

Data Analysis in Business Research: Key Concepts

Niharika Singh, Dr. Amit Kumar Singh

Research Scholar, Dept. of Management, Mizoram University Aizawl
Assistant Professor, Dept. of Management, Mizoram University Aizawl

Abstract

The data may be adequate, valid and reliable to any extent; it does not serve any worthwhile purpose unless it is carefully analyzed. There are a number of techniques can be used while analyzing data. These techniques fall into two categories; descriptive and inferential, constituting descriptive and inferential analysis. They can serve many purposes: to summarize the data in a simple manner, to organize it so it's easier to understand and to use the data to test theories about a larger population. Given the ready availability of computer software, tedious formulae and calculations can be avoided today. But, there is no substitute for having a good understanding of the conceptual basis of analytic methodologies that one applies in order to draw inferences about hard-own research data. Hence, an effort has been made in this paper to provide theoretical introduction of few but most widely used analytical tools, will allow one to produce meaningful data analysis in business research.

Keywords

Descriptive analysis, Inferential analysis, Hypothesis Testing, Estimation, Measures of Central Tendency, Measures of Dispersion, Relationship etc.

I. Introduction

Data Analysis is used in many industries to allow companies and organization to make better business decisions. Data Analysis is one of the many steps that must be completed when conducting a research but it assumes special significance. Data when collected from various primary and secondary sources is in its raw form are incredibly useful but also overwhelming. It is almost impossible for researcher to deal with all this data in its raw form. Through data analysis such data is presented in a suitable and summarized form without any loss of relevant information so that it can be efficiently used for decision making. Data can be presented in the tabular or graphic form. The tabular form (tables) implies numerical presentation of data. The graphical form (figure) involves the presentation of data in terms of structure which can be visually interpreted, e.g., Bar charts, Pie charts, Histograms, Line charts etc.

Charts, graphs and textual write ups of data are all form of data analysis. These methods are designed to refine and distill the data so that readers can have clean interesting information without needing to sort through all of the data on their own. It has to be noted that research data analysis provides the crucial link between research data and information that is needed to address research questions. Data Analysis has multiple facets and approaches, encompassing diverse statistical techniques, under a variety of names in different businesses, science and social science domains.

Analysis of data means studying the tabulated material in order to determine inherent facts or meaning. A plan of analysis can and should be prepared in advance before the actual collection of material.

Processing and analysis of data is always found to be interwoven. Many experts are of the view that analysis of data is different from processing of data. Prof. John Gatting had made distinction between analysis of data and processing of data. According to him processing of data refers to concentrating, recasting and dealing with the data so that they are as amenable to analysis as possible, while analysis of data refers to seeing the data in the light of hypothesis of research questions and the prevailing theories and drawing conclusions that are as amenable to theory formation as possible. (Gupta 2010). But there are experts who do not like to make difference between processing and analysis. Technically speaking,

processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis.

II. Objectives

The main objective of this paper is to provide a detailed summary of data analysis, and its uses in understanding the concept of data analysis for behavioural research.

III. Methodology

For this study, data is collected from Secondary sources and available literature has been reviewed and analyzed for understanding the concept and use of data analysis for behavioral research.

IV. Discussion On Data Analysis

A. Definition of Data Analysis

The term analysis refers to the computation of certain measures (such as measures of central tendency, variation etc) along with searching of patterns of relationship (such as correlation, regression) exist among data groups. Apart from that, in the process of analysis, relationships or differences, supporting or conflicting with original or new hypothesis should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions. Analysis, therefore, be categorized as *descriptive* and *inferential analysis* (inferential analysis is also known as *statistical analysis*). Descriptive analysis deals with computation of certain indices from raw data with and establishing relation between two or more variables. Whereas, inferential analysis is concerned with the: (a) the estimation of population parameters, and (b) the testing of statistical hypothesis or test of significance.

As Prof. Wilkinson & Bhandarkar quoted "Analysis of data involves a number of closely related operations that are performed with the purpose of summarizing the collected data and organizing these in such a manner that they will yield answer to the research questions or suggest hypothesis or questions if no such questions or hypothesis had initiated the study." (Mohan & Elangovan 2011) In general, data analysis is the science of examining raw data with the purpose of drawing conclusions about the information.

B. Goals of Data Analysis:

1. Giving a feel to research data: After data collection, the first step towards understanding the huge mass of data has been gathered, is to arrange the materials in a concise and logical order. The procedure is referred to as classification and tabulation of data. However, these forms of presentation may not be very interesting to the common man. Too many figures are often confusing and may fail to convey the message effectively for which they are meant. Hence, another important convincing and easily understood method of presenting the data is the use of graphs and diagrams. Constructing tables and graphs for the concerned data is a major part of analysis, which will facilitate in better understanding and comparison of data.

2. Identifying average values and variability: Most of the research studies result in a large volume of raw data which must be suitable reduced so that the same can be read easily and can be used for further analysis. One of the most important goals of data analysis is to get one single value that describes the characteristic of the entire mass of unwieldy data. Such a value is called central value or average values. The most important averages are mean, median and mode. The various measures of average values alone cannot adequately describe a set of observations, unless all the observations are same. To identify the measurement of scatteredness or variability of the mass of data in a series from the average is equally important to describe the data.

3. Identifying relation between variables: One of the ways that one can get better insights into the data is by discovering that variables are related to each other i.e. with increase in one variable there is an increase in other and vice versa. Also effort is made to know the cause and effect relation between two or more than two variables.

4. To make inferences about population parameter: In most of the research studies, it is not possible to enumerate whole population in the study. Hence, a part of the population i.e. sample is taken to consider for the study. One of the goals of data analysis of these samples is to use information contained in sample of observation (such as sample mean, sample standard distribution) for drawing conclusion or making inference about the larger population (such as population mean, standard deviation etc).

5. To test the hypothesis: A statistical hypothesis is an assumption about any aspect of a population. For e.g., there is no relationship between compensation and job satisfaction (i.e. null hypothesis) Analysis of data is carried out to test a hypothesis on the basis of sample values, so that hypothesis can be accepted or rejected. Ultimate decisions are taken on the basis of the collected information and the result of the test.

C. Types of Analysis:

As mentioned earlier, in section 1, statistical analysis can be categorized into descriptive and inferential analysis.

1. Descriptive analysis: is mostly concerned with computation of certain indices or measures from the new data. Zikmund has quoted "...with descriptive analysis, the raw data is transformed into a form that will make them easy to understand & interpret." It is largely the study of distributions of one variable. This sort of analysis can be analyzed data in three different ways:

- ❖ **Univariate analysis:** When a single variable is analyzed alone, e.g., statistic 'such as "mean" which might refer to age group of students, it is known as univariate analysis.
- ❖ **Bivariate analysis:** When some association is measured

between two variables simultaneously, e.g., cross-classification of age group, it is known as bivariate analysis.

- ❖ **Multivariate analysis:** In multivariate analysis, three or more variables are investigated simultaneously, allowing us to consider the effects of more than one variable at the same time. For example, identifying job satisfaction in terms of age, sex, salary and so on. Multivariate analysis includes techniques like multiple regression analysis, multiple discriminant analysis, multivariate analysis of variance (MANOVA), factor analysis and canonical analysis. Some of these terms are briefly described in upcoming sections.

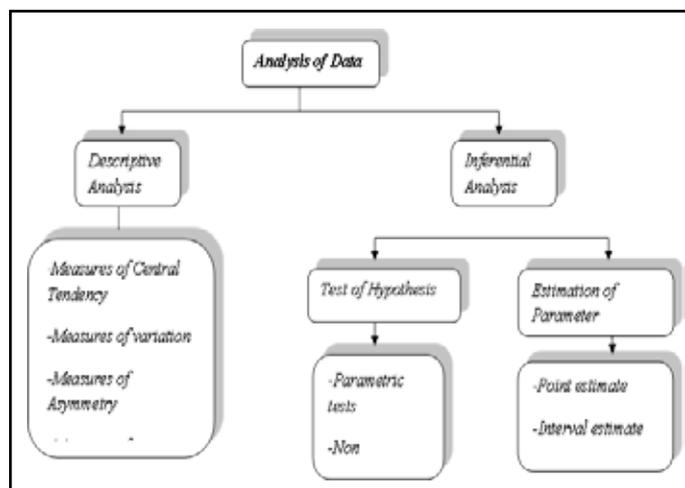
The important statistical measures that are used in descriptive analysis are:

a) Measures of Central Tendency or location or average values : The simplest type of statistical analysis of a data containing a set of observations, is the calculation of a single value which could be taken as the representative of data. There are various measures for arriving at this value, and are known as measure of central tendency or location or average values. These measures indicate a value where all the observations can be assumed to be located or concentrated. They are important to use further higher statistical calculations. There are three such measures:

i. Mean: There are three types of means:

Arithmetic mean: (typically referred to as mean) is the most common measure of central tendency. It is considered highly valuable since it considers the whole data set and gives equal importance to all. But there is one exclusive case in arithmetic mean i.e. **weighted arithmetic mean**, where, unequal weight or importance is given to all observations. We calculate the arithmetic mean by adding together all the values in the data set and then dividing that sum by the number of values in the data.

Chart 1. Types of Analysis of Data



Geometric mean: In many cases giving equal importance to all observations may lead to misleading answers. One of the measures of location can be used in such cases is geometric mean. The Geometric Mean(G.M.) of a series of observations is defined as the n^{th} root of the product. It is to be noticed that if any of observation is zero calculating G.M. is not possible since the product of various values becomes zero.

Harmonic Mean: The Harmonic Mean (H.M.) is defined as the reciprocal of the arithmetic mean of the reciprocals of the observations. The mean is used in averaging rates when the time factor is variable and the act being performed is the same, such as, for calculating average speed of car H.M. is used. The main

limitation of H.M. is it cannot be calculated if any value is zero.

ii. Median: There are situations in which data set have extreme values at lower or higher end, termed as **outliers** in statistical language. In such cases arithmetic mean is desirable to use, since it easily gets affected by those extreme values. For example, if the data is such as 2, 3, 5, 2, 22, the mean will be 16.4 which cannot be considered as good representative of data. Hence, another measure of location i.e. median is used in such cases. Further, whenever, exact values of some observations are not available median is used.

Median is the *point* that divides a distribution of scores in two equal parts, one part comprising all values greater and the other all values less than median. To be remembered, the median is a hypothetical *point* in the distribution; it may or may not be an actual score.

iii. Mode: The third central tendency statistic is the mode. Mode is defined as the 'most fashionable' value, which observation is most frequently occurring in a set of data. For example, a data series is as 2, 3, 4, 2, 2, 6 and 9, the mode is 2 because 3 observations have this value. Mode is frequently used in cases where complete data are not available, as well as, when the data is in quantitative form where only getting a data regarding presence/ absence of the observation is possible.

b). Measures of variation or dispersion: In addition to central tendency, every data can be characterized by its variation and shape. In two or more data sets central tendency may be the same but there can be wide disparities in the formation of the set. Variation measures the dispersion, or disparities, of values in data set. Dispersion may be defined as statistical summaries throwing light on the differences of items from one another or from an average. Most commonly used in statistics are the standard deviation and variance, but there are many other, discussed below:

i. Range: The range is the simplest measure of variation in a set of data, and is defined as the difference between the maximum and minimum values of the observations. However, it only depends on minimum and maximum values, and does not utilize the full information in the data, it is not considered very reliable.

ii. Semi Inter-Quartile Range or Quartile Deviation: Quartiles split a set of data into four equal parts- the first Quartile Q_1 , carries 25 % of the data set values; the second Quartile Q_2 , carries 50% of the data set values; the third Quartile Q_3 , carries 75% of the data set values. The interquartile range (also called midspread) is the difference between third and first quartiles in a data set i.e., $Q_3 - Q_1$. The interquartile range measures the spread in the middle 50% of the data.

However, a much more popular measure of variation is Semi Inter-Quartile Range or Quartile Deviation, and is defined as $Q_3 - Q_1 / 2$.

iii. Mean or Average Deviation: Mean Deviation is defined as the average of difference of individual items from some average of the series, can be mean, median or mode. Such difference of individual items from some average value is termed as deviation. While calculating mean deviation all deviations are treated as positive ignoring the actual sign.

iv. Variance and Standard Deviation: In mean deviation negative sign is ignored, otherwise the total deviation comes out to be zero, since similar values with opposite signs will cancel each other. However, another way of getting over this problem of total deviation being zero is to take the squares of deviations of the observations from the mean. The sum of squares of deviation divided by number of observations is known as variance and

its square root is known as the standard deviation. Karl Pearson introduced these terms.

c) Measure of Asymmetry (skewness): A distribution of data values is either symmetrical or skewed. In symmetrical distribution, the values below the mean are distributed exactly as the values above the mean. In this case, the low and high values balance each other out. In a skewed distribution, the values are not symmetrical around the mean. This results in an imbalance of low or high values. Skewness is a measure of asymmetry in data. The data can be **negatively (or left)** skewed or **positively (or right)** skewed. In left skewed, most of the values are in the upper portion of the distribution, whereas, in right skewed, most of the values are in the lower portion of the distribution.

If the distribution is skewed, the extent of skewness could be measure by, **Bowley's Coefficient of Skewness** or **Pearson's Measure of Skewness**.

Kurtosis is an *indicator of peakedness of a distribution*. Kark Pearson called it "Convexity of Curve". A bell shaped or normal curve is Mesokurtic, whereas more peaked curve than the normal curve is Leptokurtic and a curve more flat than the normal curve is Platokurtic.

Knowing the shape of distribution is necessary since some assumptions about their shape is made for the use of certain statistical methods.

d) Measures of Relationship

Very often, researchers are interested to study the relationship between two or more variables, which is done with the help of **correlation** and **regression analysis**. The ideas identified by the terms *correlation* and *regression* were developed by Sir Francis Galton in England.

Correlation is a statistical technique that describes the degree of relationship between two variables in which with the change in value of one variable, the value of other variable also changes. The degree of correlation between two variables is called **simple correlation**. The degree of correlation between one variable and several other variables is called **multiple correlation**.

The simplest and yet probably the most useful graphical technique for displaying the relationship between two variables is scatter diagram (also called scatter plot). Here, the data for two variables are plotted on x and y axis of graph. If the points are scattered around a straight line, the correlation is **linear** and if the points are scattered around a curve, the correlation is **non-linear (curvi-linear)**.

The scatter plot gives a rough indication of the nature and strength of relationship between two variables, The quantitative measurement of the degree/extent of correlation between two variables, is performed by **coefficient of correlation**. It was developed by Karl Pearson, the great biologist and statistician, hence referred as "Pearsonian Correlation Coefficient" (also known as Product moment correlation coefficient) It is denoted by greek letter ρ (**rho**), when calculated from population values, 'r' when calculated from sample values. The value of coefficient of correlation varies between two limits +1 and -1. The value +1 shows perfect positive relationship between variables, -1 shows perfect negative correlation and 0 indicates zero correlation. If the relationship between two variables is such that with an increase in the value of one, value of other increases or decreases, in a fixed proportion, correlation between the variables is said to be **perfect**. Similarly, **perfect positive correlation** means increase in one variable bring increase in other, in same proportion and vice versa. **Perfect negative** means increase in one variable decreases

the other variable in same proportion. **Zero correlation** shows there is no linear relationship between two variables. It is to be noted that 'r' indicates the extent of only linear relationship. Zero value only indicates there is no linear relationship, but there could be other type of non-linear relationship.

Above discussed, Pearsonian Correlation Coefficient is applicable only when data is in **interval or ratio form** i.e. quantitative measurement of variables such as height, weight, temperature, and income is possible. In some cases such as beauty, honesty or in similar cases where data is only available in **ordinal or rank form**. Karl Pearson's formula of correlation coefficient is not possible. Hence, Charles Edward Spearman in 1904 developed a measure called '**Spearman Rank Correlation**' to measure correlation between ranks of two variables. It is denoted as r_s . Value of correlation coefficient also ranges between +1 and -1. Spearman rank correlation is said to be a **non-parametric or distribution free** method, since it doesn't fulfil the assumption of normal distribution for both variables. One similar kind of method used for getting association between ranks of variables is **Kendall Tau rank correlation**.

When one or both of the variables is in categorical form i.e., not measurable but only on the basis of their presence or absence in each case it is possible to know their frequency or total number of occurrences, the data said to be on **nominal scale**. In such cases, to know the association between two attributes '**coefficient of contingency**' or '**coefficient of mean square contingency**' introduced by Karl Pearson is used.

Correlation analysis deals with exploring the correlation between two or more variables. Whereas, regression analysis attempts to establish the nature of relationship between variables, that is, to study the functional relationship between the variables and thereby provide a mechanism for **predicting**, or forecasting. Example, correlations tells there is a strong relation between advertisement and sales. Regression will predict this much increase in advertisement will give this much of increase in sales. Regression analysis can be of two types- simple (deals with two variables) and multiple (deals with more than two variables). If the relationship between two variables, one independent or predictor or explanatory variable and other dependent or explained variables, is a linear function or a straight line, then the linear function is called simple **regression equation**, and the straight line is known as **regression line**. It is a "line of best fit" i.e. the line on which the difference between the actual and estimated values will be minimum. The simple regression equation is used to make predictions.

$y = a + bx$ OR $x = a + by$ are two possible regression equations in case of two variables involved in regression analysis. First equation said as **regression equation of y on x** and so on. In first equation y and in second equation x is dependent variable, whereas x in first equation and y in second equation is independent variable. Here, 'a' and 'b' are constants, 'a' is **intercept** and 'b' is **slope** or **inclination** or most popularly known as **regression coefficient**. Regression coefficient gives the change in dependent variable when independent variable changes by 1 unit. To estimate the relationship between x and y it is vital to determine 'a' and 'b' respectively. This is done through the **Principle of Least Squares Method**. Apart from that the Principles of Least Squares provide criterion to select "line of best fit" mentioned in the last paragraph.

In case, **curved relationship** is found between variables, then correlation ratio eta (η) gives the degrees of its association. It

may be noted that as many methods discussed until now involved only two variables i.e., simple correlation analysis and regression analysis. However, very often, one is required to study the relation between more than two variables, impact of several independent variables, jointly together, on dependent variable. This is possible through **multiple correlation and multiple regression analysis** respectively. Here, multiple correlation coefficients are obtained which indicate the relation between one dependent variable and several independent variables by using multiple regression equations

When correlation between any two variables is analyzed where the effect of the third variable on these two variables is held constant or removed, then such analysis is known as **partial correlation analysis** and such correlation is termed as **partial correlation coefficient**. Similarly, **partial regression coefficient** is the value indicates the change that will be caused in dependent variable with a unit change in independent variable when other independent variables held constant. But, as a matter of fact multiple correlation coefficients and multiple regression coefficients are applicable only in case of ratio or interval data. In case of **ordinal data** such correlation can be enumerated by **Kendall partial rank correlation** & in case of **nominal data** **discriminant analysis** is used. (See Table 1)

Table 1: Choice of relationship analysis tool based on number of variables and scale of measurement

For two variables(i.e. simple correlation) For interval or ratio data For ordinal data For nominal data	Pearson product moment correlation coefficient Spearman rank order correlation coefficient or Kendall Tau rank correlation Contingency coefficient
For more than two variables(i.e. multiple correlation) For interval or ratio data For ordinal data For nominal data	Multiple regression analysis Kendall partial rank correlation Discriminant analysis NA

Source: Compiled by Authors

2. Inferential Analysis: Inferential analysis is mainly concerned with (a) estimation of population values such as population mean, population standard deviation, and (b) various tests of significance/ testing of hypothesis. Inferential analysis plays a major role in statistics since mostly it is not possible to go for whole population while conducting researches, hence, a sample is chosen and using inferential analysis the sample values obtained are used to infer about the population. The objective of inferential analysis is to use the information contained in a small sample of observations for drawing a conclusion or making an inference about the larger population. Such inference may be in the form of estimation or Testing of Hypothesis or assumptions. For example, either one could estimate population parameter based on sample statistic, like ' mean life of a car battery ', or one could test the claim of company that 'mean life of car battery is 3 years'. In both the cases an inference about population is made. There are various

methods of estimation and testing of hypothesis.

a) Estimation: It deals with the estimation of parameters such as population mean based on the sample values. The method or rule of estimation is called an *estimator* like sample mean, the value which the method or rule gives in a particular case is *estimate* of population parameter. In other words, estimator is a function of sample values to estimate a parameter of population. With the help of samples of observation, an estimate in the form of a specific number like 25 years can be given or in the form of an interval 23-27 years. In the former case it is referred as *point estimate*, whereas in the latter case it is termed as *interval estimate*.

i. Point estimate: It is used to estimate a population parameter, with the help of sample of observations. A point estimate is a single value, say 50. This number is taken as the best value of unknown population parameter. An estimator is said to be efficient, if it has minimum variance such as sample arithmetic mean.

There are several methods of estimating the parameters of a distribution, such as, maximum likelihood, least squares, methods of squares and minimum chi-square.

ii. Interval estimate: Point estimate gives a single value, taken as best estimate of parameter. However, if another data is collected from same population, the point estimate may change. In real life situation population parameter may not be exactly equal to sample statistic, and could be around this value. Thus it may be more logical to assume that the population values lies in an interval containing the sample, such as 48-52, known as interval estimate. It is expected that the true value of population will fall within this interval with the desired level of confidence; hence the name 'confidence interval' is given.

The interval should be in reasonable limits. These limits are statistically calculated. The limits or intervals, so arrived, are referred to as *confidence intervals* or *confidence limits*. Since we are estimating population parameter from sample values, we can never make any estimation with 100% confidence. Desired confidence for estimation is termed as *confidence interval*. Usually, 95% level of confidence is considered adequate. One can state as 'with 95% confidence can say that population parameter will fall somewhere between confidence interval of 40-50'.

b). Testing of Hypothesis/Test of Significance: In most of the cases, it is almost impossible to get knowledge about population parameter, therefore, hypothesis testing or test of significance is the often used strategy for deciding whether sample offers such support for a hypothesis or assumptions that generalizations about population can be made. In other words, test can find the probability that a sample statistic would differ from a parameter or another sample .

Hypothesis testing typically begins with some assumptions or hypothesis or claim about a particular parameter of a population. It could be the parameters of a distribution like mean, describing the population; the parameters of two or more population, correlations or associations between two or more characteristics of a population. Hypothesis can be of two types, null and alternative hypothesis. Null hypothesis is considered to be a hypothesis of "no relationship". Such as 'there is *no significant* difference between sample means'. The term Null hypothesis is said to have been introduced by R. A. Fisher. The word Null is used because the nature of testing is that we try our best to nullify or reject this hypothesis based on sample collected. When null hypothesis is rejected the opposite of null hypothesis i.e. *alternative hypothesis* is automatically accepted. Alternative hypothesis is the statement which is intended to be accepted if the null hypothesis is rejected.

Null hypothesis is denoted as H_0 and alternative hypothesis is denoted as H_a . It has to be kept in mind that, we cannot prove a hypothesis to be true. We may find the evidence that supports the hypothesis. Suppose, we have failed to reject the null hypothesis, doesn't mean null hypothesis have been proven to be true, because the decision is only made on the basis of sample information.

Once the null and alternative hypothesis has been set up, the next step is to decide on the level of significance. It is used as a criterion for rejecting the null hypothesis. It is expressed as a percentage like 5% or 1%, or sometimes as 0.05 or 0.01. It is that level, at which we are likely to reject null hypothesis even if it is true. Now decision on the appropriate statistic such as t, z, f etc is taken. Based on the level of significance critical or tabulated value is found. After calculating the statistic from the given sample of observation, the test statistic is compared with the critical value. If calculated value (statistic) is equal to or less than critical value, the difference between result and expected value is insignificant and this insignificant difference can be subjected to sampling error, hence, null hypothesis is accepted. Whereas, if calculated value is higher than critical value, the difference is said to be significant, and can't be subjected to sampling error, therefore null hypothesis is rejected.

Whenever we take a decision about population based on sample, the decision cannot be 100% reliable. The possibilities can be, we would reject null hypothesis even if it is true, termed as *Type I error, denoted as α* or we could accept the null hypothesis even if it is false, termed as *Type II error, denoted as β* .

Type I error is also referred as *level of significance*, as discussed above. The quantity $1 - \beta$ is called the '*power*' of test, signifying the test ability to reject null hypothesis when it is false, and $1 - \alpha$ is called *confidence coefficient*.

Various tests of significance have been developed to meet various types of requirements. They may be heavily classified into, parametric and non-parametric tests. *Parametric tests* are based on the assumptions that the observations are drawn from a normal distribution. Since the testing procedure requires assumptions about the type of population or parameters values these tests are known as 'parametric tests'. The test of significance developed for situations when this condition is not satisfied, known as '*non-parametric tests*' or '*distribution-free tests*'. As a matter of fact, parametric tests are more powerful test than non-parametric tests.

Various parametric and non-parametric tests of significance performing different functions in different conditions is mentioned below in a tabular form:

Table 2: Choice of parametric/non-parametric test based on function to perform & scale of measurement

Function	Non-Parametric Tests (Interval/Ratio Data)	Parametric tests (Ordinal/ Nominal Data)
Test of Significance of one sample test	't' (mean known, S.D. * unknown) 'z' (mean known, S.D. known)	Sign test

Test of Significance for difference between two independent sample	Independent samples 't' test (S.D. unknown) 'z' test (S.D. known)	Kolmogorov-Smirnov two sample test or Mann Whitney U Test or Wilcoxon Sum of Rank Test (<u>Ordinal Data</u>) & Chi-Square test (<u>Nominal Data</u>)
Test of Significance for difference between two paired samples (series of samples taken from same population)	Paired 't' test	Wilcoxon matched-pairs signed ranks test (<u>ordinal data</u>) & Mc Nemar test for the significance of changes. (<u>nominal data</u>)
Test of Significance for difference between series of independent samples	'F' Test & One way ANOVA (Analysis of Variance)	Kruskal-Wallis Rank Sum Test – H Test or Wilcoxon-Wilcox multiple comparison test (<u>ordinal data</u>) & chi-square for k independent samples (<u>nominal data</u>)
Test of Significance for difference between series of paired independent samples	Repeated measures ANOVA (Analysis of Variance)	Freidman two-way analysis of variance (<u>ordinal data</u>) & Cochran Q Test (<u>Nominal data</u>)

Source: Compiled by Authors

(*S.D. is standard deviation of population)

Apart from above functions Chi-square test, denoted as χ^2 used as test of goodness of fit i.e. how well observed values fits with of expected values and test of independence i.e. it tests existing association between two categorical variables. Non-parametric alternative for χ^2 test as goodness of fit is Kolmogorov-Smirnov test. Apart from that, 't' test is used for testing significance of correlation & regression coefficient or slope; 'F' test & ANOVA is used for testing significance of multiple regression coefficient.

V. Conclusions

For any successful study, analysis of data is one of the most crucial step. It is always advisable that it should be designed before the data are actually collected. Otherwise there is always danger of being too late and the chances of missing out relevant facts. There are a number of analytical tools can be used for summarizing the data and inferring about the population based on sample values. But to use any tools some of the assumptions have to be fulfilled, therefore, this assumptions always have to be kept in mind by the researcher before applying any analytical tool. That is why it is said that analysis requires a lot of experience and knowledge in the field of data analysis.

References

- [1] Bhattacharyya, D. K. (2006). *Research Methodology*, New Delhi, Excel Books, pp. 127- 166
- [2] Dunn, O.J. & Clark, V.A. (2009). *Basic Statistics-A Primer for the Biomedical Sciences*, New Jersey, John Wiley & Sons, pp. 50-58 & 189-198.
- [3] Gupta, S. (2010). *Research Methodology*, New Delhi, Deep & Deep Publications Pvt. Ltd, pp. 192-209
- [4] Kumari, A. (2008). *An Introduction to Research Methodology*, Udaipur, Agrotech Publishing Academy, pp. 79-112.
- [5] Kumar, C.R. (2008). *Research Methodology*, New Delhi, APH Publishing Corporation, pp. 113-131
- [6] Kothari, C.R. (2009). *Research Methodology-Methods & Techniques*, New Delhi, New Age International (P) Limited, pp. 122-149.
- [7] Levine, D. M., Krehbiel, T.C., Berenson, M.L. & Vishwanathan, P.K. (2011). *Business Statistics*, New Delhi, Dorling Kindersley (India) Pvt. Ltd., pp. 82-117 & 290-310
- [8] Mohan, S. (2011). *Analysis & Interpretation of Data*. Pages 145- 163 in *Research Methodology in Commerce*, New Delhi, Deep & Deep Publications Pvt. Ltd.
- [9] Shrivastava, T.M., Rego, S. (2010). *Statistics for Management*, New Delhi, Tata Mc Graw Hill Pvt. Ltd., pp. 4.1-4.56 & 10.3-10.80
- [10] Spatz, C. (2008). *Basic Statistics-Tales of Distribution*, USA, Thomson Wadsworth, pp. 41-55 & 86-97
- [11] Taylor, B., Sinha, G. & Ghosal, T. (2007). *Research Methodology- A Guide for Research in Management and Science*, New Delhi, Prentice Hall of India, pp. 135-151.
- [12] Verma, R. K. & Verma, G. (2010). *Research Methodology*, New Delhi, Commonwealth Publishers, pp. 107-118.

Author's Profile

Niharika Singh is a research scholar (SRF) in the department of Management, Mizoram University, Aizawl. She is pursuing her Ph.D. in the area of employee retention. She has published 1 article in international journal and contributed 3 chapters for edited books. Miss Singh has presented 6 papers and participated in 8 national and international seminars.

Dr. Amit Kumar Singh is assistant professor in the department of Management, Mizoram University, Aizawl. He got his education from BHU, Varanasi, VBS PU Jaunpur and IIM Ahmedabad. He has published 2 books and more than 35 research papers in International/National journals and edited books and also presented/delivered lecture in more than 50 seminar/workshop/conferences. He also served as member of different academic bodies like BOS, School Board, Academic Council (as Head i/c), etc. Currently he is in the advisory board of 12 International journals. Dr. Singh is also working as Japan International Co-operation Agency (JICA) service consultant for department of Forest, Government of Mizoram.