ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

DESCRIPTIVE STATISTICS IN BUSINESS RESEARCH

Carolyn Vanlalhriati & E Nixon Singh
1.Research Scholar, Department of Management, Mizoram University, Mizoram, India
2.Professor & Head, Department of Management, Mizoram University, Mizoram, India

Manuscript Info

Manuscript History:

Received: 12 April 2015
Final Accepted: 22 May 2015
Published Online: June 2015

Key words:

Descriptive statistics, Measures of Central Tendency, Measures of Dispersion, Measures of Asymmetry, Measures of Relationship

Corresponding Author

Abstract

The analysis of data is the most skilled task in the business research process which requires the researcher own judgment and skill. The different statistical techniques were available to enrich the researcher decision. Choices of appropriate statistical techniques were determined to a great extent by the research design, hypothesis and the kind of data that was collected. These techniques were categorized into descriptive and inferential statistics. This paper focuses only on descriptive statistics which summarizes large mass of data into understandable and meaningful form.

INTRODUCTION

Statistics is concerned with the scientific method by which information is collected, organized, analyzed and interpreted for the purpose of description and decision making. A wider scope and comprehensive definition was framed by two well known statistician Croxton and Cowden "Statistics may be defined as the science of collection, presentation and analysis and interpretation of numerical data". It is a sound techniques or method for handling the collected data, analyzing the data and used for drawing valid inferences from them.

There are different statistical approaches available to a researcher. Choices of appropriate statistical techniques are determined to a great extent by the research design, hypothesis and the kind of data that will be collected. When the data are collected, edited, classified and tabulated, they are analyzed and interpreted with the help of various statistical tools based on the nature of investigation. Thus, the researcher is expected to have basic knowledge of statistics for carrying out the systematic analysis as well to provide accurate and precise interpretation of data.

OBJECTIVE OF THE STUDY

To study the different types of descriptive statistics which are used for describing the data in business research.

METHODOLOGY

The study was entirely based on secondary data. The required data for the present study were collected from secondary source such as books.

TYPES OF STATISTICAL METHODS

The statistical methods adopted for such analysis falls into two categories

1. Descriptive statistics, and

2. Inferential statistics.

Descriptive Statistics focuses on summarizing and describing the characteristics of a data set while no attempt is made to analyze and interpret the data. The methods of descriptive statistics include graphic methods such as bar charts, line graphs and pie charts as well as numeric measures which includes measures of central tendency, dispersion, skewness and kurtosis. **Inferential Statistics** on the other hand consist of methods which are used to make inferences about population characteristics on the basis of sample results. They are also known as sampling statistics which is concerned with the process of generalization. It is further categorized as parametric or non-parametric. Parametric statistics is based on the assumption that the population is normally distributed while non-parametric statistics based on which data are collected on a nominal or ordinal scale. This paper focuses on the descriptive statistics for which the different methods of descriptive statistics are summarized as follows.

TYPES OF DESRIPTIVE STATISTICS

The process of data analysis begins with the application of descriptive statistics. Descriptive statistics provides simple summaries of the sample rather than learning the population characteristics from which the sample was ought to represent. It further permits the researcher to significantly describe many pieces of data with a few indices. The major types of descriptive statistics are measures of central tendency, measures of dispersion, measures of asymmetry and measures of relationship.

1. Measures of Central Tendency

The term 'Central Tendency' refers to the middle point of all observations. The observations for most of the data set shows a distinct tendency to cluster on a value which falls somewhere in the middle value of observations which is known as the central tendency. It provides a single value to represent the average characteristics of its distribution. The methods which are employed for measuring the values are known as measures of central tendency which is also known as statistical average.

Measures of central tendency are useful for reducing the complexity of the data. They summarize the entire data set into single representative value known as the average. The 'average' facilitates comparison between two or more data sets since it represents the entire data set of distribution. There were different types of measures of central tendency and all of these measures focus on measuring the central location which serve the purpose of summing up the entire data into a single representative figure. The most popular measures of central tendency are as follows;

- The Arithmetic Mean
- The Median
- The Mode
- The geometric Mean
- The harmonic Mean

1.1The Arithmetic Mean

The most popular measure of central tendency is the Arithmetic mean which is also known as arithmetic average. It can be defined as the value (figure) which we obtained by dividing the total values of observation by the numbers of observations. For example, the mean of a series 3,5,7,9 is 24/4 = 6. It can be work out by using the given formula:

For Ungrouped data, $\bar{X}= \frac{\sum X_i}{n} = \frac{X_{1+}X_{2+}\ldots\ldots+X_n}{n}$

For Grouped data, $\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{f_i X_{i+} f_2 X_{2+\cdots..}+f_n X_n}{f_{1+} f_{2+\cdots+} f_n}$

Weighted arithmetic mean can also be calculated, $\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$

Where, X_i= Value of the ith item X,i= 1,2,….,n

n= total number of items

= Weighted item

= weight of ith item X

= value of the ith item X

The arithmetic mean is a stable measure of central tendency. Unlike other measures, all the data are taken into considerations for finding the perfect representative figure for the data set. Every data set can have one and only mean which makes it unique from other measures. It is the only common measures in which all the values play an equal role. Therefore, it can be greatly affected by any value which has a great difference from other values. The presence of e xtreme values should be avoided to get a perfect mean for a given data set. Otherwise, due to the presence of extreme values a mean can be considered as a poor measure of central tendency

1.2 The Median

The median refers to the value of the middlemost or most central item in a distribution which divides the distribution into two equal parts. Half part comprising all values greater than or equal to the median and the other part comprised of all the values smaller or equal to the median. For calculating a median, it is necessary to arrange all the values in either ascending or descending order of magnitude.

For calculating a median from ungrouped data, if the numbers of observation are odd, the middle item of the array +is the median. On the other hand, if the numbers of observations are even, the two middle observations should be taken and the average (arithmetic mean) of the two selected items will be the median.

$$\text{Median} = \frac{\frac{n}{2}th\ observation + (\frac{n}{2}+1)^{th}}{2}$$

For calculating median from grouped data, the formula is:

$$\text{Median} = \text{l} + \frac{\left(\frac{n}{2}\right)-cf}{f}\,\text{h}$$

where, l= lower class limit of the median class interval

cf = cumulative frequency

f = frequency of the median class

w = width of the median class interval

n = total number of observations in the distribution.

The median is used in the context of qualitative phenomena. Unlike the arithmetic mean, extreme values do not affect the median.

1.3 The Mode

The mode is the observations which occur most frequently in the data set. In order to calculate the mode, one has to count the number of times various values are occurring in the data set. Like the median and unlike the mean, the mode is not affected by extreme values. For example, the mode of the given series 13, 21, 16, 13, 17, 25, 13, 10, 14, 25, 22, 13, 11, 18, 13 is 13. From this example, we can see that the value 13 is occurring maximum number of times i.e., 5 times in the data set and therefore, the mode is 13.

The mode can be calculated from the grouped data by:

$$\text{Mode}= \text{l} + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} Xh$$

Where, l= lower limit of the model class interval

f_{m-1}= frequency of the class preceeding the mode class interval

f_{m+1} =frequency of the class following the mode class interval

h = width of the mode class interval

1.4 The Geometric Mean

A measure which is employed for measuring the rate of change of variables over time is known as the geometric mean. Sometimes arithmetic mean is inappropriate for calculating index numbers such as average rate of change, average growth rate of population, etc whose quantities is changing over a period of time. In such cases, geometric mean must be one of the best averages to which can offer the perfect mean for the data set. Symbolically,

$$\text{G}= \sqrt[n]{x_1 . x_2 . x_3 \dots \dots . x_n}$$

Where, G= Geometric mean

$x_{1,} x_2 .. x_3$ = values of 'n' items

'n'= number of items

1.5 The Harmonic Mean

The reciprocal of the arithmetic mean of reciprocals of the values of its item in the data set is known as the harmonic mean. It should be preferred for calculating the average rate, average speed and average price, etc. Symbolically,

$$\text{H} = \text{Reciprocals of } \frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}{n}$$

$$= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}} = \frac{n}{\Sigma(\frac{1}{x})}$$

2. Measures of Dispersion

The measures of central tendency focus on measuring the average of a data set and represent it by a single value. These measures are one of the simplest measurements for finding out the middle point of distribution. However, they are not sufficient to fully explore the characteristics of the entire distribution. For example, two sales distributions may have an identical mean but they may have a different variability or dispersion. One may concentrate (cluster) somewhere around the middle point while the other may have a widely spread distribution. Such spread may differ from one distribution to another. In this case, measuring the variability of the distribution is required to explore how much the values in the data set are dispersed (spread or scattered).

Measures of dispersion are necessary to test the reliability of an average. The less variability among the values in the data set proved the reliability of an average by showing high uniformity of values in the distribution while the high variability shows the unreliability of average in the distribution. Moreover, they help to make out the nature and causes of variations which is essential for controlling the variations. Furthermore, they assist the use of other statistical tools such as correlation and regression, forecasting, hypothesis testing and so on. Some of the important measures of dispersion are

- The range
- Interquartile Range
- Mean Deviation
- The Variance and Standard Deviation

2.1 The Range

The range is the difference between the maximum and the minimum values observed in the distribution. It is based on the location of the maximum and minimum scores in the data set. It is denoted by R.

Range®= $(Maximum\ values\ of\ an\ item\ in\ the\ data\ set - Minimum\ Values\ of\ an\ item\ in\ the\ data\ set)$

For example, in a data set 2, 4, 6, 8, 10, 12. The range is the difference between maximum values 12 and the minimum values 2 i.e., 12- 2 =10. In other words, the length of an interval which covers the maximum and minimum values in the data set is 10.

The range is easy to calculate and understand. It is quite useful for cases which intended to find out only the extent of extreme variation such as temperature and so on. On the other hand, being based only on two values- maximum and minimum values, all the other values were failed to consider which makes it unreliable measure for serious research studies.

2.2 Interquartile Range

Interquartile range is a measure of dispersion of values in a data set which gives the difference between the third quartile Q3 and the first quartile Q1 i.e., IQR = Q3-Q1. In order to compute this range, quartiles split the data set into four equal parts Q1,Q2,Q3 and Q4 each containing 25 percent of the observed value. Interquartile range measures the data spread in the middle 50 percent of the distribution. Symbolically,

Interquartile Range(IQR) = $Q3 - Q1$

The measure of half the difference between the third and the first quartile is also known as Semi-quartile range or quartile deviation.

Quartile Deviation (QD) = $\frac{Q3-Q1}{2}$

2.3 Mean Absolute Deviation or Average Deviation

The degree to which values within a data set deviate from the central values is known as mean deviation. By deviation, it refers to the difference between the values of item and average which is taken as a standard such as mean, median and mode which is a representatives of the distribution. In other words, the quantity which is obtained by deducting the average from each item in the distribution is known as the deviation. Even though all the central tendency measure can be used as a standard, mean is commonly used because of its mathematical properties. While calculating mean deviations, algebraic signs of deviation are ignored (viz. + and -). Symbolically,

Mean Deviation = $\frac{\sum|m-a|}{n}$ or $\frac{\sum d}{n}$

Where, m = the variables
a= mean
n= total number of values

2.4. The Variance and Standard Deviation

The problem of negative signs in mean deviation can be disregarded by squaring them. The square of the deviations from mean is computed instead of calculating the values of each deviation from mean. The sum of such squared deviation which is divided by the total number of values in a data set is known as the variance. Symbolically,

$$\text{Variance} = \frac{\sum(d)\,2}{n}$$

The Square root of such variance is known as the standard deviation. It is always computed from the arithmetic mean since the sum of squares of deviations is always the least if the deviations are taken from it. Symbolically,

$$\text{Standard Deviation} = \sqrt{\frac{\sum(d)\,2}{n}}$$

3. . Measures of Asymmetry (Skewness)

A distribution of values in a data set which is not symmetrical (normal) is called asymmetrical or skewed. Skewness specifies the direction of dispersion as well as the extent to which the items are concentrated around the mean value. It is a measure which clearly explains the shape of a distribution.

A distribution is considered to be symmetrical when the values of mean, median and mode are all same and the curve is presented in a perfectly bell shaped which is also describe as a normal curve. In such a normal curve, the skewness is absent but if the curve is distorted whether on the right or the left side presenting an asymmetrical distribution, it indicates the presence of skewness. The data can be either positively or negatively skewed. In a positively skewed distribution, the curve is distorted towards the right side expressing the values i.e z<m < while in a negatively skewed distribution we have z>m> with the curve throwing towards the left side. The absolute skewness can be measured by finding the difference between mean and mode. Symbolically,

$$Absolute\ sk\ =\ Mean - Mode$$

The absolute skewness is limited for measuring the values expressed in the same unit as the distribution. Therefore, in order to compare the skewness of two or more distributions having different units of measurement, the relative measures of skewness must be adopted such as;

Karl Pearson's coefficient of skewness is given by

$$Sk_p = \frac{Mean - Mode}{Standard\ Deviation}$$

Bowley's coefficient of skewness which is based on the relative positions of median and quartiles in a distribution and is given by:

$$Sk_b = \frac{(Q3 - Md) - (Md - Q1)}{(Q3 - Md) + (Md - Q1)} = \frac{Q3 + Q1 - 2Md}{Q3 - Q1}$$

Kelly's coefficient of skewness which is based on percentiles and deciles and is given by:

$$Sk_k = \frac{P10 + P90 - 2P50}{P90 - P10} = \frac{D1 + D9 - 2D5}{D9 - D1}$$

Kurtosis is the measure of flatness or peakedness in the mode region of a curve. It is a fourth device for describing the frequency distribution characteristics. Two distributions having the same average, dispersion, and skewness can still be differentiated by measuring the concentration of values near the mode region which is known as kurtosis. Karl Pearson introduced three types of kurtosis such as Leptokurtic Curve which is peaked in nature, mesokurtic curve which is neither flat-topped nor peak in nature and Platykurtic Curve which is a flat-toped curve having positive kurtosis.

4. Measures of Relationship

So far, the statistical tools describe earlier were focusing on data involving only on one variable i.e., univariate population. In case of bivariate (two variables) or multivariate (two or more variables) populations, analysis of data which involves two or more variable is often required to describe the extent of relationship between these variables. Such a degree of relationship between two or more variables can be measured by using statistical techniques such as correlation and regression.

4.1 Correlation

A statistical technique which measures degree of relationships between two or more variables is known as correlation analysis. According to Croxton and Cowden, " When the relationship is of quantitative in nature, the appropriate statistical tool for discovering and measuring the relationship and expressing in a brief formula is known as correlation". The correlation between two variables is represented by the letter r whose values should vary in between -1 and +1. When the value falls in +1 range in the scale, it is a perfectly positive relationship while -1 shows a perfectly negative relationship and 0 shows no relation. In a perfectly positive correlation, the increase in one variable causes corresponding increase in the other variable while perfectly negative correlation indicates the decrease in one variable ca uses the corresponding decrease in the other variable. There are several methods of applying correlation techniques but the most important ones are:

4.1.1. The Scatter Diagram

A scatter diagram is represented by a graph which helps the researcher in visualizing the relationship between two variables. It can be obtained by plotting the observed values of two variables x and y in a graph paper where the independent variable values lie on the x axis and dependent variable values on the y axis. A researcher must draw a line through data points enhancing equal number of points lie on either side of the line. This line is used to depicts the type of correlation existed among the variables x and y. The straight line represents the linear correlation while the curve line represents non-linear correlation.

4.1.2 Karl Pearson's correlation coefficient

Karl Pearson's correlation coefficient also known as simple correlation is one the most popular method for determining the extent of relationship between two variables. It is based on the following assumptions:

- For calculating Karl Pearson's Coefficient, both the variables x and y must be measured on an interval or a ratio scale.
- There is a linear relationship between these variables.
- The cause and effect relationship existed between the two variables influencing their pattern of distributions.

Karl Pearson's Correlation coefficient is given by

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n.\sigma_x\sigma_y}\text{where,}$$

x_i= ith value of x variable

$\bar{x}$= mean of x

y_i= ith value of y

$\bar{y}$= mean of Y

n =number of pairs of observations of x and y

σ_x=standard deviation of x

σ_y=standard deviation of y

4.1.3 Spearman's Rank Correlation Coefficient

British Psychologist Charles Edward Spearman developed this method in 1904 for determining the degrees of correlation between two variables which consist only of ordinal data. Ranks are given to the different values of variables either in ascending order or descending order to determine the similarities or dissimilarities of the two set of ranks. Symbolically,

$$r_s= 1 - \left\{\frac{6\sum d_i{}^2}{n(n^2-1)}\right\}$$

Where, r_s= Spearman's rank correlation coefficient

d_i= the difference between a pair of ranks

n = number of paired observations.

4.2. **Regression**

A statistical technique which determines the functional (or algebraic) relationship between two variables is known as regression. It is presented in the form of an algebraic equation whereby the value of one variable (dependent) is predicted or estimated based on the value of the other variable (independent). Regression determines the cause and effect relationship between two variables which indicates that the change in the value of an independent variable also causes a change in the value of the dependent variable. Regression can be either simple (deals with two variables) or multiple (deals with two or more variables) in nature.

4.2.1 **Simple Regression Analysis**

The relationship between two variables that is, the independent variable and dependent variable which are expressed in a linear function or straight line is known as simple regression. The straight line is a regression line which is termed as the 'line of best fit' where the difference between the actual and estimated value is minimum. The line can be represented by the equation

$$Y= a + bX$$

From the above equation, y is the dependent variable whose value is estimated from the independent variable x. 'a' is known as the y-intercept of the place at which the line crosses the y-axis while 'b' represents the slope of line across the group. By adopting least squares method the 'a' and 'b' in the regression line can be calculated by:

$$b= \frac{N(\sum XY)-(\sum X)(\sum Y)}{N(\sum X^2)-(\sum X)^2}$$

$$a= \bar{Y} - b\bar{X}$$

Once the value of 'a' and 'b' are determined, the regression line must be fitted to the data by using least square method with the purpose of achieving the best fit by minimizing the difference between the actual and estimated value of Y. Therefore, the estimated value of y will be represented by

$$Y_e= a + bX$$

4.2.2 **Multiple Regression Analysis**

Unlike simple regression, multiple regression deals with two or more independent variables in an equation. The equations which describes the relationship between one dependent variable and two or more independent variables is known as multiple regression equation. Symbolically,

$$Y= a + b_1X_1+b_2X_2$$

From the above equation, we have two independent variables and , Y as dependent variables and three constants a, and . The increase in independent variables (and shows a high degree of correlation between themselves creating a problem which is described as a problem of multicollinearity. Such a problem can reduce the reliability of regression coefficients (namely and) therefore it is suggested to use only one set of independent variable for making estimation.

CONCLUSION

Data, facts and figures are silent but they do have complexities. The important characteristics hiding beneath them can be explored by using systematic analysis. It is always crucial for the researcher to have a careful planning of an analytical framework to ensure appropriate techniques are applied for the appropriate research.

BIBLIOGRAPHY

Gupta, S.(2010). Research Methodology, New Delhi, Deep & Deep Publications Pvt. Ltd, pp. 192-258

Kothari, C.R.(2009). Research Methodology-Methods & Techniques, New Delhi, New Age International (P) Limited, pp. 132-142

Khishnaswami, O.R & Ranganatham, M.(2011). Methodology of Research in Social Sciences, Mumbai, Himalaya Publishing House Pvt. Ltd., pp. 299-312

Levine, D. M., Krehbiel, T.C., Berenson, M.L. & Vishwanathan, P.K.(2011). Business Statistics, New Delhi, Dorling Kindersley (India) Pvt. Ltd., pp. 96-106

Levin, Richard I & Rubin, David S(1998). Statistics for Mangement, Noida, Dorling Kindersley Pvt. Ltd, pp.70-113

Saunders, Mark., Lewis, Philip & Thornhill, Adrian(2011). Research Methods for Business Students, Noida, Dorling Kindersley India Pvt. Ltd, pp. 444-449

Sharma, J.K(2007). Business Statistics, Noida, Dorling Kindersley India Pvt. Ltd, pp. 81-121 & 134-146 &172-186.