DRAWING INFERENCES ON THE BASIS OF LEVEL OF STATISTICAL TOOLS USED IN DATA ANALYSIS

Murari Karki*

Abstract

Statistics is the science which deals with the collection, organisation, tabulation, manipulation, analysis and interpretation of observed results(Winters R, Winters A, Amedee RG. Statistics: A brief overview. Ochsner J2010;10:213-6.).Using all the statistical procedure, we draw some inferences about population parameter from the study sample. Before drawing inferences about population parameters, we use different levels of statistical tools. This article is meant to provide some information to reader about various statistical methods that are useful for the process of making inferences about parameters. This article is focused on the various levels of statistical tools, difference between descriptive and inferential statistics and various test statistic to make conclusion about population parameters.

Keywords: *Parameters, Statistic, Descriptive statistics, Inferential statistics, Hypothesis testing, Parametric test.*

Introduction

In ancient time, statistics was used for the collection of data that are collected and maintained for the welfare of the people belonging to the state. We may define statistics either in a singular sense or in plural sense. When used as a plural sense, it is defined as data (Qualitative as well as Quantitative) that are collected, usually with a view of having statistical analysis. However, statistics, when used in a singular sense it may be defined as scientific method that is used for collecting, presenting and analyzing data, leading finally to drawing statistical inferences about some important characteristics (Horace Secrist. *Introduction to stataistical methods*: The economic journal 1926,141,115-116). It clearly indicates the following four stages in a statistical inquiry: -

- Collection of data
- Presentation of data
- Analysis of data and
- Interpretation of data

A sufficient and proper knowledge about statistics leads to give a desirable and meaningful result. On the other hand, inadequate use of statistics may cause various accidents and failure of various planning (Ali Z, Bhaskar SB. *Basic statistical tools in research and data analysis*. Indian J Anaesth 2016; 60:662-9).

* Assistant Professor of Saraswati Multiple Campus (Management Faculty: Statistics Department), Tribhuvan University, Nepal

Objective of the study

The primary objective of the study is to define various statistical tools used in pragmatic field of statistics. In addition, an illustration is also shown by using various level of statistical tools by taking secondary data of Pashmina export in the period of twelve years from fiscal year 2000/01 to 2011/12 published by 'Trade and Export promotion centre' Nepal.

Limitation of the Study

This study doesnot include the inferences based on non – parametric test.

Methodology

This study is primarily descriptive which describes various statistical procedures which are useful in pragmatic field. Some statistical tools such as diagrammatic representation, descriptive statististics, slope coefficients and hypothesis testing using t-test has been implemented.

Classification of data

Generally, there are four types of classification

- Geographical classification
- Chronological classification
- Qualitative classification
- Quantitative classification

Geographical classification

In this classification data are classified with respect to geographical region. i.e., regionwise, zone wise, districtwise etc. For example:

Table 1.

Population density of Nepal

Development region	Population density (per sq.km)
Eastern	188
Central	293
Western	155
Mid-western	71
Far-western	12

(Source: Census year 2001, CBS)

Chronological classification

A classification in which data are classified on the basis of differences in time e.g. the production of an industrial concern for different periods, population of a country in different decades, etc. For Example:

Table 2.

Growth rate of Nepal

Census year	Growth Rate (%)
1961	2.24
1971	2.10
1981	2.66
1991	2.10
2001	2.24

(Source: Census year 2001, CBS)

Qualitative classification

A Classification in which data are classified on the basis of some attribute or quality such as honesty, beauty, intelligence, occupation, sex, literacy etc. are known as qualitative classification.

In qualitative classification the data are classified according to the presence or absence of the attributes in the given units. If the data are classified into only two class with respect to an attribute the classification is called simple or dichotomous classification. When two or more attributes are classified into more than two classes it is known as manifold classification. For example:

Figure 1. Simple or dichotomous classification

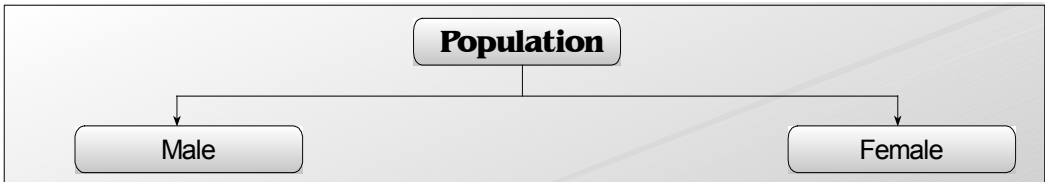
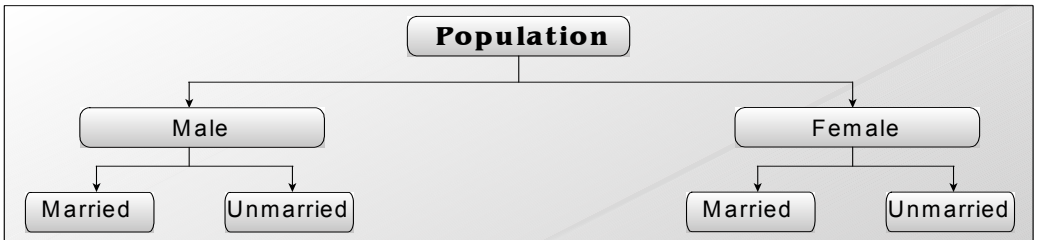


Figure 2. Manifold classification



Quantitative classification:

A Classification in which data are classified on the basis of numerical figures which represents the measures of certain characteristics is known as quantitative classification. For example: - age, height, weight, production, income, expenditure etc.

Table 3.

The wage distribution of workers:

Wage (in '00 Rs.)	Below 10	10–20	20–30	30–40	40–50
No. of workers	10	15	30	40	5

Variable

A quantity which can take different values at stated condition is known as variable (Kaur SP. *Variables in research*. Indian J Res Rep Med Sci 2013; 4:36-8.). For example, height, weight, marks obtained by a student's number of children's in a family etc. Basically there are two types of variables: (a) Discrete variable (b) Continuous variable.

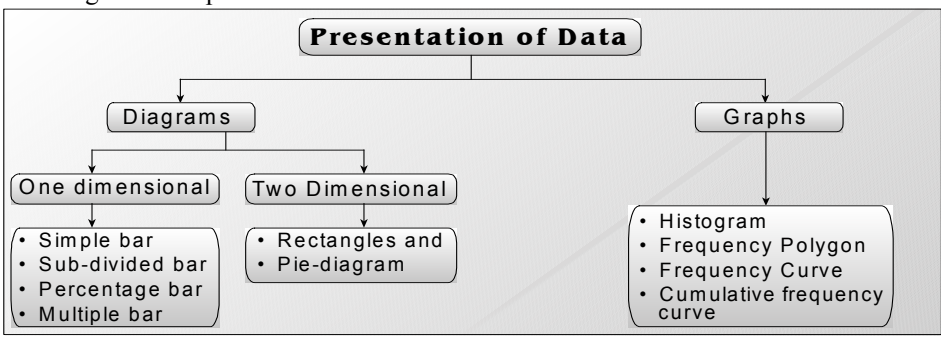
Discrete variable: A variable which can take only integer as its value or exact value is known as discrete variable. In other word, a variable which cannot be written in fraction is known as discrete variable. For example: - number of goals scored in a soccer game, number of children in a family, number of students in a class etc.

Continuous variable: A variable which can take all the possible values in a specified range is known as continuous variable. In other word a variable which can be written in fraction is known as continuous variable. For example, height, weight age temperature etc.

Presentation of data

Generally, people do not have time to go through mass data and understand its nature, in such cases diagrammatic and graphical representation are more intelligible, attractive and appealing. These presentations give a prompt view of data and facilitate comparison of various aspects of data.

Figure 3. Diagrammatic presentation of data



Statistical Tool

Measures of central tendency

It is the statistical tool which measures the central value of the data and represents whole data. It is an attempt to find one single figure to describe whole data. The following are the measures of central tendency which are commonly used in practice.

1. Arithmetic Mean
2. Geometric Mean
3. Harmonic Mean
4. Median
5. Mode

Arithmetic mean

The arithmetic mean (A.M.) is more popular than other means. It is helpful simply to analyze the data and in other further analysis of statistics. Generally, arithmetic mean is defined as the sum of observations divided by the number of observations. Let x_1, x_2, x_3,…,x_n be 'n' variate values of a random variable X. Then, arithmetic mean is computed by the following formula:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n} = \frac{\Sigma X}{n}$$

Geometric Mean

Geometric mean is defined as the "nth root of their product" of n observations. It is denoted by G.M. or G. If data are expressed into rate, ratio and percentage then G.M. is appropriate average.

Let, x_1, x_2,…,x_n are n non-negative, non-zero observations of a variable X then their geometric mean is given by

$$\text{Geometric mean (G.M.)} = \sqrt[n]{x_1 x_2 \ldots\ldots x_n}$$

$$\text{G.M.} = \text{Antilog}\left(\frac{\Sigma \log X}{n}\right)$$

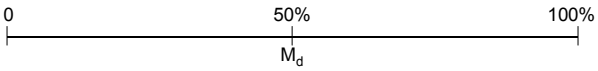
Objectives of Measures of Central Tendency

The main objectives of measures of central tendency are:

1. To get a single representative value of all the items of a series.
2. To facilitate comparison between two or more groups of data.
3. To facilitate the computation of other statistical measures.
4. To generalize on the basis of sample of data.
5. To help in decision making.
6. To find out the statistical relationships and classify them as required.

Median

Median is defined as the middle value of ordered data set, which divides the distribution into two equal parts. The number of observations below the median and number of observations above the median are equal i.e. 50% observations lies on both sides of the median. It is denoted by Md. It is also called positional average.



First of all, the given data should be arranged in ascending or descending order (generally ascending order) of their magnitude. If the number of observations is odd, the middle value gives the median and if the number of observation is even, there will be two middle values. In such case the arithmetic average of two middle values gives the median. The formula for calculating the median in case of individual series with n observations

$$\text{Position of median} \quad = \quad \text{Value of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{item}$$

Measures of dispersion

The term dispersion is used to indicate the facts that within a given group the items differ from one another in size or in other words, there is lack of uniformity in their sizes.–"W.I. King"

The common methods of measuring dispersion are:

1. Range
2. Quartile deviation or Semi- inter quartile range
3. Mean deviation or average deviation
4. Standard deviation
5. Lorenz curve

Quartile Deviation or Semi-inter Quartile Range

Semi Inter quartile range or Quartile deviation: Half of difference between largest and smallest quartile the is called semi-inter quartile range, which is also known as quartile deviation. Generally, it is denoted by quartile deviation (Q.D.) and given by,

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

Quartile deviation is an absolute measure of dispersion. For comparative studies of variability of two or more distributions having different units of measurement we need relative measure which is known as coefficient of quartile deviation and is givenby

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Standard Deviation

Standard deviation is the best measure of dispersion. It satisfies most of the characteristics of ideal measure of dispersion (It was first suggested by Karl Pearson in 1893). Usually it is denoted by Greek alphabet σ (sigma) and defined as "the positive square root of the arithmetic mean of the squares of the deviations of the given set of observations from their arithmetic mean". It is given by:

$$\text{S.D. } (\sigma) = \sqrt{\frac{\sum(X - \bar{X})^2}{n}}, \text{ n = Total number of observation}$$

$$= \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{1}{n}\sum X^2 - \bar{X}^2}$$

Objectives of Dispersion

The main objectives of Dispersion are:

1. To determine the reliability of central tendency.
2. To determine variation.
3. To compare consistency.
4. To determine cause of variability and control it.
5. To control quality.

Measures of Skewness and Kurtosis

Skewness measures the tailness and kurtosis measures the peakness of curve fitted for any data. Following are relative measures of Skewness.

1. Karl Pearson's coefficient of skewness
2. Bowley's coefficient of skewness
3. Kelly's coefficient of skewness

Karl Pearson's coefficient of skewness: This is relative measures of skewness based on $\bar{X}$, M_o and σ. It is denoted $S_k(P)$ and is given by

$$S_{k(P)} = \frac{\bar{X} - M_o}{\sigma}$$

When mode is ill-defined then

$$S_{k(P)} = \frac{3(\bar{X} - M_d)}{\sigma}$$

Bowley's Coefficient of Skewness: It is based on quartiles it is denoted by $S_k(B)$ and is given by

$$S_k(B) = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

Kelly's Coefficient of Skewness: It is based on Percentiles and deciles.

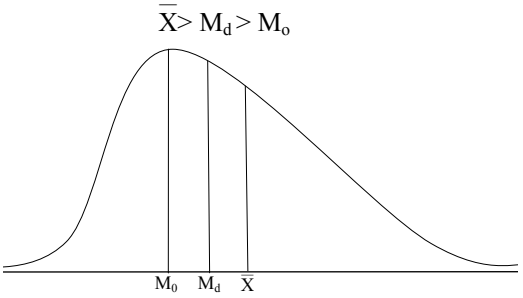
$$S_k(K) = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1} \quad \text{(Based on deciles)}$$

$$\text{Or, } S_k(K) = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \quad \text{(Based on percentile)}$$

Types of Skewness

Positive skewness: If the given frequencies curve has longer tail towards right side then it is said to be positive skewness.

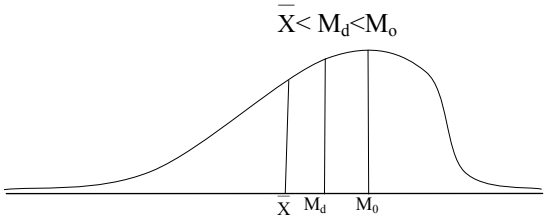
In this case,



Positively skewed curve

Negative skewness: If the given frequencies curve has longer tail towards left side then it is said to be negative skewness.

In this case,

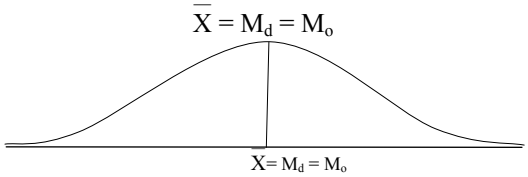


Negatively skewed curve

Symetrical Distribution: If the mean, median and mode of frequency distribution are equal then it is said to be symmetricali.e. $\bar{X}$= Md = M0.

In this case frequency distribution curve can be divided in two identical parts about the central value.

In this case,



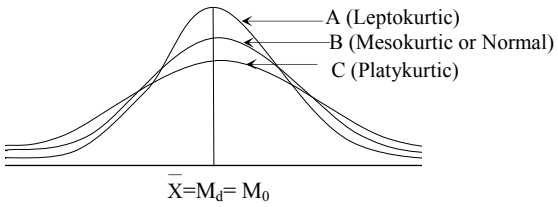
Symmetrical distribution

Percentile measures of kurtosis

The relative measure of kurtosis based on quartiles and percentiles is denoted by k and computed as

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{(P_{90} - P_{10})} = \frac{Q.D.}{P_{90} - P_{10}}$$

Karl Pearson's introduced following three patterns of peakedness.



Objectives of Skewness and Kurtosis

1. It helps in finding nature and degree of concentration of observations towards higher or lower values.
2. It gives the measures of extent of relationship between mean, median and mode.
3. It helps to measure whether curve is symmetrical or not.
4. Test the normality of the distribution.

Difference between Descriptive and Inferential Statistics

Descriptive statistics is used to describe the characteristics of the data gathered to study its various characteristics (Ali Z, Bhaskar SB. *Basic statistical tools in research and data analysis*. Indian J Anaesth 2016; 60:662-9). In descriptive statistics, there is no any uncertainity because we explore or describe the characteristic of observed data.

Inferential statistics refers to make predictions or conclusion about population charactristics by the study of samples drawn (F.A. Adesoji&M.A. Babatunde: *Datacollection, Management and analysis in academic research* ,2009). In the process of making inferences about population parameter we encounter certain level of risk that's why there is some uncertainity associated with the process. The risk factor associated in making inferences about parameter is known as level of significance.

Hypothesis testing

Hypothesis is the assumptions about population parameters which is tested by using various test statistic obtained from the study of sample. We use technique of testing of hypothesis to examine the truth or falsity of the statement about population parameter based on sample observations. There are two types of hypothesis.

- Null Hypothesis
- Alternative Hypothesis

Null Hypothesis refers to assumption of no significant difference between population parameter and its specified value and is denoted by H_0. Alternative Hypothesis is the assumption about population parameter other than Null hypothesis and is denoted by H_1.

Let us suppose that two different concerns manufacture electric light bulbs of brand A and brand B respectively. Each company claims that its brand is superior in terms of average life hours. For testing if out of the two brands, one is better than other or not , we set uphypothesis as

H_0: $\mu_A = \mu_B$ i.e. the two brands don't differ significantly in terms of average life hours. In other words there is no significant difference between the average life hours of two brands.

Vs H_1: $\mu_A \neq \mu_B$ i.e. the two brandsdiffer significantly in terms of average life hours.

Or H_1: $\mu_A > \mu_B$ (Right tailed test) i.e. Brand A is superior to Brand B in terms of average life hours.

Or H_1: $\mu_A < \mu_B$(Left tailed test) i.e. Brand B is superior to Brand A in terms of average life hours.

Test statistic

According to nature of sample, sample size, parameter to test , known/unknown values of population standard deviation e.t.c., we use different test such as Z – test, t – test, F – test, chi square test etc.

Z – test (Large sample test n > 30)

Z-test as a test of significance is used for: (i) sampling of variable and (ii) sampling of attributes.

For Sampling of Variables

1. To test the significance of a single mean.
2. To test the significance of difference between two independent sample means

For sampling of attributes

1. To test the significance of single proportion
2. To test the significance of difference between two independent sample proportions.

t – test (Small Sample test n ≤ 30)

t-distribution has many applications in testing of hypothesis. Major applications of t – test are: -

1. To test the significance of a sample mean, population variance being unknown.
2. To test the significance of difference between two independent sample means, the population variances being equal but unknown.
3. To test the significance of difference between two dependent sample means or paired t-test for difference of two means.
4. To test the significance of an observed sample correlation coefficient.

F – test

F-distribution has following applications in statistical theory. F-test is used:

1. For the test of equality of two population variances.
2. For test of equality of several population means.
3. For testing the significance of an observed sample multiple correlation coefficient.
4. For testing the significance of an observed sample correlation ratio.
5. For testing the linearity of regression.

$\chi 2$ - test (Non – parametric test)

$\chi 2$–test is one of the most popular statistical test in statistical inference. It has

a number of applications. Some important applications are as follows:

1. $\chi 2$–test for goodness of fit
2. $\chi 2$–test for independence of attributes
3. $\chi 2$–test for population variance

After the computation of value of test statistic, we make inferences about

population parameter by accepting or rejecting our Hypothesis by comparing calculated and tabulated value of respective test statistic. We may use concept of p – value for making decision about whether to accept or reject null hypothesis.

Illustration

Following data represents the value of export of Pashmina from Nepal to various countries in the period of fiscal year 2000/01 to 2011/12.

Table 4.

Value of Export of Pashmina

Fiscal year	Value of export in '000'
2000/01	3877965
2001/02	5269548
2002/03	1852220
2003/04	1534081
2004/05	1473675
2005/06	1460411
2006/07	1665949
2007/08	1789507
2008/09	1635343
2009/10	1317065
2010/11	1635629
2011/12	1908766

Source: Nepal overseas Trade Statistics (2000/01 to 2011/12), TEPC, Kathmandu

Overview of data and simple comparision of given data can be done by diagrammatic representation can be shown as

Figure 4. Line graph showing given data

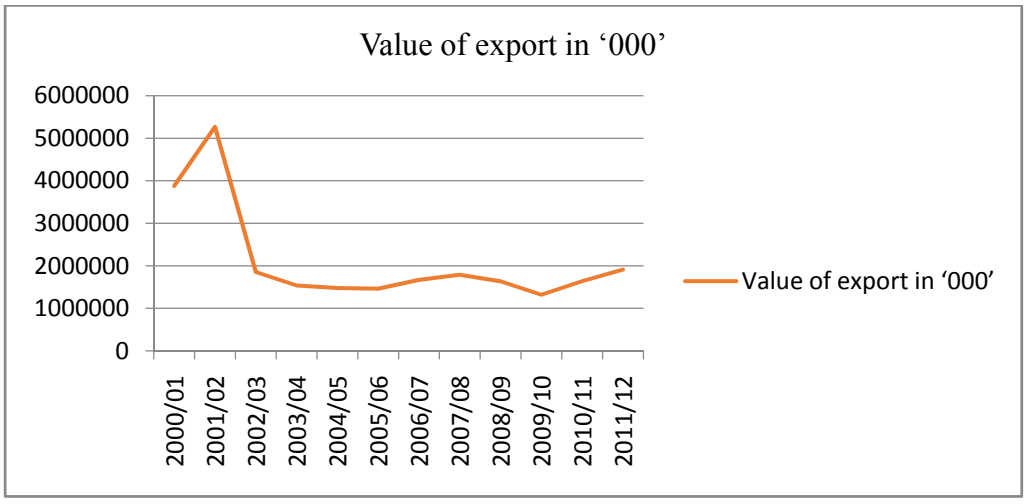


Figure 5. Bar diagram for given data

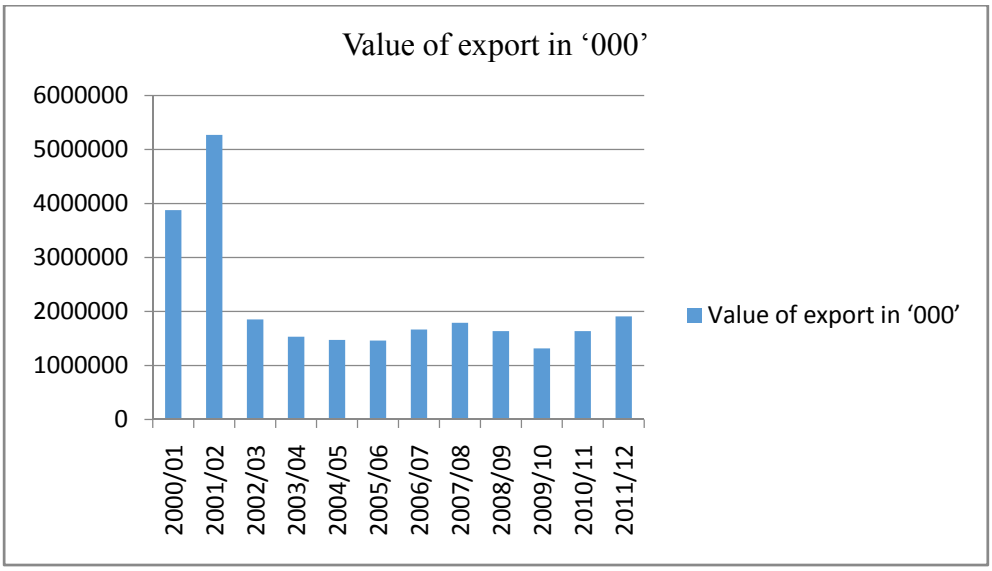
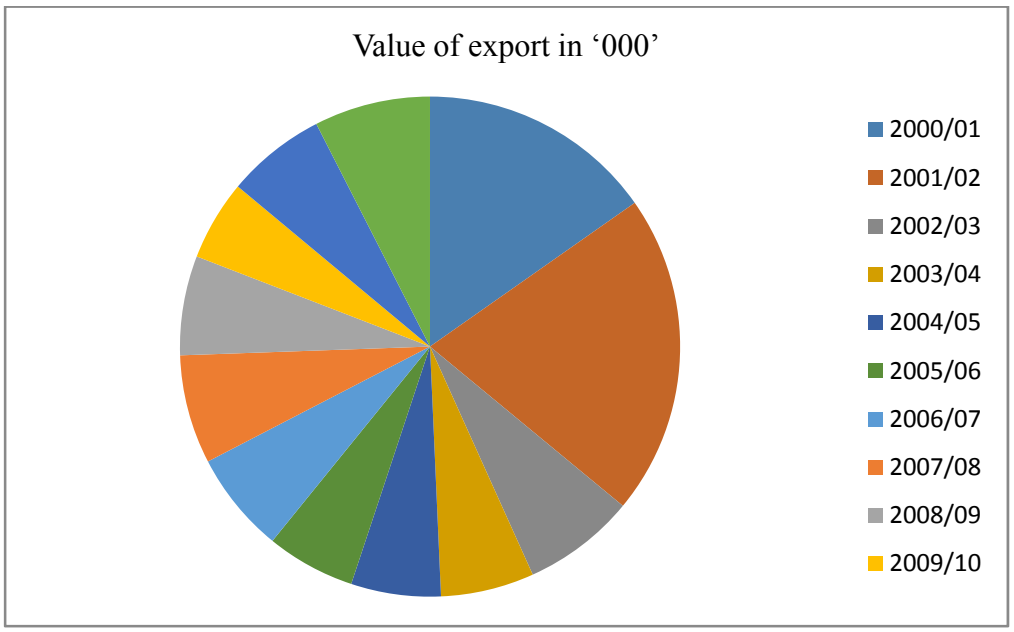


Figure 6. Pie – Chart for given data



Elementary level interpretation can be drawn by various level of statistical tools computed from given data. Here, statistical based computer software Ms – Excel and SPSS has been used to analyse the data.

Descriptive Statistics (Value of export in "000")

Statistical tools	Value
Range	3952483
Minimum	1317065
Maximum	5269548
Sum	25420159
Mean	2118346.58
Std. Deviation	1196665.124
Skewness	2.234
Kurtosis	4.362
Slope Coefficient	-98694.4

Hypothesis testing

Here, we are interested to test whether the average export is 200,00,00,000 or not. By using SPSS following result has been obtained.

One sample test

Value of export in '000' with Test Value μ= 2000000		
$t_{cal.}$	d.f.	Sig. (2-tailed)
0.343	11	0.738

Since, p – value is greater than level of significance (Generally taken as 5%) we accept Null hypothesis and reject alternative hypothesis with the conclusion that average value of export of Pashmina was 200,00,00,000 (Two Hundred crores).

Conflict of interest

There are no conflicts of interest.

References

Ali Z, Bhaskar SB. Basic statistical tools in research and data analysis. *Indian J Anaesth* 2016;60:662-9.

F.A.Adesoji&M.A.Babatunde:*Datacollection,Management and analysis in academic research*,2009

Kaur SP. *Variables in research*. Indian J Res Rep Med Sci 2013;4:36-8

Horace Secrist. *Introduction to stataistical methods*: The economic journal 1926,141,115-116

Winters R, Winters A, Amedee RG. Statistics: A brief overview. *Ochsner J*2010;10:213-6.