categorized depending on its **structure, time, and dimensions**. **Unstructured** data is text, photos, or videos without a specified format; **structured** data is data kept in rows and columns. **Cross-sectional** data—that is, that is, a snapshot at a single point in time—or time series, which illustrates changes over a period —can also be categorized. Data can also **multivariate**, involving several variables—or **univariate**, meaning involving a single variable(Robert & John, 2022; Thomas, 2021). These categories enable efficient data organization and analysis in many different settings.

D. Population and Sample

A **population** is defined as an entire group of individuals, objects, elements, anything that we want to study. In other words, it represents the entire set of datasets that come under the research or study(Moore & Notz, 2021; Thomas, 2021). For example, the population of the USA, all the students at the University, all the employees of the company, etc.

Sample, on the other hand, is the sub-set of the population. Sample is used when the entire population is impractical in the research or study (Moore & Notz, 2021; Thomas, 2021). For example, 100 students are selected from the University, 100 employees are selected from the company, population from Mankato, Minnesota, USA, etc. Sampling techniques are essential for data collection in research or analysis, and Table 1 provides a detailed summary of the key

methods along with their definitions and examples.

1. Probabilistic, and
2. Non-Probabilistic

Probabilistic Sampling: Methods of sampling in which each person in the group has a known, non-zero chance of being chosen(Moore & Notz, 2021; Thomas, 2021).

Non-Probabilistic Sampling: Sampling techniques where the selection is based on convenience, judgment, or quotas, and not every member has a chance of being selected(Moore & Notz, 2021; Thomas, 2021).

Table 1: Sampling Techniques

Type of Samplig	Definition	Example
Simple Random Sampling	Every member of the population has an equal chance of being selected.	Randomly selecting 100 students from a school of 1,000 students.
Systematic Sampling	Selecting every member of the population at a regular interval.	Surveying every 10th person on a list of 500 employees.
Stratified Sampling	Dividing the population into subgroups (strata) and sampling proportionally from each.	Selecting 50 students from each grade level in a high school.
Cluster Sampling	Dividing the population into clusters and randomly selecting entire clusters for the sample.	Surveying all households in 5 randomly selected neighborhoods.
Conveniene Sampling	Selecting individuals who are easiest to reach or access.	Surveying customers entering a specific store on a given day.
Quota Sampling	Selecting a sample that reflects specific characteristics or quotas of the population.	Ensuring 60% of the sample are women and 40% are men in a study.
Purposive Sampling	Choosing participants based on the purpose of the study and their relevance to the research.	Interviewing only experts in supply chain management for a study.
Snowball Sampling	Existing participants recruit future participants from their network.	Using initial survey respondents to identify others in hidden populations (e.g., freelancers).

APPROACH OFSTATISTICAL DATA ANALYSIS AND STEPS

A. Qualitative and Quantitative Approaches:

There are two approaches that can be used in statistical data analysis-

1. Qualitative
2. Quantitative

Qualitative - is a data analysis approach that uses non-numerical data like text, images, or videos to understand concepts, themes, or experiences(Moore & Notz, 2021; Thomas, 2021). For example, analyzing customer satisfaction interviews to understand customers' opinion. Thematic, content, and narrative analysis are the common techniques of qualitative data analysis approach.

Quantitative– is a data analysis approach that uses numerical data to analyze patterns, relationships or trends(Moore & Notz, 2021; Thomas, 2021). For example, analyzing customer satisfaction surveys on a scale of 1 to 5. Descriptive and inferential are common techniques for quantitative data analysis methods. These will be discussed in detail in the next section.

B. Steps of Data Analysis:

The steps in statistical data analysis are given the following table 3.

Table 2: Steps of Data Analysis

Step	Description	Example
Define the Problem	Identify the question or problem	What factors affect customer satisfaction?
Collect the Data	Collect information from surveys, experiments, or other sources.	Conduct a survey to measure satisfaction scores.
Clean and Prepare Data	Handle missing values, correct errors, and organize data for analysis.	Remove duplicates, replace missing values.
Explore and Summarize Data	Use descriptive statistics and visualizations to understand data patterns.	Plot satisfaction scores using a histogram.
Apply Statistical Methods	Perform analysis using techniques like hypothesis testing, regression, etc.	Test satisfaction differences by age using ANOVA.
Interpret Results	Convert statistical results into practical insights.	Younger customers have higher satisfaction.
Present Findings	Communicate results with graphs, charts, and clear summaries.	Share a report highlighting trends and solutions.

DESCRIPTIVE STATISTICS

Descriptive statistics are a simple way to summarize and describe the most essential parts of a dataset. They give numbers and pictures to show the data, which helps us understand patterns and trends more clearly. The median (middle value), mode (most common value), and standard deviation (spread of data) are all common ways to measure things. Descriptive statistics include graphs and charts like histograms, bar charts, and box plots (Moore & Notz, 2021; Peter & Andrew, 2017; Robert & John, 2022; Thomas, 2021; Uddin et al., 2024; Weihs & Ickstadt, 2018).

A. Measures of Central Tendency

Measures of central tendency are used to describe the center or typical value of a dataset. They summarize the dataset with a single value, representing the "average" or most common characteristic of the data (Peter & Andrew, 2017). The three main measures are:

1. Mean,
2. Median, and
3. Mode.

Mean is simply the average of a data set. This is the common measure of central tendency. This can be calculated by adding all the values in a dataset together and then dividing by the number of the values or observations (Moore & Notz, 2021).

Median is the number that is right in the middle, or at the halfway point of the set of values. Putting the numbers in increasing order and finding the middle number is one way to find the median(Moore & Notz, 2021). If there are an even number of values, we need to find the median by taking the mean of the two values in the middle.

Mode is the number that shows up most often in the data distribution. The data is put in order from lowest to highest, and then the numbers are counted to find the mode(Moore & Notz, 2021). As the name suggests, mode is the figure that shows up most often.

B. Measures of Spread

The level of difference in the data is called dispersion. Dispersion, which is also called variability or spread, measures how far apart data points in a set are from the mean, median, or mode, which is the central trend (Patten & Newhart, 2023; Thomas, 2021; Weihs &

Ickstadt, 2018). They tell us a lot about how the data is spread out, scattered, or distributed. When it comes to Tech Classes, two datasets with the same variable may have similar values of the center but be very different when it comes to variability. The four key measures are:

1. Variance,
2. Standard deviation,
3. Range, and
4. Interquartile range.

Variance is the average differences of the squared from the mean. It indicates how spread out the data points are around the mean(Moore & Notz, 2021). A higher variance means greater dispersion.

Standard deviation is the square root of variance. It provides a more intuitive measure of dispersion in the same units as the data(Moore & Notz, 2021).

The **range** is the most basic measure of dispersion and is computed as the difference between the maximum and minimum values in a dataset(Moore & Notz, 2021). Range = Max – Min

Interquartile Range (IQR) - The quartiles show the numbers at certain percentages of the data, such as 25%, 50%, and 75% of the sorted data distribution. By splitting the data into quartiles, we can see how different it is(Moore & Notz, 2021). It is less sensitive to outliers than the range.

C. Measures of Position

Builds on this by showing where individual data points lie relative to the entire dataset.

Quartiles divide data into four equal halves. Q1 is the 25th percentile, meaning 25% of data values are below it (Lind *et al*., 2018). The median, or second quartile (Q2), divides the dataset into two equal pieces with 50% of data points below it. Q3 is the 75th percentile, meaning 75% of data values fall below it. If Q1 is 60, Q2 is 75, and Q3 is 85 in a dataset of test scores from 0 to 100, 25% will score below 60, 50% below 75, and 75% below 85.

Deciles divide a dataset into ten equal sections to refine this idea. A decile represents 10% of data(Lind *et al*., 2018). First decile (D1) is 10th percentile, indicating 10% of data points are below it. D2, the 20th percentile, is followed by D9, the 90th percentile. An income dataset with D1 = $20,000 and D9 = $80,000 shows that 10% of people earn less than $20,000 and 90% less than $80,000.

Percentiles, like quartiles and deciles, partition the dataset into 100 equal segments. Percentiles indicate the proportion of data points that fall below a specified value (Lind *et al*., 2018). The first quartile (Q1) corresponds to the 25th percentile (P25), whereas the median (Q2) represents the

Descriptive statistics aid in summarizing and characterizing a dataset's salient aspects, but they do not reveal information on the correlations between variables or the probability of future occurrences. We use Probability Basics, which establishes the framework for comprehending randomness, likelihood, and the fundamental ideas of inferential analysis, to answer such queries.

PROBABILITY BASICS

Probability is a mathematical idea that tells us how likely a particular event is to occur. It is the basis for many statistical studies and is needed to understand how data can be random. It forms the foundation for many statistical analyses and is essential for understanding randomness in data (Moore & Notz, 2021; Robert & John, 2022; Thomas, 2021; Weihs & Ickstadt, 2018). Probability ranges from 0 (impossible event) to 1 (certain event).

A. Key Concepts

- **Event**: A certain result or set of results of a chance experiment. For example, rolling a die and getting a "4."
- **Sample Space**: The set of all possible outcomes of an experiment. For example, rolling a die: {1,2,3,4,5,6}.
- **Outcome**: A single result from a random experiment. For example, rolling a die and getting "3."

B. Probability Distributions

Understanding the likelihood of events and patterns in data is the focus of probability basics, which build on the knowledge obtained from descriptive statistics. We may measure uncertainty and become ready for inferential analysis, which involves making predictions and testing hypotheses, by investigating ideas like probability distributions, events, and sample spaces. Probability distributions describe how probabilities are distributed across possible outcomes (Robert & John, 2022).

1. **Discrete Distributions**(Thomas, 2021)
   - Deal with outcomes that are countable.
   - **Example**: Number of heads in coin flips (Binomial), number of calls in an hour (Poisson).

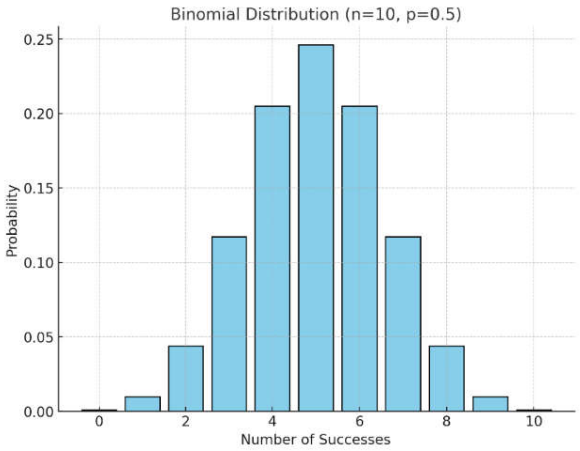

Figure 1: Binomial Distribution

Here is the **Binomial Distribution** graph. It represents the probability of getting a specific number of successes in 10 trials, where the success probability is 0.5.

2. **Continuous Distributions** (Thomas, 2021)
   - Deal with outcomes that are measurable and not countable.
   - **Example**: Heights of individuals (Normal distribution), time between events (Exponential distribution).

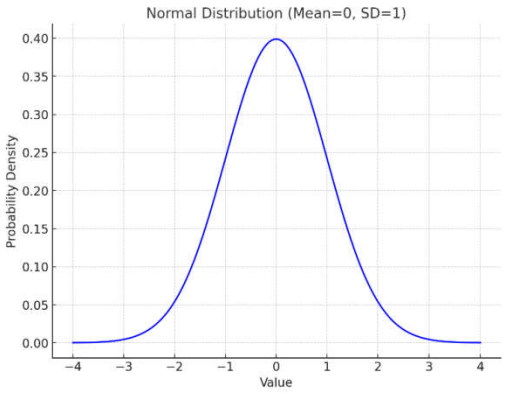

Figure 2: Normal Distribution

This is the **Normal Distribution** graph, showing a symmetric bell curve centered at a mean of 0 with a standard deviation of 1.

C. Normal Distribution

The normal distribution, a key example of continuous probability distribution, assumes symmetry and a specific tail behavior. To assess whether a dataset follows a normal distribution, skewness and kurtosis are critical measures. Skewness evaluates asymmetry, while kurtosis quantifies the peakedness and presence of outliers. Real-world data often deviates from these ideal properties, highlighting the importance of these measures for ensuring the suitability of statistical analyses based on normality. The two concepts of normal distribution should be clarified:

1. Skewness
2. Kurtosis

Skewness is computed using the third standardized moment and is essential for assessing the suitability of different statistical tests, as several tests presume data normality (Kim, 2013). Skewness measures the extent of asymmetry of a distribution in relation to its mean.

A distribution is classified as positively skewed (or right-skewed) if it has a larger tail on the right side, signifying that the majority of data points are concentrated on the left. A negatively skewed (or left-skewed) distribution features an extended tail on the left, indicating that the majority of data points are clustered on the right.

Kurtosis quantifies the "tailedness" or peakedness of a distribution. It is characterized as the fourth standardized moment and conveys information regarding the existence of outliers in the dataset. A distribution exhibiting high kurtosis signifies that data points are more densely clustered in the tails and the peak, indicating an increased probability of extreme values(K. I. Khan *et al*., 2020; Kim, 2013; Zhao *et al*., 2013).

D. Bayes' Theorem

Bayes' Theorem enables the calculation of an event's probability using prior information about associated events(Moore & Notz, 2021). This technique is used for conditional probability.

Formula: P(A|B) = $P(B \mid A) \cdot P(A)/P(B)$

Example: A test for a disease is 95% accurate. Given that 1% of the population is afflicted with the disease, what is the likelihood that an individual who tests positive truly has the condition? Using Bayes' formula, we can calculate $P(Disease \mid PositiveTest)$

E. Law of Large Numbers and Central Limit Theorem

Law of large number - As the number of trials increases, the sample mean approaches the population mean (Moore & Notz, 2021).

Example: Tossing a fair coin multiple time. Initially, the ratio of heads might vary, but as trials increase, it converges to 0.5.

Central Limit Theorem - No matter how the population is distributed, as the sample size grows, the sampling distribution of the mean will get closer to a normal distribution (Moore & Notz, 2021).

Example: Heights of individuals sampled in increasing numbers form a normal distribution, even if the population distribution is skewed.

INFERENTIAL STATISTICS

Inferential statistics tell us about the general trend or behavior of data in a qualitative way. Hypotheses are tested in inferential statistics to see if there is enough evidence in the sample to draw the conclusion that a specific condition is valid for the whole community (Moore & Notz, 2021; Peter & Andrew, 2017; Robert & John, 2022).

A. Hypotheses Testing

A statistical method called hypothesis testing is used to make decisions based on evidence from experiments. The hypothesis test is based on a guess we make about the population measure. We do hypothesis testing in research or data analysis to determine if there is enough evidence in a sample to support or reject a claim about a population. It helps to check assumptions, find connections, and make decisions based on facts that you can be sure of statistically(Moore & Notz, 2021; Peter & Andrew, 2017; Robert & John, 2022).

Key Terms & Concepts

- **Null Hypotheses:** The claim that the effect being studied does not exist is called the null hypothesis, which is also written as H0.
- **Alternative hypothesis:** It is the opposite of null hypotheses. Here, the result is entirely contrary to the assumption. It is denoted by H1.
- **Level of significance:** It refers to the degree of significance for accepting or rejecting a null hypothesis.
- **Type I error:** This type of error occurs when we reject a null hypothesis even if it is correct. It is denoted by alpha.
- **Type II errors:** It occurs when we accept the null hypothesis even if it is false. Type II error is denoted by beta.
- **Test Statistic**: A test statistic compares groups and looks for links between factors when testing a hypothesis—calculated value (e.g., t-value, chi-square) based on the data.
- **P-value**: If the null hypothesis is true, the p-value tells us how likely a sample number will be one or more extremes by chance alone. Our conclusions about the acceptance or rejection of the hypothesis are based on the p-value and the threshold significance level—the probability of observing the data if H0 is true. Compare p-value to α.
- **ANOVA**: ANOVA is used to compare the means of three or more groups to determine if at least one group's mean is significantly different. It helps test the null hypothesis (H0H_0H0) that all group means are equal against the alternative hypothesis (H1H_1H1) that at least one group mean differs.

B. Confidence Intervals

Based on sample data, confidence intervals (CI) estimate the range within which a population parameter (e.g., mean or proportion) is likely to fall. It reflects the degree of certainty or uncertainty in the estimate (Moore & Notz, 2021; Patten & Newhart, 2023; Thomas, 2021). A confidence level, often 95% or 99%, shows the likelihood the interval contains the real parameter if repeated sampling were performed.

Example: A sample of 100 customers yields an average satisfaction score of 80, with a standard deviation of 10. The 95% confidence interval for the mean is calculated as:

$$CI = 80 \pm 1.96 \times \frac{10}{\sqrt{100}}$$
$$= [78.04, 81.96]$$

Interpretation: We are 95% confident that the true average satisfaction score lies between 78.04 and 81.96.

Confidence intervals are crucial in research and data analysis because they provide a range, rather than a single point estimate, for understanding population parameters.

C. Chi-square Tests

Chi-square tests are used to see how two **categorical** data/factors are related to each other. They look at the difference between the frequencies of observed data and those predicted to see if it is just a coincidence or if there is a real connection (Moore & Notz, 2021; Thomas, 2021; Weihs & Ickstadt, 2018).

Example: Testing if gender influences product preference using the following data:

Gender	Product A (Observed)	Product B (Observed)
Male	30	20
Female	25	25

Steps:

1. Calculate expected frequencies: Assuming no association, both genders should equally prefer products.
   - Expected for Product A (Male): $55/100 \times 50 = 27.5$.
   - Expected for Product B (Male):$45/100 \times 50 = 22.5$.
2. Compute the chi-square statistic using: $x^2 = \sum \frac{(O-E)2}{E}$
3. Compare $\chi^2$with the critical value from the chi-square distribution table or use the p-value.

A significant result suggests a relationship between gender and product preference. Chi-square tests are widely used in social sciences, market research, and business to explore relationships between categories, making them an essential tool in inferential statistics(Thomas, 2021).

REGRESSION ANALYSIS

A statistical technique that explains the relationships between one or more independent and dependent variables is known as regression analysis. It is one of the most important tools for estimating what will happen, seeing how things are connected, and finding patterns in data(Moore & Notz, 2021; Robert & John, 2022; Thomas, 2021). Here are the different kinds of regression and what they are used for:

A. Linear Regression

Linear regression uses the relationship between an independent variable and a dependent variable to predict what will happen in the future. It assumes that there is a straight-line relationship between variables (Moore & Notz, 2021; Thomas, 2021).

Example: Predicting customer satisfaction scores based on the response time of a service team. As the response time gets faster, customer satisfaction tends to increase proportionally.

B. Logistic Regression

We use logistic regression analysis when the outcome variable is binary (e.g., yes/no, true/false, 0/1)(Robert & John, 2022). It helps classify data into one of two categories based on independent variables.

Example: Determining whether a loan application will be approved or rejected based on the applicant's income, credit score, and other factors.

C. Multiple Regression

Multiple regression is a regression technique that predicts the value of a dependent variable using two or more independent variables (Moore & Notz, 2021; Peter & Andrew, 2017). It helps identify the combined effect of multiple predictors on the target variable.

Example: Analyzing how advertising spending, product pricing, and market conditions collectively affect sales revenue.

For reliable regression analysis, certain assumptions must be met:

- **Linearity**: There is linear relationship between variables.
- **Homoscedasticity**: The variance of errors should be constant across all levels of the independent variables.
- **Independence**: Observations should be independent of each other.
- **Normality**: The residuals should be normally distributed.

DATA CLEANING AND EXPLORATION

Exploration and cleaning of the data are important steps in data analysis that make sure the information is correct, consistent, and ready for more analysis (Chai, 2020). We can make statistical results better and more reliable by fixing problems with data like outliers, missing values, and data changes. These are important things to keep in mind when cleaning up and exploring data:

Outliers:

Data points that substantially deviate from other observations are known as outliers (Kwak & Kim, 2017). They have the potential to skew statistical analysis and produce inaccurate findings. Z-scores or the Interquartile Range (IQR) can be used to find the outliers. Depending on its reason, we must choose whether to exclude, modify, or examine the outlier independently during the analysis.

Example: In a dataset of employee ages, if most are between 25 and 40, but one value is 90, it's an outlier.

Missing Values

Missing data occurs when some observations lack values for certain variables (Kwak & Kim, 2017). If not handled properly, it can bias the results. We can replace the mission values with the mean, median, or

any predictive value based on other variables.

Example: If 10% of a dataset's income entries are missing, we might replace them with the average income of the existing data.

Data Transformations

Data transformations make the data more suitable for analysis by improving its structure or distribution. Log transformation and square root transformation are the two methods that help reduce skewness or variability in data(Osborne, 2002).

Example: Transforming highly skewed income data using a log transformation to make it more normally distributed for regression analysis.

MULTIVARIATE ANALYSIS

Multivariate analysis is a way to use statistics to look at data sets that have more than one variable. It helps find structures, relationships, and patterns in large amounts of complicated data, which leads to better insights and decisions (Moore & Notz, 2021). The most important and common techniques in multivariate analysis are stated below:

1. Principal Component Analysis
2. Cluster Analysis
3. Correlation Analysis

A. Principal Component Analysis

A method called Principal Component Analysis (PCA) is used to simplify large datasets by finding the most important variables (principal components) that explain the most of the data's variation. It reduces complexity while retaining essential information(M. M. Khan *et al*., 2024).

Example: In a marketing dataset with hundreds of customer attributes, PCA can reduce these variables to a smaller set of principal components for better visualization and analysis.

B. Cluster Analysis

Cluster analysis groups data points according to their commonalities. It is often used to segment data for targeted analysis (Pérez-Ortega *et al*., 2022).

Example: Grouping customers into clusters (e.g., high spenders, occasional buyers) using k-means clustering for personalized marketing strategies.

C. Correlation Analysis

Correlation analysis examines how strong and in what way two or more variables are connected. Common methods include Pearson (for linear relationships) and Spearman (for rank-based relationships) coefficients (Weihs & Ickstadt, 2018).

Example: Analyzing the correlation between hours of study and test scores to determine if increased study time leads to better results.
There are other techniques such as Kernel Principal Component Analysis (KPCA), Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA), and Multivariate Gaussian Process Regression used in data science and research purposes (Abbas & Ghous, 2022; M. M. Khan *et al*., 2024; Medeiros *et al*., 2020).

TIME SERIES ANALYSIS

Time series analysis emphasizes analyzing sequential data over a period of time(Thomas, 2021). This analysis uses three common methods:

1. Trends analysis
2. Seasonality analysis, and
3. Residual analysis.

A. Trends, Seasonality, and Residual Analysis

Trends analysis focuses on long-term patterns in data—for example, a consistent increase in sales volume over a five-year period. On the other hand, **seasonality** examines repetitive patterns or fluctuations, such as increased demand in sales volume during festival time (Weihs & Ickstadt, 2018). Upon eliminating trends and seasonality, the resultant variations are termed **residuals**, signifying random noise or unaccounted alterations, demonstrated by daily sales moves lacking an identifiable pattern.

B. ARIMA Model

For forecasting, **ARIMA (Auto Regressive Integrated Moving Average)** models are widely used. They are made up of three parts: Auto Regressive (AR), which predicts future values based on past ones; Integrated (I), which takes into account trends; and Moving Average (MA), which uses past mistakes to make forecasts more accurate. Predicting monthly energy needs by looking at how it was used in the past is one example (Gebhard & Wolters, 2012). This method gives us a strong way to understand and guess what time data will mean.

DATA VISUALIZATION

Data visualization is a useful tool that lets you look at a dataset and see how it is organized. It makes finding patterns, trends, and outliers in data easy to understand (Peter & Andrew, 2017). Here are three popular ways to visualize things:

1. Histograms,
2. Boxplots, and
3. Scatterplots.

A **histogram** shows how the frequencies of the data in a set are spread out by separating it into bins and showing how many data points are in each bin. It provides a quick understanding of the data's shape and spread (Moore & Notz, 2021; Thomas, 2021). **Example**: For the dataset [5,7,8,10,10,15], a histogram shows the frequency of data points within bins.

A **boxplot** (or whisker plot) represents the distribution of data based on five key metrics: minimum, first quartile (Q1), median, third quartile (Q3), and maximum (Moore & Notz, 2021; Peter & Andrew, 2017). It also highlights outliers. **Example**: For the dataset [5,7,8,10,10,15], the boxplot shows the median (9), the interquartile range (IQR = 5), and potential outliers.

A **scatterplot** shows how two factors are related to each other. Every point on the plot is an observation from the collection (Moore & Notz, 2021). **Example**: Plotting data values the dataset [5,7,8,10,10,15] against their indices (1,2,3,…) reveals patterns or trends in the data.

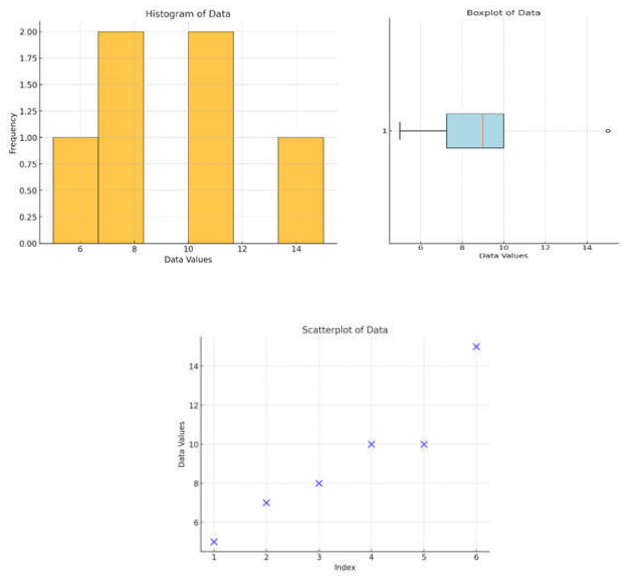

Figure 3: Histogram, Boxplot, and Scatterplot

STATISTICAL SOFTWARE AND TOOLS

In the contemporary world of data science and analytics, we cannot do accurate and useful analyses without statistical software and tools. Python has many useful tools for doing statistical calculations and working with data, such as NumPy, pandas, SciPy, and stats models. In the same way, R is a powerful program known for having a huge number of built-in functions and packages that are used for statistical analysis and data display. Tools like Excel and Tableau are best for simple or quick studies because they are easy to use and can show results visually. Researchers and experts who know how to use these tools well can work with large datasets and get useful information from them.

REAL-WORLD APPLICATIONS

Statistical methods are extensively utilized across diverse disciplines, including business domains such as accounting, finance, marketing, supply chain, and operations, as well as other areas of study, to tackle real-world problems and enable data-driven decision-making that enhances both operational and organizational performance (Uddin *et al.,* 2023b). Businesses use statistics to divide their customers into groups so they can better understand how those groups act and make their plans fit those groups. Statistical algorithms are used by fraud detection models to find strange patterns in financial activities. In industry, predictive maintenance uses statistical tools to figure out when equipment will break down and cut down on downtime. A/B testing is used by marketing teams to compare strategies and make efforts more effective. These uses show how important statistical methods are for dealing with problems in the real world and creating good plans in many areas.

CONCLUSION

The paper emphasizes the use of statistics in training researchers, professionals, and students in domains such as business analytics, data science, and doctorate research. Combining knowledge with practical applications demonstrates how statistical techniques and tools support well-informed decision-making and spur innovation in a range of industries. Addressing challenges such, as grasping procedures integrating knowledge from different areas and ensuring data precision through cleansing and examination highlights the need to equip individuals with a robust statistical acumen. In today's data driven landscape statistics play a role, in solving problems developing models and uncovering valuable insights.

REFERENCES

Abbas, M., & Ghous, H. (2022). Early Detection of Breast Cancer Tumors using Linear Discriminant Analysis Feature Selection with Different Machine Learning Classification Methods. Computer Science & Engineering: An International Journal, 12(1), 171–186. https://doi.org/10.5121/cseij.2022.12117

Arıkan, U., Kranz, T., Schmitt, S., Sal, B., & Witt, J. (2023). Human-Centric Parcel Delivery at Deutsche Post with Operations Research and Machine Learning. INFORMS Journal on Applied Analytics. https://doi.org/53. 333-387. 10.1287/inte.2023.0031

Balakrishnan, N. (2010). Methods and applications of statistics in business, finance, and management science.

Bonett, D. G., & Price, R. M. (2020). Confidence Intervals for Ratios of Means and Medians. Journal of Educational and Behavioral Statistics, 45(6), 750–770. https://doi.org/10.3102/1076998620934125

Chai, C. P. (2020). The Importance of Data Cleaning: Three Visualization Examples. Chance, 33(1), 4–9. https://doi.org/10.1080/09332480.2020.1726112

Ellis, A. R., & Slade, E. (2023). A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. Journal of Statistics and Data Science Education, 31(2), 128–133. https://doi.org/10.1080/26939169.2023.2223609

Gastelum, V. S., Almaguer, C. A. G., Rabanales, E. G. A., Acosta, A. C. A., Montes, E. C., & Ramirez, C. Z. (2023). The importance of statistics in data science, how to redesign classroom learning in the TEC21 educative model. 2023 Future of Educational Innovation-Workshop Series Data in Action, FEIWS 2023, 10105024. https://doi.org/10.1109/IEEECONF56852.2023.10105024

Gebhard, K., & Wolters, J. (2012). Introduction to Modern Time Series Analysis. In Sustainability (Switzerland) (Vol. 11, Issue1). Springer. http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

Gupta, P., & Tawar, N. (2020). The Impact and Importance of Statistics in Data Science. International Journal of Computer Applications, 176(24), 10–14. https://doi.org/10.5120/ijca2020920215

Hofmann, L. A., Lau, S., & Kirchebner, J. (2022). Advantages of Machine Learning in Forensic Psychiatric Research—Uncovering the Complexities of Aggressive Behavior in Schizophrenia. Applied Sciences (Switzerland), 12(2). https://doi.org/10.3390/app12020819

Hsieh, W. W. (2023). Introduction to Environmental Data Science.

Khan, K. I., Naqvi, S. M. W. A., Ghafoor, M. M., & Akash, R. S. I. (2020). Sustainable portfolio optimization with higher-order moments of risk. Sustainability (Switzerland), 12(5), 1–14. https://doi.org/10.3390/su12052006

Khan, M. M., Islam, I., & Rashid, A. B. (2024). Fault Diagnosis of an Industrial Chemical Process using Machine Learning Algorithms: Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA). IOP Conference Series: Materials Science and Engineering, 1305(1), 012037. https://doi.org/10.1088/1757-899x/1305/1/012037

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative Dentistry & Endodontics, 38(1), 52. https://doi.org/10.5395/rde.2013.38.1.52

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. Korean Journal of Anesthesiology, 70(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407

Liu, Z. (2023). Review on the influence of machine learning methods and data science on the economics. Applied and Computational Engineering, 22(1), 137–141. https://doi.org/10.54254/2755-2721/22/20231208

Lytsy, P., Hartman, M., & Pingel, R. (2022). Misinterpretations of P-values and statistical tests persists among researchers and professionals working with statistics and epidemiology. Upsala Journal of Medical Sciences, 127. https://doi.org/10.48101/UJMS.V127.8760

Martha, S. (2024). The Importance of Statistics in Medical Healthcare. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/12. 303- 304. 10.22214/ijraset.2024.63920

Medeiros, A. D. de, Silvia, laercio J. da, Ribeiro, J. P. O., Ferreira, K. C., Rosas, J. tadeu F., Santos, A. A., & Silva, C. B. da. (2020). Machine Learning for Seed Quality Classification : An Advanced Approach Using Merger Data from. Sensors, 20, 1–12.

Moore, D. S., & Notz, W. I. (2021). The Basic Practice of Statistics (Ninth Edit). Macmillan Learning.

Osborne, J. W. (2002). Notes on the use of data transformations . Data transformation and normality. Practical Assessment, Research & Evaluation, 8(6), 1–7.

Patil, S., & Satagopan, J. (2022). Building and Teaching a Statistics Curriculum for Post-Doctoral Biomedical Scientists at a Free-Standing Cancer Center. CHANCE : New Directions for Statistics and Computing, 35, 55–56.

Patten, M. L., & Newhart, M. (2023). Understanding Research Methods. In Understanding Research Methods. https://doi.org/10.4324/9781003092049

Penchev, D. (2021). Role of Statistical Methods in Pedagogical Research. Pedagogical Almanac, 29(2), 169–178. https://doi.org/10.54664/chkk3190

Pérez-Ortega, J., Almanza-Ortega, N. N., Torres-Poveda, K., Martínez-González, G., Zavala-Díaz, J. C., & Pazos-Rangel, R. (2022). Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico. Mathematics, 10(13). https://doi.org/10.3390/math10132167

Peter, B., & Andrew, B. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts (1st Editio, Vol. 63). O'Reilly Media.

Priya, M., Punithavall, M., & Rajesh, K. (2021). Conceptual Review on Machine Learning Algorithms for Classification Techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(1), 215–222.

Robert, S. W., & John, S. W. (2022). Stastistics (Ninth Edit). Wiley.

Salloum, S. J., Young, T., & Brown, R. D. (2016). Teaching Introductory Quantitative Research Methods to Doctoral Students in Educational Leadership: Using Real Data to Improve Statistical Literacy. 1, 203.

Sarraf, S. (2023). Formulating A Strategic Plan Based On Statistical Analyses And Applications For Financial Companies Through A Real-World Use Case. http://arxiv.org/abs/2307.04778

Thomas, C. G. (2021). Research Methodology and Scientific Writing. In Aslib Proceedings (Second Edi). Springer. https://doi.org/10.1108/eb051376

Uddin, M. S., Eltahir, O., Mohamed, B., & Ebert, J. (2024). Effect of Sustainable Supply Chain and Management Practices on Supply Chain Performance : The Mediating Roles of Ethical Dilemmas and Leadership Quality. 4, 13–32.

Uddin, M. S., Habib, M. M., & Mohamed, O. E. B. (2023a). Exploring the Interconnectedness of Supply Chain Management Theories: A Literature Review. International Supply Chain Technology Journal, 9(4). https://doi.org/10.20545/isctj.v09.i04.03

Uddin, M. S., Habib, M., & Mohamed, O. E. B. (2023b). The Role of Supply Chain Finance on Supply Chain Management and Firm ' s Performance : A Conceptual Framework. https://doi.org/https://doi.org/10.20545/isctj.v09.i06.01

Weihs, C., & Ickstadt, K. (2018). Data Science: the impact of statistics. International Journal of Data Science and Analytics, 6(3), 189–194. https://doi.org/10.1007/s41060-018-0102-5

Zhao, H., Zhang, J. E., & Chang, E. C. (2013). The relation between physical and risk-neutral cumulants. International Review of Finance, 13(3), 345–381. https://doi.org/10.1111/irfi.12013

****