Damak Campus Journal 2023, **11 (1): 53-67**
ISSN : 2565-4772 (Print)

SEVERAL APPROPRIATE STATISTICAL TECHNIQUES IN RESEARCH

Shantiram Subedi

Lecturer, Department of Physics' and Mathematics
Damak Multiple Campus
E-mail: shantiramsubedi2020@gmail.com

ABSTRACT

This paper has the main purpose to assist researchers and students in choosing the appropriate statistical test for studies that examine one variable or more variables. The aim of this article is to consider the role which statistical methods can sensibly take in some relevant fields. Other objective of writing this paper is to provide the algorithm of choosing the statistical techniques in the data analysis. This article is concentrated to define data analysis and the concept of data preparation. Then, the data analysis will be discussed. The article covers a brief outline of the variables, an understanding of quantitative and qualitative variables and measure of central tendency. This article will also try to apprise the reader with the basic research techniques that are utilized while conducting various studies. Finally, there is a focus on parametric and nonparametric tests for data analysis and various strategies in this concept.

Keywords: Vairable, dependent variable, independen variable, tex, significace

INTRODUCTION

The statistical analysis gives meaning to the meaningless numbers, thereby breathing life into a lifeless data. Nowadays statistics is the one of the most important parts in all sectors. Without of the knowledge on the tools and techniques of statistical nobody can write the quantitative research paper. Choosing an appropriate test is one of the most important tasks in search. So, the right test will give the valid conclusion and wrong test give the misleading inference. To choose the right statistical test, we should be familiar with different variables and their nature. The results and inferences are precise only if proper statistical tests are used. There are various rules in the statistics for the data analysis and condition for the choosing of suitable test statistics. The data collected for the information are not useful because they are referred to as raw data or intreated data until they are analyzed by using appropriate statistical tools.
The collected data and research design of the study must fit appropriate data analysis.

Research objectives

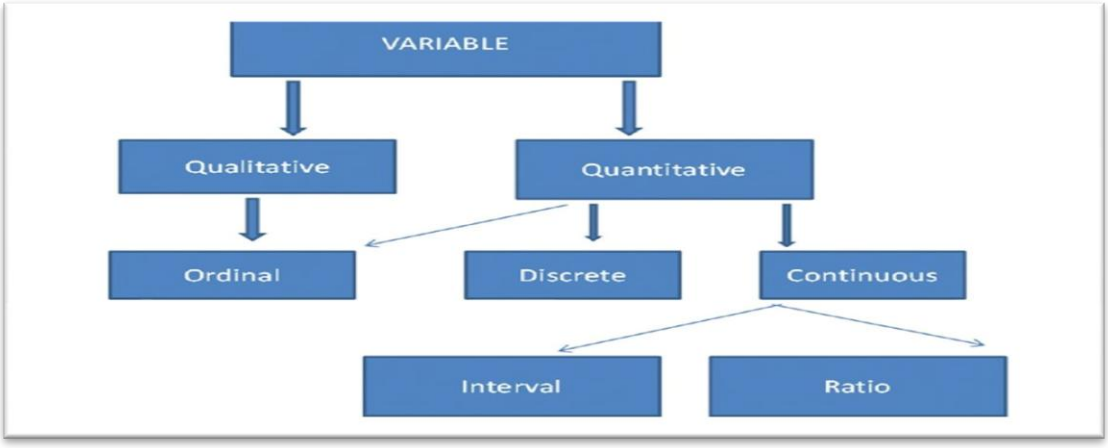
The following three basic questions to be answered:

i. What type of research question did the study formulate or asked?
ii. What type and number of variables does the study want to analysis?
iii. What nature and characteristics of variables does the study have?

Nature of Variables and Types of Data

A variable is any characteristics, number, or quantity that can be measured, or counted. A variable may also be called a data item. A variable may be also called a data item. Age, sex, business, income, expenses, capital, eye color, vehicle type, are examples of variables. Any research that deals with the manipulation of variables which are basically of two types; these are numerical (**quantitative**) and categorical (**qualitative**). Numerical variables are recoded as numbers such as height, age, score, weight etc. Categorical variables could be dichotomy (for example: male or female), trichotomy (for example: high, low, medium and low economic status) or polychotomy (for example: birth places).

Quantitative variable can be further classified into two groups: discrete and continuous. Discrete variables assume values that can be counted such as 0, 1, 2, 3, … using integers. Continuous variables, by comparison, can assume all values in an interval between any two specific values. Pages is discrete variable and temperature is a continuous variable since it can assume all values between any two given temperatures.



MATERIAL AND METHODS

Dependent Variable: A variable that may depend on the other factor is termed as dependent variables. e.g., exam score is a variable may change depending on the student's age.

Independent Variable: A variable that does not depend on the other factor is termed as independent variable. e.g., student's age does not change depending upon exam score.

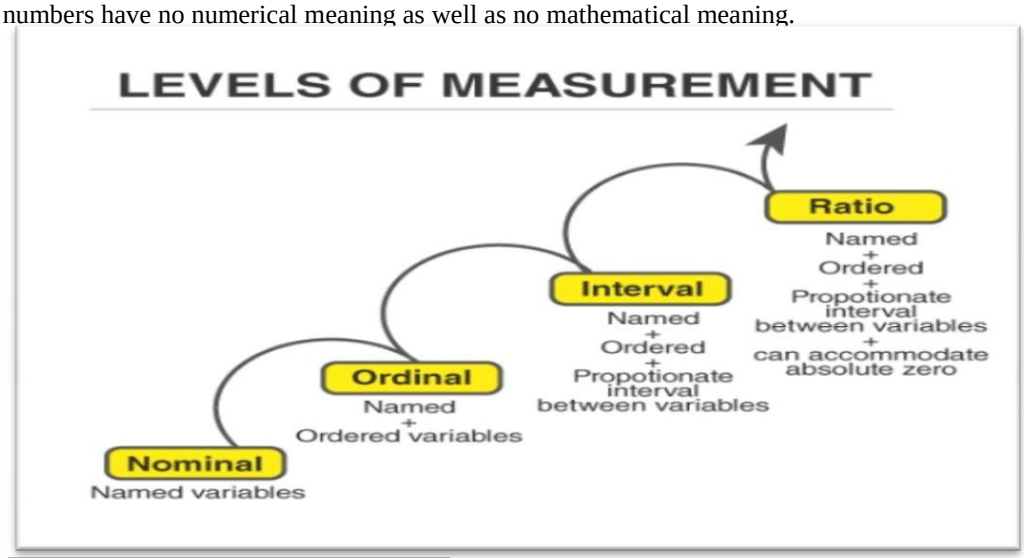
Random Variable: A random variable is a rule that assigns a numerical value to each outcome in a sample space. Random variable may be either discrete or continuous.

Data Measurement Scales

Measurement of statistical data is essential for further statistical analysis. Measurement is a process of assigning numbers or symbols to any facts or objects or product or terms

according to some rule. It is a tool by which individuals are distinguished on the variables of area under study. Scale is simply a range of levels or numbers used for measuring something. It is a set of all the different levels of symbols or numerals. Different measurement scales are used based on nature of data. These measurement scales of variables under study are (a) Nominal scale (b) Ordinal scale (c) Interval scale (d) Ratio scale.

- **Nominal Scale:** It is the simplest or lowest type of scale. It is simply a system of assigning number or the symbols to objects or events, to distinguish one from another or to label them. The symbols or the numbers have no numerical meaning as well as no mathematical meaning.



Lowest level of measurment **Highest level of measurement**

- **Ordinal Scale:**

It is the quantification of items by ranking. In this scale, the numerals are arranged in some order but the gaps between the positions of the numerals are not made equal. It represents qualitative values in ascending or descending order. The rank orders represent ordinal scale and mostly useful in scaling the qualitative phenomena.

- **Interval Scale:**

In addition, ordering the data, this scale uses equidistance units to measure the difference between scores. This scale does not have absolute zero but only arbitrary zero. For example, scale of temperature is an ordinal scale. The temperature 320F and 40F are not viable to express in ratio because the zero is not a true zero but is an arbitrary point.

- **Ratio Scale:** Ratio scale is the ideal scale and an extended form of interval scale. It is most powerful scale of measurement. It possesses the characteristics of nominal, ordinal, interval scale. Ratio scale has an absolute zero or true zero or natural zero of measurement. The true zero point or the initial point indicates the completely absence of that property of an object what is being measured. Numbers on the scale indicates the actual amount of property being measured.

Statistics and its area

Statistics is a branch of mathematics that deals with collecting, organizing, analyzing, and interpreting, and presenting data. It is used in a wide range of fields, including business, economics, psychology, biology, and engineering, to make informed decisions based on data. There are several areas within statistics, including:

- Descriptive Statistics: deals with summarizing and describing the main features of a set of data.
- Inferential Statistics: deals with drawing conclusions about a population based on a sample of data.
- Probability: deals with quantifying the likelihood of an event occurring.
- Hypothesis Testing: deals with the process of testing claims or hypothesis about a population based on sample data.
- Regression Analysis: deals with finding the relationship between two or more variables.
- Bayesian Statistics: deals with the incorporation of prior knowledge and updating beliefs based on new data.
- Time Series Analysis: deals with analyzing and modeling data collected over time.
- Multivariate Statistics: deals with analyzing and modeling data with multiple variables.

Statistical Analysis and its types

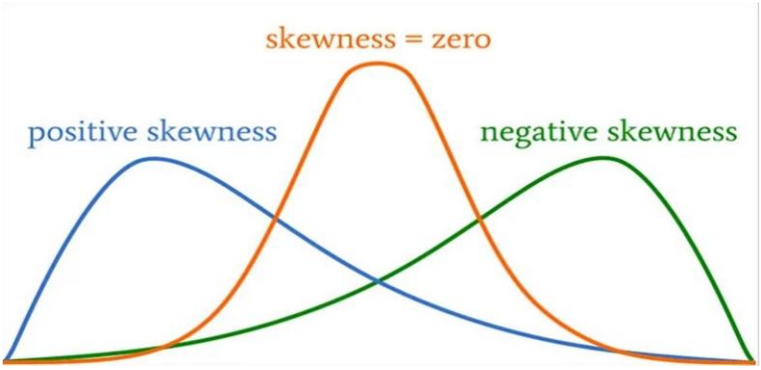
Statistics is a set of scientific principles and techniques that are useful in reaching conclusions about population and process when the available information is both limited and variable; that is, statistics is the science of learning from data. The objective of statistics is to make inferences about a population of interest based on information obtained from a sample of measurements from that population. Statistical Analysis is the process of collecting and analyzing data to discern pattern and trends. In simple words, statistical analysis is a data analysis tool that helps draw meaningful conclusions from raw and unstructured with numbers and is used by business and other institutions to make use of data to derive meaningful information. Here we discuss six types of statistical analysis.

- **Descriptive Analysis:** Descriptive statistical analysis involves collecting, interpreting, analyzing, and summarizing data to present them in the form of charts, graphs, and tables. Rather than drawing conclusions, it simply makes the complex data easy to read and understand.
- **Inferential Analysis:** The inferential statistics analysis focuses on drawing meaningful conclusions o the basis of the data analyzed. It studies the relationship between different variable or makes prediction for the whole population.
- **Predictive Analysis:** Predictive statistical analysis is a type of analysis that analyzes data to derive past trends and predicts future events on the basis of them. It uses machine learning algorithms, data mining, data modelling and artificial intelligence to conduct the statistical analysis of data.

- **Prescriptive Analysis:** The prescriptive analysis conducts the analysis of data and prescribes the best course of action based on the results. It is a type of statistical analysis that helps to make as informed decision.
- **Exploratory Data Analysis:** Exploratory analysis is like inferential analysis, but the difference is that it involves exploring the unknown data association. It analyzes the potential relationships within the data.
- **Casual Analysis:** The casual statistical analysis focuses on determining the cause-and-effect relationship between different variables within the raw data. In simple words, it determines why something happens its effect on the other variables. This methodology can be used by business to determine the reason for failure.

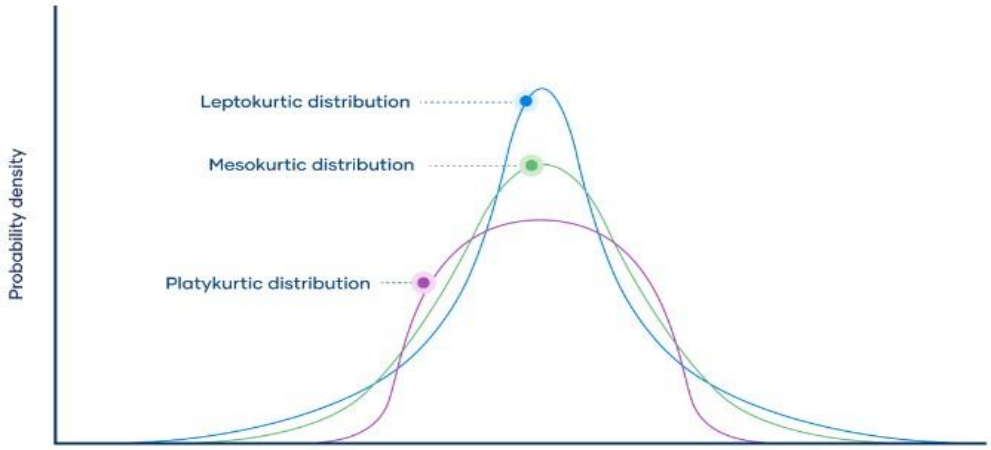
Statistics can be very broadly classified into two categories, viz, descriptive statistics and inferential statistics. Descriptive statistics refers to the type of statistics which deal with collection, organizing, summarizing describing qualitative data. It deals to any forms whereby data are displayed for easier understanding.

Descriptive statistics summarize the numerical data through frequency distribution, central tendency, variation and data shape. Skewness is an important measure of the shape of a distribution. It measures the degree of departure from symmetry. It is used to determine the nature and extent of the concentration of the observations towards higher or lower values of the variable. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean ≠ median ≠ mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could be either be positively skewed or negatively skewed.



Depiction of positive skewness, zero skewness and negative skewness

Kurtosis measure provide information about the peakdness of the distribution. Zero or near to zero value indicate that the distribution is normal or mesokurtic. Positive kurtosis (Leptokurtic) values indicates that the distribution is relatively peaked because many cases cluster in the center. Negative kurtosis (Platykurtic) values indicates that the distribution is a relatively flat because too many cases in the extremes.



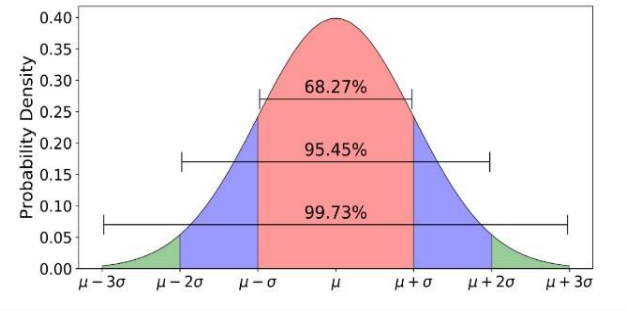
General forms of Kurtosis

Inferential statistics deals with the methods by which inferences are made on the population based on the observations made on the smaller sample. Any procedure of making generalization that goes beyond the original data is called inferential statistics. These statistics provide a way of testing the significance of results obtained when data are collected. Examples of inferential statistics are Student t-test, Analysis of variance, Analysis of covariance, Correlation Analysis, Multiple regression analysis, multivariate Analysis of variance etc. The attempt to choose the right test to compare measurements may however a bit difficult, since we must choose between two families of tests: Parametric and Non-parametric.

Parametric and Non-parametric Tests

Many statistical tests are based upon the assumption that the data are sampled from a normal distribution. Parametric statistics are statistics where the population is assumed to fit any parametrized distributions (mostly typically the normal distribution). These tests are referred to as parametric tests. The normal distribution, also called the Gaussian distribution is a continuous probability distribution for a real valued random variable X with the probability density function.

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} dx\,; \quad -\infty < x < \infty.$$



Normal Distribution

In this distribution, each member of the family may be defined by two parameters, the mean(μ) and the variance (σ^2) wherein values lie in a symmetrical fashion mostly situated around the mean (μ). Commonly used parametric test include the mean, the standard deviation, the t-test, the one-way ANOVA, the Pearson product moment correlation, the simple linear regression, the Multiple linear regression etc.

Although parametric techniques are robust, that is, they often retain considerable power to defect difference or similarities even when these assumptions are violated, some distributions violate the assumptions so markedly that a non-parametric alternative is more likely to defect a difference or similarity. Hence, tests that do not make assumptions about the population distribution are referred to as non-parametric tests. Specially, non-parametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population (hence, the name non-parametric. In more technical terms, non-parametric methods don't rely on the estimation of parameters (such as the estimation or the standard deviation) describing the distribution of the variable of interest in the population. Therefore, these methods are also sometimes (and more appropriately) called parameter-free methods or distribution-free methods. Commonly used non-parametric test include the median, the interquartile range, the Wilcoxon test, the Mann-Whitney test, the Kruskal-Walli's test, The Freidman test for dependent samples, the Chi-square test, the Spearman correlation etc.

Basic Statistical Techniques

Statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypothesis and make estimates about population measures. Some well-known statistical test and procedure for research observations are discussed here briefly.

FINDINGS AND DISCUSSION

The t-test

The t-test is the mostly commonly used method to estimate the difference in means between two groups. Theoretically, the t-test can be used even if the sample sizes are very small (e.g., as small as 10) as long as the variables as the variables are normally distributed within each group and the variables of score in the two groups is not reliably different. There are specific assumptions underlying the use of the t-test:

- The sample data should be normally distributed.
- The sample must be representative of the population so that we can generalized at the end of the analysis.
- Equal variances are assumed when two independent samples are used to test a hypothesis.
- The dependent measurement involved in the calculation of the means must come from either interval or ratios.

A fundamental issue in the use of the t-test is often whether the samples are independent or dependent. Independent samples typically consist of two groups with no relationship while dependent samples typically consist of a matched sample or one group that

has been tested twice (repeated measures). The p-value reported with a t-test represents the probability of error involved in accepting the research hypothesis about existence of a difference. It is the probability of error associated with rejecting the hypothesis of no difference between the two categories of observations in the population when, in fact, the hypothesis is true. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.05 level); then the null hypothesis which usually states that the two groups do not differ is rejected in favor of an alternative hypothesis; which typically states that the groups do differ.

ANOVA/ANCOVA/MCA

The purpose of analysis of variance (ANOVA) is to test differences in means (for groups or variables) for statistical significance. This is accomplished by analyzing the variance, that is, by partitioning the total variance into the component that is due to true random error and the components that are due to differences between means. These latter variances components are then tested for statistical significance, and if significant, we reject the null hypothesis of no difference between means and accept the alternative hypothesis that means (in the population) are different from each other.

Analysis of covariance (ANCOVA) is a general linear model with one continuous explanatory variable and one or more factors. It is a merger of ANOVA and regression for continuous variables.

Multiple Classification Analysis (MCA) is a technique for examining the interrelationship between several predictor variables and one dependent variable in the content of an additive model.

Multivariate Analysis of Variance (MNOVA)

The Multivariate Analysis of Variance (MNOVA) is designed to test the significance of group difference. Multi various analysis (MNOVA) is a is a statistical method to analyze the relationship between multiple dependent and one or more independent variables. MNOVA can include several dependent variables, whereas ANOVA can handle only one dependent, whereas ANOVA can handle only on dependent variable. In MANOVA, the dependent variables are analyzed together as a set, rather than separately as in ANOVA. The goal is to determine whether there is significant differences between the groups on the set of dependent variables, whole taking into accounts the correlations between the dependent variables. MNOVA is based on the following assumptions:

- The observations within each sample must be randomly sampled and must be independent of each other.
- The observations on all dependent variables follow a multivariate normal distribution in each group.
- The observations on all dependent variables follow a multivariate normal distribution in each group.
- The population covariance matrices for the dependent variables in each group must be equal.
- The relationship among all pairs of dependent variables for each cell in the data matrix must be linear.

Regression Analysis

Regression analysis is a statistical technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (responsive variables) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters (constants), and an error term. The error terms represent unexplained variation in the dependent variable. The parameters are estimated so as to give a best fit of the data. Most commonly the best fit is estimated by using the least squares methods. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. The best of underling assumptions in regression analysis is that:

- The sample must be representative of the population for the inference prediction.
- The dependent variable is subject to error. This error is assumed to be a random variable with mean zero.
- The independent variable is error free.
- The predictors must be linearly independent. That is, it must not be possible to express any predictor as a linear combination of the other.
- The errors are uncorrelated.
- The variable of the errors is uncorrelated.
- The variance of the errors is constant.
- The errors follow a normal distribution.

For example, a multiple regression model is

$$y = m_0 + m_1 x_1 + m_2 x_2 + \cdots + m_n x_n \quad ---[1]$$

Where, y is the dependent variable; $x_1, x_2, \ldots, x_n$ are the independent variables; and $m_0, m_1, \ldots, m_n$ are the parameters. Furthers, m_0 is the intercept of the regression line and $m_1, m_2, \ldots, m_n$ are referred to as the partial regression coefficients. These values are then used to create predicted values of the outcome, with the observed or true value from the data designated as y and the predicted value as $\hat{y}$.

Furthermore, in equation (1), the value m_1 measures the casual effect of a one unit increase of x_1 on the value of y. The parameter m_1 is also referred to as the regression coefficient for x_1 and is the average amount of the dependent variable increase when the independent variable increases one unit and other independent variables are held constant. Thus, when the independent variable x_1 increases by 1, the dependent variable y increases by m_1 units.

Hypothesis testing are conducted to show whether the parameters that have been estimated are statistically significant or, whether the independent variables contribute to the explanation of variation in the dependent variable. If we are able to reject the null null hypothesis at an acceptance significance level, then we conclude that the parameter is not statistically significant.

The quality of fitness of the model is determined by the R^2. It lies in between 0 and 1. High value of R^2 will indicate that the model fits the data well. A limitation in the use if R^2 is that its value increases with the number of explanatory variables. The power of the test is therefore affected. Thus, the adjusted R-square $(\bar{R}^2)$ was developed to take care of the inadequacies.

Correlation

Correlation is a measure of the relation between two or more variable. It indicates the strength and direction of a linear relation between two or more variables. It is denoted by the symbol 'r' and can take values between -1 and 1. If the value of r is closer to 1, it indicates a strong positive correlation, which means that when one variable increases, the other variable also tends to increase. On the other hand, if the value of r is closer to -1, it indicates a strong negative correlation, which means that when one variable increases, the other variable tends to decreases. A value of r close to 0 indicates that there is no linear relationship between the two variables. The closer the value of r is to 0, the weaker the correlation between the variables.

Regression Analysis

Regression analysis is a statistical method that is used to investigate the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to create a mathematical model that can be used to predict the values of the dependent variables based on the values of the independent variables. Regression analysis can be used for both linear and non-linear relationships between the variables**.** regression analysis involves estimating the coefficients of the mathematical model, which can be done using various methods such as least squares, maximum likelihood, or Bayesian estimation. Once the coefficients are estimated, they can be used to predict the values of the dependent variable for any given values of the independent variables.

There several types of regression analysis, including simple linear regression, multiple linear regression, logistic regression, and polynomial regression. Simple linear regression involves a single independent variable, while multiple linear regression involves two or more independent variables. Logistic regression is used when the dependent variable is categorical, and polynomial regression is used when the relationship between the variables is non-linear.

Choosing Appropriate Statistical Techniques in a research Enterprise:

Statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypothesis and make estimates about population measures. Not only it is important to have sample size that is large enough, but also it is necessary to see how the subjects in the sample were selected. Volunteers generally do not represent the population at large.

The computer merely gives numerical answer and save time and effort of doing calculation by hand. It will be our duty to understand and interpret computer printout correctly. Note that data can be subjected to parametric and nonparametric statistics depending on the nature of data.

The variable type determines to some extent the type of statistical (descriptive or inferential) method that it will support. To find out whether the performance is significantly different, there need to make inference of the parametric statistics: t-test. When we compare two mean scores, determination f F- ratio is involved: ANOVA or ANCOVA depending on the design employed for the study. If a study involves determination of relationship, we can use Spearman Rank order correlation, Pearson Product moment correlation, Chi-square statistics or even multiple regression analysis. All depends on the nature of research. Chi-square shows the degree of association between two different bases of classification. The Z-test is used only when the population parameters are known and the variable of interest is normally distributed in the parent population. If the two condition are met, Z-test is used as an exact test, even for small samples (n<30). However, if the variable is not normally distributed, a large sample permits the use of a Z-test. In most research, the Z-test for single mean is rarely encountered because the conditions of normality and known parameters (σ) are rarely met. Normality, Z-test is used to test for mean of a large sample, and t-test for the mean of small sample.

The collected data and research design of the study has to fit appropriate analysis. The following three basic questions to be answered.

i. What type of research question did the study formulate or asked?
ii. What type and number of variable of variables does the study want to analysis?
III. What nature and characteristics of variable does the study have?
iv. Does it need for parametric or non-parametric?

Independent Sample T-test

An independent sample t-test, also known as a two-sample t-test, is a statistical hypothesis test used to compare the means of two independent groups. It is parametric test, meaning it assumes that the data follows a normal distribution and that the variances of the two groups are equal. Example: Are there significant difference in SME's outputs between male and female? For this we need two variables

- One categorical independent variable with two groups {Gender: male and female}
- One continuous dependent variable {SME's output}

Its non-parametric alternative is Mann-Whitney U Test.

Paired Sample (or dependent) T-test

It is used to compare how a group of subjects perform under two different test conditioned. It could involve before and after measurement of the same continuous variable. Example: What effect does training classes have on performance results of staff? For this we need two variables

- One categorical independent variable {scores from two different periods}
- One continuous dependent variable {performance result of staff}

Its non-parametric alternative is Wilcoxon Signed Tank-test.

One sample T-test

It is used when we have data from a single sample of participants and we wish to know whether the mean of the population from which the sample is drawn is the same as the hypothesized mean. It is used to determine whether a sample comes from a population with a specific mean. Example: A researcher might to test whether the average IQ score for a group of students differs from 100 at 95% confidence level.

Its non-parametric alternative is one- sampled Wilcoxon Signed Rank test.

One way ANOVA Test

This is used to compare the difference between the means of a continuous outcome variable of three or more groups. Example: What impact does food choices have on body weights?

For this we need two variables

- One categorical independent variable with three or more distinct categories of food such as fast food, fruits, protein packed food etc.
- Other continuous dependent variable such as body weights

Its non-parametric alternative is Kruskal – Wallis Test.

Two-way ANOVA Test

This technique allows allow us to look at the individual and joint effect of two independent variables on one dependent variables. It is used for comparing mean combination of two independent categorical variables on a continuous dependent variable. Example: What impact does gender and work experience have on sales volume? For this we need three variables.

- Two categorical independent variables such as gender and work experiences
- And one continuous dependent variable such as sales volume.

MNOVA Test

Multivariate analysis of variance (MNOVA) is used when we want to compare groups on a number of different, but related dependent variables. Example: What impact does job categories have on salaries and expenses of workers?

For this we need

- One categorical independent variable {Job categories: Clerk, Supervisor, Manager etc.}
- Two (or more) continuous dependent variables {Salaries and expenses of workers}

Pearson Correlation Coefficient Test

The Pearson correlation coefficient describes the strength and polarity of a linear relationship between two continuous variables. Example: What is the correlation between income and savings? For this we need

- Two continuous variables {Income and Savings} The t-test is used to check the significance of the correlation coefficient. Its non-parametric alternative is Spearman Rank correlation.

Linear Regression Test

In a linear regression, the relationship between two or more continuous or categorical variables is modelled with a line of best fit. It is used to explore the predictive ability of a single or set of independent variables. Example: What is the impact of advertisement budget and taxes on sales revenue? For this we need

- Two continuous independent variables {advertisement budget and taxes}
- One continuous dependent variable {sales revenue}. Its non-parametric alternative is Ordinal or Multinomial logistic regression.

Multivariate Linear Regression Test

This method is used to measure the degree at which more than one independent variable (predictors) and more than one dependent variable (responses) are linearly related. It is broadly used to predict behavior of the response's variables associated with changes in the predictor's variables, one a desire degree of relation has been established. Example: What effect does taxes, salaries and expenses have on revenue, sales output and profits? For this we need

- Three continuous independent variables {taxes, salaries, expenses}
- Three continuous dependent variables {revenue, salaries, outputs, and profits}

Its non-parametric alternative is Multivariate Logistic Regression.

Logistic Regression Test

In logistic regression, the relationship between two or more continuous or categorical outcome (or dependent) variable is modelled with a line of best fit. The types of logistic regression are:

- Binary logistic regression is utilized in those cases when a researcher is modeling a predictive relationship between one or more independent variables and a dichotomous dependent variable.
- Simple logistic regression analysis applies when there is a single dichotomous outcome and more than one independent variable.
- Multiple logistic regression analysis applies when there is a single dichotomous outcome and more than one independent variable.

Multinomial Logistic Regression Test

This is used when we have a categorical dependent variable with two or more uncorrelated levels (i.e., two or more discrete outcomes) and one or more continuous or categorical independent variables. This is a simple extension of binary logistic regression that allows for more than two categories of a dependent or outcome variable. It is often considered an attractive analysis because it does not assume normality, linearity or homoscedasticity. Example: What effect does salaries and education levels have on food choices?

For this we need:

- Two independent variables (continuous and categorical) {Salaries and education levels}.
- One unordered (categorical) dependent variable: different food choices {fast food, fruits, protein, packed food etc.} Its parametric alternative is Linear Regression.

Ordinal Logistic Regression (OLR)

This is generally used when we have categorical categories for the dependent variables that are ordered (i.e., are ranked) and one or more continuous or categorical independent variables. OLR yields only a single set of regression coefficients to estimate relationship between independent (continuous or categorical) variables and ranked (categorical) dependent variables. Example: What effect does salaries and education levels have on job performance? For this we need:

- Two independent variables (continuous and categorical) {salaries and education levels}
- One ordered (categorical) dependent variable: ranked job performance levels: very high, high, moderate, low, very low etc.}

Its parametric alternative is Linear Regression.

Chi-square test for independence

This test is ideal for comparing two categorical (nominal) variables regardless of the number of subgroups or levels per variable. This chi-square test identifies whether the observations differ significantly from what would be expected by chance and thus tests for statistical significance. Example: What is the relationship between gender distribution and job categories?

For this we need:

- Two categorical variables with one or more categories in each {gender distribution}
- Independent variables {male and female and job categories}
- Dependent variables {clerical, custodian, manager etc.}

Chi-square test for Goodness of fit

This test explores the proportion of cases that fall into the various categories of a single variable and compares these with hypothesized values. It is also referred to as one- sample chi-square. For this we need one categorical variable with fine distinct categories.

CONCLUSION

This article provided a summary of the most common data analysis techniques. The common methods are revived, and the tools for the most important techniques are discussed. Also, there is a summary of parametric and non-parametric tests for data analysis.

REFERENCES

Nasser Said G.A., Muhhamad A.S. et. al., 2022. "Statistical Analysis Tools: A Review of Implementation and effectiveness of Teaching English." *ResearchGate, 2—4.*

F.A. Adesoji and M.A. Babatonde. 2009. "Basic Statistical Techniques in Research." *ResearchGate,4-20.*

J. Karthikeyan, H.P. Horizan, et.al., 2018. "Statistical Techniques and Tools for Describing and Analyzing data in Elt Research." *International Journal of Civil engineering and Technology.*

D.R. Cox, G. Gudmundsson, et.al., 1981." Statistical Analysis of Time Series. *"Scandinavian Journal of Statistics,93-115.*

Karim Elbahloul. 2019. "Stock Market Prediction Using Various Statistical Methods Volume I."*ResearchGate.*

Gianluca Malato. 2020. "Statistical analysis of Stock price."*Toward Data Science.*

Refael A. Irizarry. 2019. "Introduction to Data Science: Data Analysis and Prediction Algorithm with R." *Chapman and Hall/CRC;1st edition.*

Lyle F. Bachman. 2005. "Statistical Analyses for Language Assessment Workbook and CD-ROM." *Cambridge University Press, 29-54.*

Susan Trocoso Skidmore and Bruce Thompson. 2010. "Statistical Techniques Used in Published Article: A Historical Review of Reviews." *SageJournals.*

Bangert, Q.W.,&Baumberger, J.P. 20055."Research and Statistical Techniques used in the journal of Counseling & Development: *Journal of Counseling & Development, 83,480- 487.*Agresti, A. 2007. "An Introduction to Categorical Data Analysis." *The Statistical Analysis of Composition Data, Chapman & Hall.*

Douglas A.Lind, Willam G. Marchal and Samuel A. Wathen. 2012." Statistical Techniques in Business & Economics." *MaGraw-Hill/Irwin.*