F.Y.B.Com.
(Entrepreneurship)

Business Statistics with MS - Excel

Contents

Unit 1

Measure of Central Tendency & Dispersion

1.1 Unit Structure

- Frequency distribution: Raw data, attributes and variables, Classification of data, frequency distribution, cumulative frequency distribution, Histogram & Ogive curves.
- Concept of central tendency, Desirable Properties for good measures of central tendency.
- Measures of central tendency: Arithmetic mean, median and mode for grouped and ungrouped data, Combined mean for two groups.
- Appropriate choice of measures.
- Measures of dispersion: Range, Standard deviation (S.D.) for grouped and ungrouped data, combined S.D., Variance.
- Measures of relative dispersion: coefficient of range, coefficient of variation
- Skewness and Kurtosis.

1.2 Frequency distribution

Raw data

Raw data is the unorganized data when we're done with the collection stage. This is because it is similar to a lump of clay with no identity and also of no practical use. It is important to realize that organized data facilitates comparison and meaningful conclusions. Further, to organize the data we need to look for similarities or group the data. In this way, we effectively convert heterogeneous data into homogeneous data. To do so, an investigator has to classify the data in the form of a series. Series refer to those data which are in some order and sequence. Thus, if we arrange the data in the example mentioned in the introduction according to the classes in your school, we will eventually classify the data in form of a statistical series. Note that we can also arrange them according to their heights. Hence, this basis of the arrangement of raw data can vary from purpose to purpose.

Variable

A variable is simply something that can vary with time and we can measure this variation. In other words, a variable is a characteristic or a phenomenon which is capable of being measured and changes its value over time. A variable is classified into two:

1.2.2.1 Discrete

Value of a discrete variable changes only in complete numbers or increases in jumps. Thus the phenomenon or characteristic, a discrete variable represents should be such that its value cannot be infractions but only in whole numbers. For example, the number of children in a family can be 2, 3, 4 etc but not 2.5, 3.5 etc.

1.2.2.2 Continuous

A continuous variable assumes fractional values or its value does not increase in jumps. For example, the heights of students, the weights of students and so on.

1.3 Classification of Data

The main objective of the organization of data is to arrange the data in such a form that it becomes fairly easy to compare and analyze. Generally, we can do this by distributing data into various classes on the basis of some attribute or characteristic. This distribution of data into classes is the classification of data. Further, each division of data is a class. All in all, through the process of classification we can group and divide data into classes according to a general attribute, which facilitates comparison and analysis.

Objectives of Classification

1) Simplification and Briefness: Classification presents data in a brief manner. Hence, it becomes fairly easy to analyze the data.

2) Utility: As classification highlights the similarity in the data, it brings out its utility.

3) Distinctiveness: With the help of grouping data into different classes, classification also brings out the distinctiveness in data.

4) Comparability: As already mentioned, it facilitates comparison of data.

5) Scientific Arrangement: Classification arranges data on scientific lines. Thus it also increases the reliability of data.

6) Attractive and Effective: Lastly, through the process of classification, data becomes effective and attractive.

Basis of Classification

Definitely, we can classify a given data according to various characteristics, depending on the purpose of our study. Evidently, there is the various basis of classification.

1.3.2.1 Geographical classification

When we classify data according to different locations, it is termed as a geographical classification of data. For example, a classification of the data about the number of children aged between 3-8 according to the various cities in India.

1.3.2.2 Chronological Classification

In chronological classification, we classify data according to time i.e., it follows a chronological sequence. For example, the classification of the data about the number of deaths in India according to the years.

1.3.2.3 Qualitative Classification

Here, we classify data according to the qualities or attributes of data. One key point to remember is that an attribute is qualitative in nature i.e. we cannot measure an attribute in quantitative terms like 5, 1, 2 etc. This qualification is further of two types:

1.3.2.3.1 Simple

In the simple qualitative classification of data, we qualify data exactly into two groups. One group has data items that exhibit the quality, the other group doesn't. Evidently, it is also known as classification according to a dichotomy. Example of classes can be educated-uneducated, male-female and so on.

1.3.2.3.2 Manifold

Here we classify data according to more than one characteristic of an attribute. This means one we classify data into two groups according to an attribute, the two groups are further divided into two according to another attribute. As a result, there can be many levels of classification couples with more than just two classes. For example, the classification of data about students in a class, according to their gender, followed by classification according to whether they are fat or not.

1.3.2.4 Quantitative or Numerical Classification

Unlike qualitative classification, quantitative classification allows numerical division of data into classes. Here, each class represents a range of numerical values for the phenomenon under consideration. Accordingly, we frame each class with a lower and higher value and according to the range of data.

Again, the phenomenon should be such that it can be expressed in numerical terms. As it is classified into classes with a different range of values, this classification is effectively the representation of the change of the value of a phenomenon over time or across different regions. Which means its

value varies. Accordingly, quantitative classification is also known as classification by variables.

The different types of frequency distributions are ungrouped frequency distributions, grouped frequency distributions, cumulative frequency distributions, and relative frequency distributions.

1.3.2.5 Grouped Frequency Distribution

Sometimes to make deriving insights from an observation easily, we group them into class intervals.

Calculate the maximum and minimum value of the data set.

Divide this range by the number of groups you intend to have in your analysis.

Segregate the data within this small sub-group basis the class width.

Calculate the frequency of data within each group.

1.4 Ungrouped Frequency Distribution

The ungrouped cumulative distribution is similar to grouped frequency distribution except for the fact that class intervals are not created, and values are ordered from minimum to maximum.

List the unique values as the first column.

Calculate the repeated instances of each unique value and record it.

1.5 Cumulative Frequency Distribution

When you add or subtract the frequencies of all the previous class intervals to determine the frequency of a particular class interval, it results in a cumulative frequency distribution. Also, another major difference is that class intervals do not denote a range but instead represent a logical conclusion like greater than a threshold value or less than a threshold value.

Calculate frequencies for every category.

Arrange in ascending or descending order according to categories/class intervals based on whether one wants to prepare an increasing/decreasing cumulative frequency distribution.

Total all the preceding frequencies. E.g., the second category's frequency is calculated by the sum of the first and second category's individual frequencies.

Third is calculated by the sum of the first, second, third category's individual frequencies.

1.6 Relative Frequency Distribution

A relative frequency distribution is extensively used in our day-to-day statistical applications, which refers to the proportion of total observations associated with each category. It is calculated for individual class intervals by dividing them by the total observed frequencies. Relative frequencies can be written as a percentage, fraction, or decimal points. Cumulative relative frequency is the total of all preceding relative frequencies. To find the cumulative relative frequency, total all the previous relative frequencies till the current category.

Solved Examples

1) A research was done in 20 homes in Chennai Avadi. People were asked how many bikes did they own?

   The results were:1, 4, 3, 0, 5, 1, 2, 2, 1, 5, 2, 3, 2, 2, 0, 1, 2, 0, 3, 2.

   Present this data in Frequency Distribution Table. Also, find the maximum number of homes owning the same number of bikes.

   Solution:

   Divide the number of bikes in every home into different intervals. Every house can own either 0,1,2,3, etc. bikes. All these numbers form the rows. Now calculate the number of homes having {0,1,2,3, etc.} bikes. This is called the frequency. When you plot this in the form of a table:

Number of Bikes	Frequency
0	3
1	4
2	6
3	3
4	2
5	2

   It can be seen from the table that 6 homes have 2 bikes and a lesser number of people own other numbers of bikes. Hence the answer is 6 homes.

1.7 Ogive Definition:

The Ogive is defined as the frequency distribution graph of a series. The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative per cent frequencies on the vertical axis.

Cumulative frequency is defined as the sum of all the previous frequencies up to the current point. To find the popularity of the given data or the likelihood of the data that fall within the certain frequency range, Ogive curve helps in finding those details accurately.

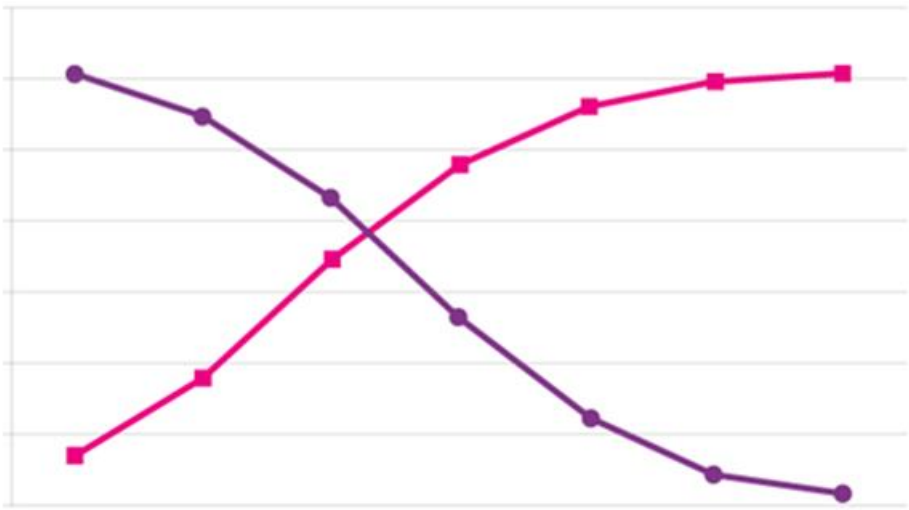
Create the Ogive by plotting the point corresponding to the cumulative frequency of each class interval. Most of the Statisticians use Ogive curve, to illustrate the data in the pictorial representation. It helps in estimating the number of observations which are less than or equal to the particular value.

Ogive Graph

The graphs of the frequency distribution are frequency graphs that are used to exhibit the characteristics of discrete and continuous data. Such figures are more appealing to the eye than the tabulated data. It helps us to facilitate the comparative study of two or more frequency distributions. We can relate the shape and pattern of the two frequency distributions.

The two methods of Ogives are:

- Less than Ogive
- Greater than or more than Ogive



The graph given above represents less than and the greater than Ogive curve. The rising curve (Brown Curve) represents the less than Ogive, and the falling curve (Green Curve) represents the greater than Ogive.

1.7.1.1 Less than Ogive

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

1.7.1.2 Greater than or More than Ogive

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class, second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

1.7.1.3 Ogive Chart

An Ogive Chart is a curve of the cumulative frequency distribution or cumulative relative frequency distribution. For drawing such a curve, the frequencies must be expressed as a percentage of the total frequency. Then, such percentages are cumulated and plotted, as in the case of an Ogive.

Below are the steps to construct the less than and greater than Ogive.

How to Draw Less Than Ogive Curve?

Step 1. Draw and mark the horizontal and vertical axes.

Step 2. Take the cumulative frequencies along the y-axis (vertical axis) and the upper-class limits on the x-axis (horizontal axis).

Step 3. Against each upper-class limit, plot the cumulative frequencies.

Step 4. Connect the points with a continuous curve.

How to Draw Greater than or More than Ogive Curve?

Step 1. Draw and mark the horizontal and vertical axes.

Step 5. Take the cumulative frequencies along the y-axis (vertical axis) and the lower-class limits on the x-axis (horizontal axis).

Step 6. Against each lower-class limit, plot the cumulative frequencies.

Step 7. Connect the points with a continuous curve.

1.7.1.4 Uses of Ogive Curve

Ogive Graph or the cumulative frequency graphs are used to find the median of the given set of data. If both, less than and greater than, cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which, both the curve intersects, corresponding to the x-axis, gives the median value. Apart from finding the medians, Ogives are used in computing the percentiles of the data set values.

Ogive Example

1) Construct the more than cumulative frequency table and draw the Ogive for the below-given data.

Marks	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	3	8	12	14	10	6	5	2

Solution:

"More than" Cumulative Frequency Table:

Marks	Frequency	More than Cumulative Frequency
More than 1	3	60
More than 11	8	57
More than 21	12	49
More than 31	14	37
More than 41	10	23
More than 51	6	13
More than 61	5	7
More than 71	2	2

Plotting an Ogive:

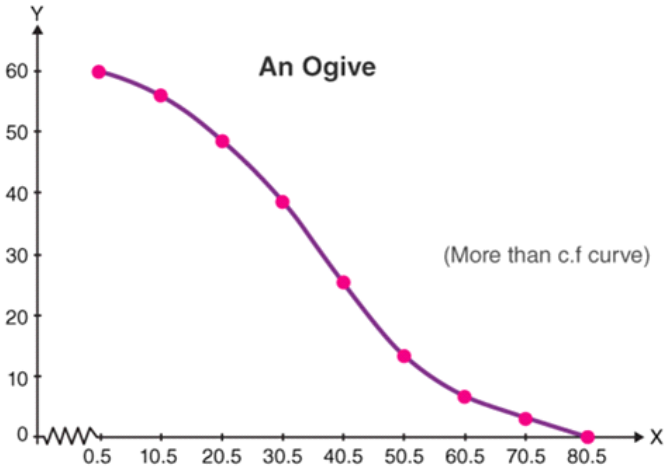
Plot the points with coordinates such as (70.5, 2), (60.5, 7), (50.5, 13), (40.5, 23), (30.5, 37), (20.5, 49), (10.5, 57), (0.5, 60).

An Ogive is connected to a point on the x-axis, that represents the actual upper limit of the last class, i.e.,(80.5, 0)

Take x-axis, 1 cm = 10 marks

Y-axis, 1 cm = 10 c.f

More than the Ogive Curve:



1.8 Central Tendency:

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution." It aims to provide an accurate description of the entire data. It is the single value that is most typical / representative of the collected data. The term "number crunching" is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

Properties of a good measure of central tendency:

1) It is rigidly defined

7) It is based on all values of the data

8) It should not be affected by the extreme values of the data

9) It should have the sampling stability

10) It should be capable of further statistical analysis.

[redacted]Arithmetic Mean (A.M)

For Simple or Ungrouped data:

(where frequencies are not given) Arithmetic Mean is defined as the sum of all the observations divided by the total number of observations in the data and is denoted by x, which is read as 'x-bar'

formula

[redacted]Median

The median by definition refers to the middle value in a distribution. Median is the value of the variable which divides the distribution into two equal parts. The 50% observations lie below the value of the median and 50% observations lie above it. Median is called a positional average. Median is denoted by M.

1.8.2.1 For Simple data or ungrouped data

Median is defined as the value of the middle item of a series when the observations have been arranged in ascending or descending order of magnitude.

Steps:

Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer).

gfjgj

1.8.2.2 For Ungrouped Frequency Distribution

Steps:

Arrange the data in ascending or descending order of magnitude with respective frequencies .

Find the cumulative frequency (c. f) less than type .

Find N/2 , N = total frequency.

See the c. f column either equal or greater than N/2 and determine the value of the variable corresponding to it . That gives the value of Median.

1.8.2.3 For Grouped Data

Steps:

Find the c.f less than type

Find N/2, N= total frequency.

See the c.f column just greater than N/2.

The Corresponding class interval is called the Median class.

To find Median, use the following formula.

[redacted]Mode

1.8.3.1 For Raw Data

Mode is the value which occurs most frequently, in a set of observations. It is a value which is repeated maximum number of times and is denoted by Z.

Example 19: Find mode for the following data.

64, 38,35,68,35,94,42,35,52,35

Solution:

As the number 35 is repeated maximum number of times that is 4 times.

Mode=35 units.

For ungrouped frequency distribution:

Mode is the value of the variable corresponding to the highest frequency.

Example 20: Calculate the mode for the following data.

Size of Shoe:	5	6	7	8	9	10
No. of Pairs:	38	43	48	56	25	22

Solution: Here the highest frequency is 56 against the size 8.

Modal size = 8.

1.8.3.2 For Grouped data:

In a Continuous distribution first the modal class is determined. The class interval corresponding to the highest frequency is called modal class.

1.9 Measures of Dispersion

Range

Range is the simplest measure of dispersion.

When the data are arranged in an array the difference between the largest and the smallest values in the group is called the Range.

Symbolically: Absolute Range = L - S, [where L is the largest value and S is the smallest value]

Amongst all the methods of studying dispersion range is the simplest to calculate and to understand but it is not used generally because of the following reasons:

Since it is based on the smallest and the largest values of the distribution, it is unduly influenced by two unusual values at either end. On this account, range is usually not used to describe a sample having one or a few unusual values at one or the other end It is not affected by the values of various items comprised in the

distribution. Thus, it cannot give any information about the general characters of the distribution within the two extreme observations.

For example, let us consider the following three series:

Series A: 6 46 46 46 46 46 46 46

Series B: 6 6 6 6 46 46 46 46

Series C: 6 10 15 25 30 32 40 46

It can be noted that in all three series the range is the same, i.e. 40, however the distributions are not alike: the averages in each case is also quite different. It is because range is not sensitive to the values of individual items included in the distribution. It thus cannot be depended upon to give any guidance for determining the dispersion of the values within a distribution.

Standard Deviation

As we have seen range is unstable, quartile deviation excludes half the data arbitrarily and mean deviation neglects algebraic signs of the deviations, a measure of dispersion that does not suffer from any of these defects and is at the same time useful in statistic work is standard deviation. In 1893 Karl Pearson first introduced the concept. It is considered as one of the best measures of dispersion as it satisfies the requisites of a good measure of dispersion. The standard deviation measures the absolute dispersion or variability of a distribution. The greater the amount of variability or dispersion greater is the value of standard deviation. In common language a small value of standard deviation means greater uniformity of the data and homogeneity of the distribution. It is due to this reason that standard deviation is considered as a good indicator of the representativeness of the mean.

It is represented by σ (read as _sigma').

σ2 i.e., the square of the standard deviation is called variance. Here, each deviation is squared.

The measure is calculated as the average of deviations from arithmetic mean. To avoid positive and negative signs, the deviations are squared. Further, squaring gives added weight to extreme measures, which is a desirable feature for some types of data. It is a square root of arithmetic mean of the squared deviations of individual items from their arithmetic mean.

The mean of squared deviation, i.e., the square of standard deviation is known as variance. Standard deviation is one of the most important measures of variation used in Statistics. Let us see how to compute the measure in different situation.

Combined Mean:

A combined mean is a mean of two or more separate groups, and is found by :

Calculating the mean of each group,

Combining the results.

Formula:

A combined mean for two sets can be calculated by the formula

$$x_c = \frac{m \cdot x_a + n \cdot x_b}{m + n}$$

Where:

xa = the mean of the first set,

m = the number of items in the first set,

xb = the mean of the second set,

n = the number of items in the second set,

xc the combined mean.

A combined mean is simply a weighted mean, where the weights are the size of each group.

Combined Standard Deviation

1.10 Skewness:

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Definition of Skewness:

For univariate data Y1, Y2, ..., YN, the formula for Skewness is:

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the Skewness, the s is computed with N in the denominator rather than N - 1.

The above formula for Skewness is referred to as the Fisher-Pearson coefficient of Skewness. Many software programs actually compute the adjusted Fisher-Pearson coefficient of Skewness

This is an adjustment for sample size. The adjustment approaches 1 as N gets large. For reference, the adjustment factor is 1.49 for N = 5, 1.19 for N = 10, 1.08 for N = 20, 1.05 for N = 30, and 1.02 for N = 100.

The Skewness for a normal distribution is zero, and any symmetric data should have Skewness near zero. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. If the data are multi-modal, then this may affect the sign of the Skewness.

Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

It should be noted that there are alternative definitions of Skewness in the literature. For example, the Galton Skewness (also known as Bowley's Skewness) is defined as

where Q1 is the lower quartile, Q3 is the upper quartile, and Q2 is the median.

The Pearson 2 Skewness coefficient is defined as

where Y~ is the sample median.

There are many other definitions for Skewness that will not be discussed here.

1.11 Kurtosis:

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the Skewness and kurtosis of data set.

Definition of Kurtosis:

For univariate data Y1, Y2, ..., YN, the formula for kurtosis is:

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using N in the denominator rather than N - 1.

Unit 2

Correlation and Regression

2

2.1 Introduction

a

2.2

Unit 3

Index Number and Time Series

3

3.1 Introduction

This.

Unit 4

Probability

4

4.1 Introduction

This

.

Unit 5

Revision

5

5.1 Introduction

This.

.

Unit 6

Question Bank

6

6.1 Introduction

This

.

Unit 7

Extras

7

7.1 Thesis Summary

This .

11)

12)