College of Defence Management

*Victory through Excellence*

# Introduction to Descriptive Statistics

## Pre Course Training Handouts

**Descriptive >>> Diagnostic >>> Predictive >>> Prescriptive**

# PRE-COURSE TRAINING HANDOUTS

# INTRODUCTION TO DESCRIPTIVE STATISTICS

## CONTENTS

**The purpose of this handout is to acquaint the participants with an overview of Descriptive Statistics, which is a Foundational Subject in the Higher Defence Management Course. A prior understanding of the basics of the subject will assist the participants in easy comprehension during the contact sessions at CDM.**

**Ver 1.3 [ Apr 2024]**

# INTRODUCTION TO STATISTICS

***"Statistical thinking will one day be as necessary as the ability to read and write"***

- HG Wells

## Introduction

1. Every day we come across a lot of information in the form of facts, numerical figures, tables, graphs, etc. These may relate to cricket batting or bowling averages, profits of a company, temperatures of cities, expenditures in various sectors of a five-year plan, polling results, and so on. These facts or figures, which are numerical or otherwise, collected with a definite purpose are called data. Our world is becoming more and more information oriented. Every part of our lives utilises data in one form or the other. So, it becomes essential for us to know how to extract meaningful information from such data. This extraction of meaningful information is studied in a branch of mathematics called Statistics.

2. In today's information-overloaded age, statistics is one of the most useful subjects everyone must learn. Knowing a little about statistics will help one to make more informed decisions about these and other important questions. On the flip side, statistics can be used to twist the truth for their own gain. For example, a Regiment may claim quite truthfully that eight out of ten personnel of their regiment come in Excellent in BPET during the Annual Inspection for the last five years. What they may not mention is that the personnel who are poor in BPET are either detailed on essential duty or are on leave. As a Commander, you should be empowered to realise the truth. Thus, for Decision Makers, it is important to develop the ability to extract meaningful information from raw data to make better decisions. It is possible only through the careful analysis of data guided by a statistical thinking.

3. ***Statistics is concerned with scientific methods for collecting, processing, presenting and analysing and modelling data and what is even more important, drawing valid conclusions for making reasonable decisions based on such analysis***. It is in general divided into two branches **Descriptive Statistics** and **Inferential Statistics**. The former focuses on the collection, summarization and characterization of a set of data and the latter estimates characteristics of a data or uncover patterns on a data set that are unlikely to occur by chance and help draw inferences.

## Stages in a Statistical Investigation

4. From the definition of statistics presented above, the stages involved in a statistical study are:-

(a) **Collection of Data**. The first step in a statistical investigation is the collection of data. The purpose of the enquiry, its scope, source of information and type of enquiry are the preliminary aspects to be considered before the actual collection of data. The purpose of the enquiry would determine the nature of the data to be collected. The scope, i.e. coverage concerning place, time and type of data to be collected, would determine the amount of data to be gathered. Source of information determines the method to be adopted for data collection, viz. primary or secondary method. It is preferable to adopt the primary method. Type of enquiry could be classified as open/confidential, original/repetitive, official/non-official, and census/sample study. When statistical methods are used, the problem is formulated in terms of population, which is defined as a collection of all the elements with a measurable attribute. However, it shall not always be possible to collect data from the entire population, then a representative subset of the population is chosen to be studied, which is called a sample. For proper analysis, data must be gathered systematically. While details about some of these aspects would be discussed subsequently, it is appropriate at this stage to distinguish between census and sample studies.

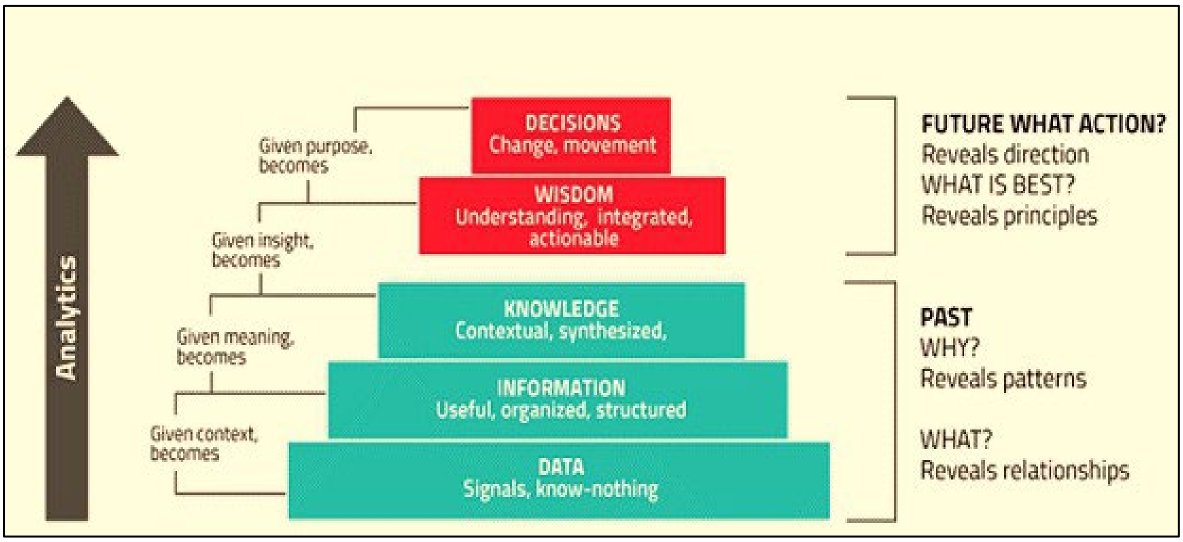
(i) **Census**. In data collection language, this process of eliciting data or information from/about every element of the population or universe is called Census. The attribute or characteristics about which information is gathered is termed as ***parameter*** in the case of a census.

(ii) **Survey**. If the data is collected from a sample, then the process is called a Survey. The characteristic in case of a sample is termed as ***statistic***.

(b) **Presentation of Data**. While the terms 'data' and 'statistics' are often used interchangeably, there is an important distinction between them. Data are individual pieces of factual information recorded and used for analysis. It is the raw information from which statistics are created. Statistics are the results of data analysis - its interpretation and presentation. The data collected must be presented in a suitable form for studying its salient features. The raw data, therefore, needs to be put into a compact form by classifying it and then presented by employing various methods like Tables, Diagrams and Graphs.

(c) **Analysis of Data**. Analysis implies studying the nature of the data. Salient features of the data are studied with the aid of statistical tools ranging from the simplest to the complicated.

(d) **Interpretation of Data**. The final step in a statistical investigation is interpreting the data collected. This implies the technique of drawing conclusions from the critical study of the collected data. While interpreting, the limitations of the original data must be duly considered. Interpretation is a task to be undertaken by skilled investigators. The Data to Wisdom hierarchy or the DIKW pyramid is shown in the figure below.



## Types of data

5. Data are a collection of facts such as values or measurements. It can be numbers, words, measurements, observations, or even just descriptions of things. Basically, data are two types: ***constant and variable***. Constant is a situation/value that does not change, while a characteristic, number, or quantity that increases or decreases over time or takes different values in different situations is called variable.

6. A **variate or random variable** is a quantity or attribute whose value may vary from one unit of investigation to another. Consumer behaviour, profit/loss of an entity, job satisfaction, smoking habits, leadership ability etc. are few examples of variables. An **observation or response** is the value taken by a variate for some given unit. There are various types of variate.

(a) **Categorical (Qualitative)**. Described by a word or phrase (e.g. Gender, colour).

(b) **Numerical (Quantitative)**. Described by a number (e.g. time till cure, number of calls arriving at a telephone exchange in 5 seconds). Quantitative data can be:

(i) **Discrete**. The variate can only take one of a finite or countable number of values (e.g. number of students present in a class, Number of Soaps sold in URC).

(ii) **Continuous**. The variate is a measurement which can take any value in an interval of the real line (e.g. Height of a student).

7. **Scales of Measurement**. Data can also be described in terms of levels of measurement or Scales of Measurement. In general, the principles of measurement have four characteristics; ***Classification, Order, Difference and Origin***. Combination of these characteristics provides four widely used classification of measurement of scales; ***Nominal, Ordinal, Interval and Ratio***.

| *Scale* | *Description* | *Example* | *Type of Data* | *Possible Operations* |
|---|---|---|---|---|
| **Nominal** | Data consists of names or categories. No ordering scheme possible. | • Jersey Number assigned to a player<br>• Aadhar Number | Discrete | Counting & % Calculation |
| **Ordinal** | Data is arranged in some order but the difference between values cannot be determined or are meaningless. | • Rank order of runners in a race | Discrete | Counting & % Calculation |
| **Interval** | Data is arranged in order and the difference can be found. However, there is no inherent starting point and ratios are meaningless. | • Temperature in centigrade ($80^0$ is $20^0$ hotter than $60^0$, which is $20^0$ hotter than $40^0$, but $80^0$ is not twice as hot as $40^0$) | Continuous | Addition & Subtraction |
| **Ratio** | An extension of an interval scale that includes an inherent Zero starting point. Both differences and ratios are meaningful. | • Temperature in Kelvin ($80^0$ is twice as hot as $40^0$)<br>• Weight, Length, Age | Continuous | Addition, Subtraction, Multiplication, Division & All Statistical techniques. |

# DATA CLASSIFICATION, TABULATION AND PRESENTATION

## Data Classification – Frequency Distribution

8. Very often, we are confronted by huge volumes of data that are difficult to process as individual items to make meaningful inferences. A useful method to process large volumes of data is to classify them into sub-categories. By this, we can reduce the large volume of data to a more manageable volume. One method to achieve this is through Frequency Distribution. Data is condensed by classifying it. The total range of observations is divided into a limited number of groups. For example, if we collect the data on the bodyweight of HDMC participants, we would be left with 160 data points from which we can infer very little. On the other hand, if these numbers are put into bins or bands, say 61 – 65 kg, 66 – 70 kg, 71 – 75 kg and so on, we can arrive at as tabulated aside.

| **Weight Band** | **Number of Participants** |
|---|---|
| <= 60 Kg | 6 |
| 61 – 65 kg | 18 |
| 66 – 70 kg | 22 |
| 71 – 75 kg | 31 |
| 76 – 80 kg | 34 |
| 81 – 85 kg | 22 |
| 85 – 90 kg | 20 |
| > 90 kg | 7 |
| **TOTAL** | **160** |

9. In this Table, the categories or weight bands are called ***Classes***. A ***class is a range of values that the variable (in this case, weight) can assume***. The ***number of participants in each category or weight band is called Frequency of that Class***. Data entered in a frequency distribution is considered as Grouped Data. This enables us to get a clear picture of the pattern of observations and to establish the characteristics of the mass of data. We can also glean some important information regarding the data set by identifying the Maximum value, the Minimum value and the Range (Maximum – Minimum) of the data set.

10. Certain terms with respect to a Frequency Distribution are as follows: -

(a) **Class Limits**. Every class is limited by what is called class limits. The class limits are the numbers that typically serve to identify the classes in the listing of a frequency distribution. The upper limit defines the highest value to be included in a class.

(b) **Class Interval**. The width of a class is called the class interval. It is determined by dividing Range with the number of classes (k).

(c) **Mid Class Mark**. It is the point dividing the class into two equal halves. Adding the lower class limits (LCL) and upper class limits (UCL) and dividing by 2 you get the Mid Class Mark (MCM). The mid-class mark or mid-point is used to represent all values in the class for statistical calculations. This is based on the assumption that the observations are spread evenly over the class interval.

(d) **Frequency**. Each class has a number of items/observations that fall within the range of its interval. This number is called the frequency of the class.

(e) **Relative Frequency**. To comprehend better and enable comparison of frequency tables of different sets of observations, we use the ratio of the frequency of a class to the total number of observations. This ratio is known as the relative frequency of that class.

(f) **Cumulative Frequency and Cumulative Relative Frequency**. The cumulative frequency and cumulative relative frequency of a class are arrived at by summing up the frequencies and relative frequencies of all classes up to and including that particular class.

**Number of Classes**.

- The number of classes into which the data is to be condensed depends upon; (a) The number of items in the entire series, (b) The lowest and highest values or Range of data, (c) Even distribution of items within the classes, and (d) A regular sequence of frequencies.
- Generally, 5 to 15 classes are chosen in practice. The rough guideline for determining the number of classes k, given by Sturges' approximation is as below:-

$$\mathbf{k = 1 + 3.3 \log (n)}$$

***where n is the total number of observations in the sample***

- The criterion to be borne in mind is that the grouped data should resemble the parent set as closely as possible. If we have too many classes, the pattern is not clearly revealed and too few classes result in excessive concentration of values.

11. **Construction of Frequency Distribution Using MS Excel**. Using the ***Frequency function***, we can convert the data into a frequency table.

**Example**; Marks obtained by fifteen HDMC participants in Statistics final exam is given below. Construct a frequency table.

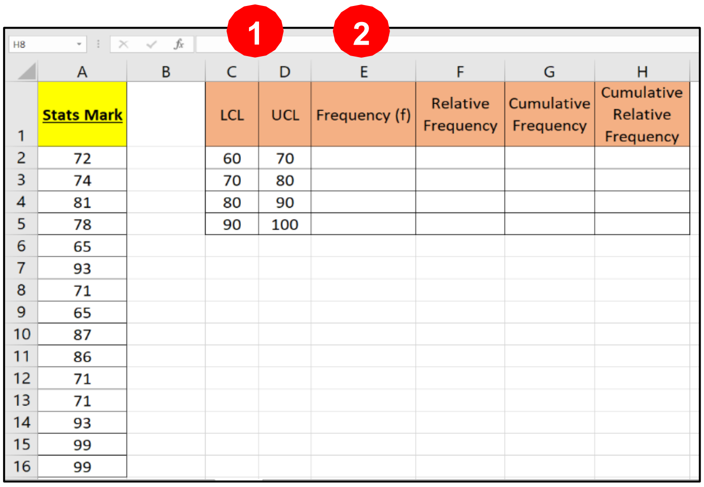
| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 72 | 74 | 81 | 78 | 65 | 93 | 71 | 65 | 87 | 86 | 71 | 71 | 93 | 99 | 99 |

1 Now in order to calculate frequency, we have to group the data with students marks as shown. Make Class like 60-70, 70-80 and so on.

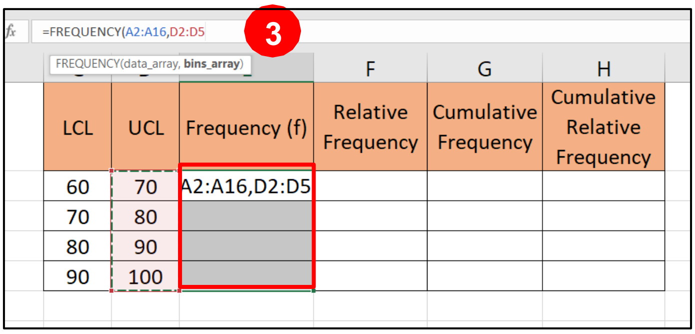
2 Create a new column named Frequency.

3 Use the frequency formulation on E column by selecting E2 to E5. Here we need to select the entire frequency column then only the frequency function will work properly or else we will get an error value.

As shown in the screenshot we have selected column as data array and Bin array as UCL in the function **=FREQUENCY(A2:A16,D2:D5)** and go for **CTRL+SHIFT+ENTER**.



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | **Stats Mark** | | LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 2 | 72 | | 60 | 70 | | | | |
| 3 | 74 | | 70 | 80 | | | | |
| 4 | 81 | | 80 | 90 | | | | |
| 5 | 78 | | 90 | 100 | | | | |
| 6 | 65 | | | | | | | |
| 7 | 93 | | | | | | | |
| 8 | 71 | | | | | | | |
| 9 | 65 | | | | | | | |
| 10 | 87 | | | | | | | |
| 11 | 86 | | | | | | | |
| 12 | 71 | | | | | | | |
| 13 | 71 | | | | | | | |
| 14 | 93 | | | | | | | |
| 15 | 99 | | | | | | | |
| 16 | 99 | | | | | | | |



=FREQUENCY(A2:A16,D2:D5

| | | | F | G | H |
|---|---|---|---|---|---|
| LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 60 | 70 | A2:A16,D2:D5 | | | |
| 70 | 80 | | | | |
| 80 | 90 | | | | |
| 90 | 100 | | | | |

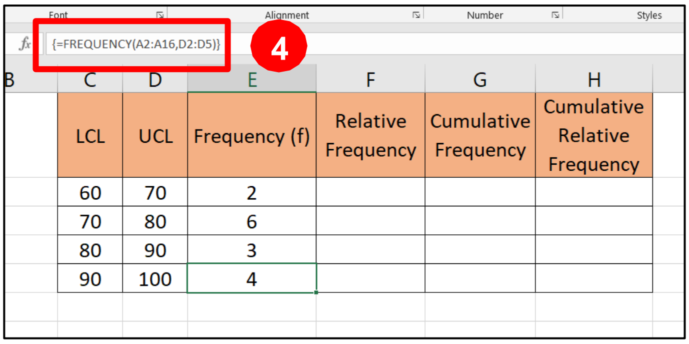
4 We have the frequency values in the column. Once we hit the **CTRL+SHIFT+ENTER** we can see the open and closing parenthesis in the function call.

5 Next, calculate the **Relative Frequency** in column F. The relative frequency of any class is the ratio of the frequency of that class to the total strength of the dataset, i.e, 2/15, 6/15, 3/15 and 4/15.

6 Now calculate the **Cumulative Frequency**. Cumulative frequency is the total number of all data points from minimum up to that class.

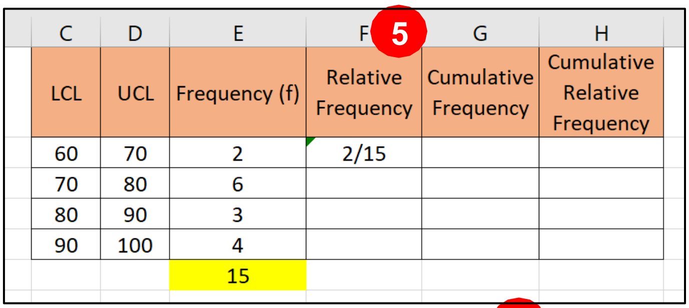
7 Lastly, calculate the **Cumulative Relative Frequency**. Cumulative relative frequency is the ratio of the cumulative frequency up to that class from minimum to the total number of all data points.

8 With this, we have converted the 15 data points to a frequency distribution with additional information like Relative Frequency, Cumulative Frequency and Cumulative Relative Frequency. Which can be used for further statistical analysis.

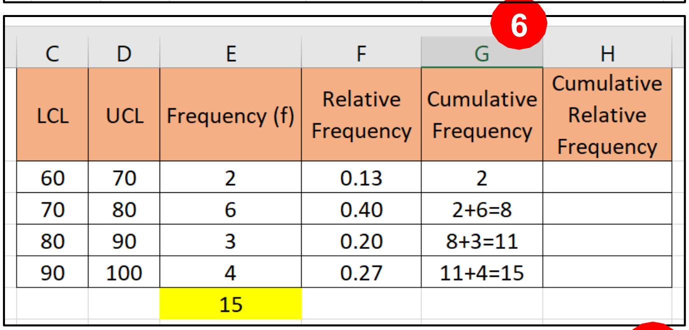


fx {=FREQUENCY(A2:A16,D2:D5)}

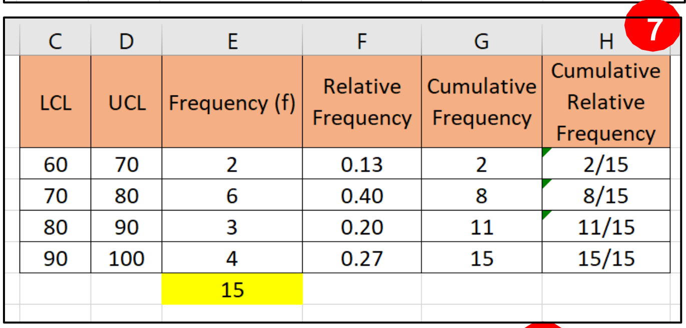
| B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|
| | LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| | 60 | 70 | 2 | | | |
| | 70 | 80 | 6 | | | |
| | 80 | 90 | 3 | | | |
| | 90 | 100 | 4 | | | |



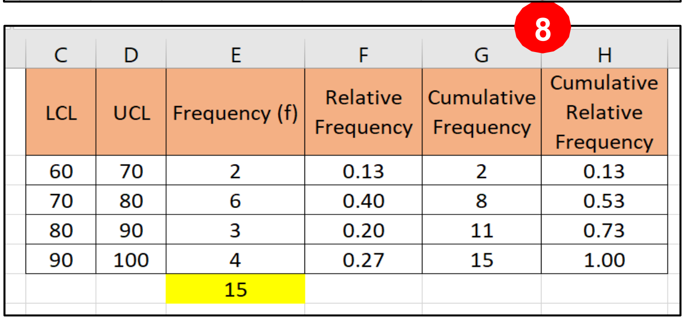
| C | D | E | F | G | H |
|---|---|---|---|---|---|
| LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 60 | 70 | 2 | 2/15 | | |
| 70 | 80 | 6 | | | |
| 80 | 90 | 3 | | | |
| 90 | 100 | 4 | | | |
| | | 15 | | | |



| C | D | E | F | G | H |
|---|---|---|---|---|---|
| LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 60 | 70 | 2 | 0.13 | 2 | |
| 70 | 80 | 6 | 0.40 | 2+6=8 | |
| 80 | 90 | 3 | 0.20 | 8+3=11 | |
| 90 | 100 | 4 | 0.27 | 11+4=15 | |
| | | 15 | | | |



| C | D | E | F | G | H |
|---|---|---|---|---|---|
| LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 60 | 70 | 2 | 0.13 | 2 | 2/15 |
| 70 | 80 | 6 | 0.40 | 8 | 8/15 |
| 80 | 90 | 3 | 0.20 | 11 | 11/15 |
| 90 | 100 | 4 | 0.27 | 15 | 15/15 |
| | | 15 | | | |



| C | D | E | F | G | H |
|---|---|---|---|---|---|
| LCL | UCL | Frequency (f) | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
| 60 | 70 | 2 | 0.13 | 2 | 0.13 |
| 70 | 80 | 6 | 0.40 | 8 | 0.53 |
| 80 | 90 | 3 | 0.20 | 11 | 0.73 |
| 90 | 100 | 4 | 0.27 | 15 | 1.00 |
| | | 15 | | | |

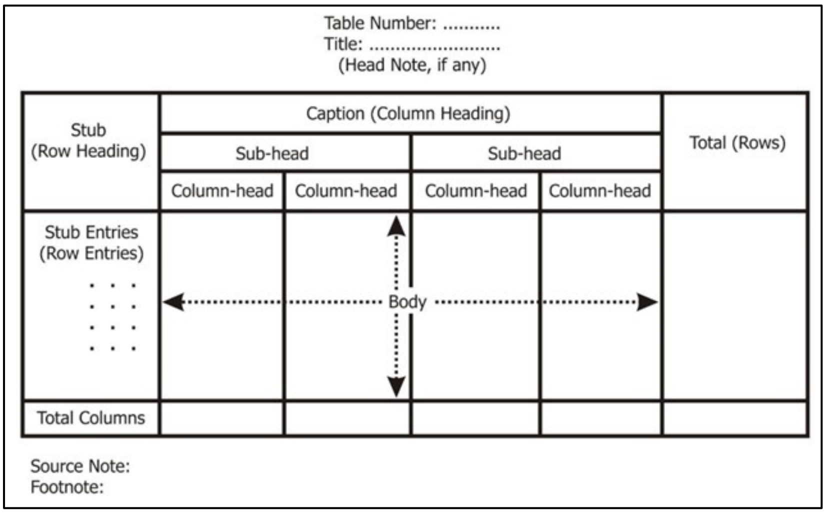
## *Quick Tip…*

***Array formulas*** are powerful formulas that enable you to perform complex calculations that often can't be done with standard worksheet functions. They are also referred to as "Ctrl-Shift-Enter" or "CSE" formulas, because you need to press Ctrl+Shift+Enter to enter them.

## Data Tabulation

12. Tabulation is another method of summarising and presenting data systematically in rows and columns. Tabulation helps in organising a set of data in an orderly manner to highlight its basic characteristics. It helps in the representation of even large amount of data in an engaging, easy to read and coordinated manner. The major objectives should be; ***Simplify the complex data, Depict Trend, Facilitates Comparison and Help as reference***.

13. Presenting the data in a tabular form is an art. The main parts of a Table are illustrated in the figure



| | |
|---|---|
| **Table Number.** Table number is the very first item mentioned on the top of each table for easy identification and further reference. | **Stubs or Row Headings.** The title of the horizontal rows is called "Stubs". |
| **Title.** Title of the table is the second item which is placed just above the table. It narrates about the contents of the table, so, it has to be very clear, brief and carefully worded. | **Body of the Table.** It contains the numeric information and reveals the whole story of investigated facts. Columns are read vertically from top to bottom and rows are read horizontally from left to right. |
| **Headnote.** It is the third item just above the Table & shown after the title. It gives information about the unit of data like, "Amount in Rupees", "Quantity in Tonnes" etc. It is generally given in brackets. | **Source Note.** It is a brief statement or phrase indicating the source of data presented in the table. |
| **Captions or Column Headings.** At the top of each column in a table, a column designation/head is given to explain figures of the column. This is called the column heading/Caption. | **Footnote.** It explains the specific feature of the table which is not self-explanatory and has not been explained earlier. For example, Points of exception if any. |

14. **Types of Tables**. The structure of tables can be classified based on ; ***(a) Objective and Scope of Investigation, (b) Nature of Data, and (c) Extent of data coverage***.

? **Self-Explore ….**
Familiarise with different types of Tables and its contextual usage.

## Graphical Presentation of Data

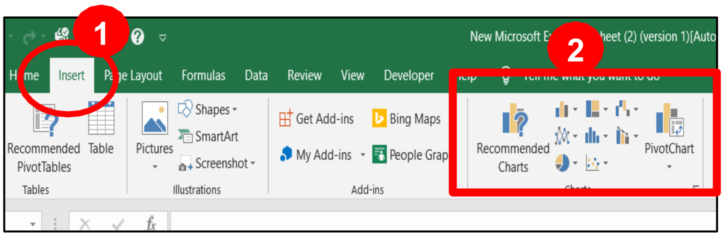
15. Graphical Representation is a way of analysing numerical data. In comparison to a tabular presentation, the graphic presentation of a frequency distribution facilitates easy understanding of the features of data and can be easily interpreted. The graphic presentation also serves as an easy technique for quick and effective comparison between two or more data distributions. It exhibits the relation between data, ideas, information and concepts in a diagram. It always depends on the type of information in a particular domain. There are different types of graphical representation. Some of them are as follows:-

| | |
|---|---|
| **Line Graphs**. Line graphs are used to display the continuous data and it is useful for predicting future events over time | **Ogives or Cumulative frequency graphs**. By plotting cumulative frequency against the respective class boundary, we get ogives. |
| **Bar Graphs**. Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities. Bar Charts can be of many types like Simple bar charts, Multiple bar charts, stacked bar charts, paired bar charts, deviation bar charts etc. | **Circle Graph or Pie Chart**. Shows the relationships between the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc. |
| **Histograms**. The graph that uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width. | **Stem and Leaf Plot**. In stem and leaf plot, the data are organised from the least value to the greatest value. The digits of the least place values from the leaves and the next place value digit forms the stems. |
| **Frequency Polygon**. Frequency Polygon is another method of representing frequency distribution graphically. | **Box and Whisker Plot**. The plot diagram summarises the data by dividing into four parts. Box and whisker show the range (spread) and the middle (median) of the data. |

16. **Charts in MS Excel**. MS Excel offers a large library of chart and graph types to help visually present your data. While multiple chart types might "work" for a given data set, it's important to select a chart type that best fits with the story you want the data to tell.

1 Once your data is highlighted in the Workbook, click the **Insert tab** on the top banner.

2 About halfway across the toolbar is a section with several **chart options**. Excel provides Recommended Charts based on popularity, but you can click any of the dropdown menus to select a different template.



**Self-Explore ….**
Familiarise with different types of Charts and its construction in MS Excel. Understand the difference between different visualisations and context.

# MEASURES OF CENTRAL TENDENCY

17. Frequency distribution and graphical representation make raw data more meaningful, yet this condensation at times may not be adequate and there may be a requirement to express the gist of data in a nutshell. A set of observations can be described satisfactorily in most cases by two characteristics: ***a measure of its location or central tendency and a measure of its dispersion or variability or spread***. Generally speaking, all data elements tend to cluster around a value in the middle region of its range and this central tendency is described by any of the following measures:-

(a) **Mean**. ***Mean or Arithmetic Mean or Average*** is defined as being equal to the sum of the numerical values of a series of items or numbers divided by the number of such items. In simple terms, mean is calculated by adding up all the numbers and divide by how many numbers there are. For example, the average of 2, 3, 3, 5, 7, and 10 is 30 divided by 6, which is 5.

The *sample mean* of the values $x_1, x_2, \dots, x_n$ is

$$\bar{x} = \frac{x_1 + x_2 + .. x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(b) **Median.** The median is defined as that value which exceeds the values of no more than one-half of the items and also is itself exceeded by no more than half of the items. In other words, the median is the value of the middle item when the series is arranged in ascending/ descending order. It is the central value in the sense that there as many values smaller than it as there are larger than it. For example, the median of 2, 3, 3, 5, 7, and 10 is 4.

All values known: if there are $n$ observations then the median is:

- the $\frac{n+1}{2}$ largest value, if $n$ is odd;
- the sample mean of the $\frac{n}{2}$ largest and the $\frac{n}{2} + 1$ largest values, if $n$ is even.

(c) **Mode**. The ***most frequently occurring value (number)***. The mode is defined as the value, which occurs most often and around which the values of the other items tend to cluster. If the data is grouped and represented pictorially as a histogram or a polygon, the mode will stand out as having the maximum height. For example, the mode of 2, 3, 3, 5, 7, and 10 is 3.

18. **Relationship between Mean, Median and Mode.** For a ***symmetrical distribution, the mean, median and mode all coincide and fall at the same value***. If the distribution is skewed, the mean gets shifted towards the skewness i.e. towards the right for right-skewed and to the left for left-skewed distributions. It is the mean that gets most affected as the extreme values shift it towards them. Median on the other hand is responsive only to the number of items considered and hence shifts towards the skewness but not as much as the mean. Mode, however, is not affected and remains the same. These three measures of central tendency are connected by the relation; **Mode = 3 Median – 2 Mean.**

**Self-Explore ….**
Understand the Merits, Demerits and characteristics of Mean, Median and Mode

19. **Partition Values – Quartiles, Deciles and Percentiles.** Median explore the characteristics of the data set by dividing the ordered data into two equal parts. However, to have more knowledge about the data set, we may decompose it into more parts of equal size. The measure of central tendency which is used for dividing the data into several equal parts is called ***partition values***. Two such methods are:-

(a) **Quartiles** - ***Fractiles that divide data into four equal parts***. The values which divide an ordered data set into four equal parts using three quartiles namely Q1, Q2 and Q3. The point which gives us 50% of the values to the left of it and 50% to the right of it is called the second quartile (Q2) or median.

(b) **Percentile** - ***Fractiles that divide data into 100 equal parts.*** The values which divide an ordered data set into 100 equal parts using 99 percentiles. The kth percentile is the value corresponding to the cumulative relative frequency of k/100 on the cumulative relative frequency diagram e.g. the 2nd percentile is the value corresponding to the cumulative relative frequency of 0.02. The 25th percentile is also known as the first quartile and the 75th percentile is also known as the third quartile.

20. **Calculation using MS Excel**. Calculation of above discussed three measures of central tendency can be easily calculated using the built-in functions in MS Excel. Consider the following example and illustration to understand the calculation of Mean, Median and Mode using MS Excel.

**Example**; Marks obtained by fifteen HDMC participants in Statistics final exam is given below.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 72 | 74 | 81 | 78 | 65 | 93 | 71 | 65 | 87 | 86 | 71 | 71 | 93 | 99 | 99 |

1. Find the Mean, Median and Mode of the marks.
2. Find the Q1, Q2 and Q3
3. Find the 25th Percentile, 50th Percentile, 75th Percentile and 99th Percentile

1 Enter the marks in cells A2 to A16

2 Enter the formula in Cell C3, C4 and C5 as under and enter:

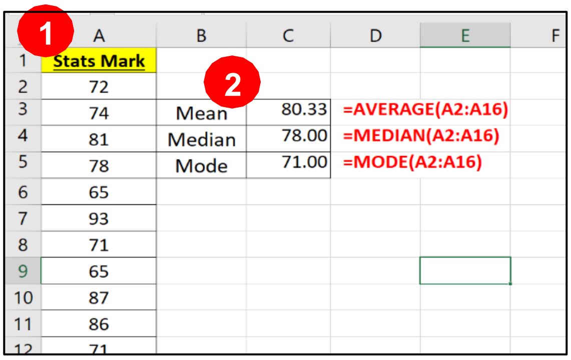
**=AVERAGE(A2:A16)**
**=MEDIAN(A2:A16)**
**=MODE(A2:A16)**

3 Enter the formula for **Quartiles** in Cell C3, C4 and C5 as under and enter:

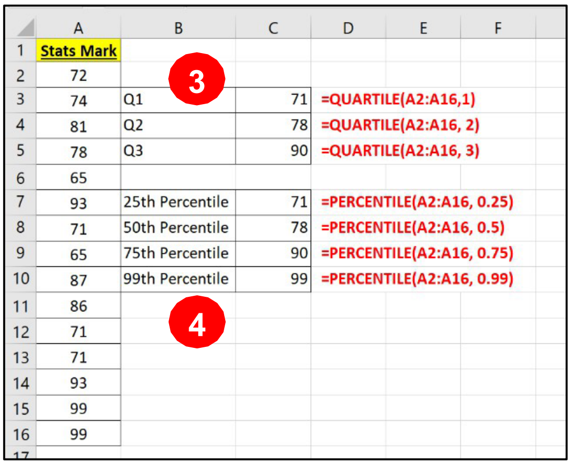
**= QUARTILE(A2:A16, 1) for Q1**
**= QUARTILE(A2:A16, 2) for Q2**
**= QUARTILE(A2:A16, 3) for Q3**

4 Enter the formula for **Percentile** in Cell C7, C8, C9 and C10 as under and enter:

**= PERCENTILE(A2:A16, 0.25)**
**= PERCENTILE(A2:A16, 0.50)**
**= PERCENTILE(A2:A16, 0.75)**
**= PERCENTILE(A2:A16, 0.99)**



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Stats Mark | | | | | |
| 2 | 72 | | | | | |
| 3 | 74 | Mean | 80.33 | =AVERAGE(A2:A16) | | |
| 4 | 81 | Median | 78.00 | =MEDIAN(A2:A16) | | |
| 5 | 78 | Mode | 71.00 | =MODE(A2:A16) | | |
| 6 | 65 | | | | | |
| 7 | 93 | | | | | |
| 8 | 71 | | | | | |
| 9 | 65 | | | | | |
| 10 | 87 | | | | | |
| 11 | 86 | | | | | |
| 12 | 71 | | | | | |



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Stats Mark | | | | | |
| 2 | 72 | | | | | |
| 3 | 74 | Q1 | 71 | =QUARTILE(A2:A16,1) | | |
| 4 | 81 | Q2 | 78 | =QUARTILE(A2:A16, 2) | | |
| 5 | 78 | Q3 | 90 | =QUARTILE(A2:A16, 3) | | |
| 6 | 65 | | | | | |
| 7 | 93 | 25th Percentile | 71 | =PERCENTILE(A2:A16, 0.25) | | |
| 8 | 71 | 50th Percentile | 78 | =PERCENTILE(A2:A16, 0.5) | | |
| 9 | 65 | 75th Percentile | 90 | =PERCENTILE(A2:A16, 0.75) | | |
| 10 | 87 | 99th Percentile | 99 | =PERCENTILE(A2:A16, 0.99) | | |
| 11 | 86 | | | | | |
| 12 | 71 | | | | | |
| 13 | 71 | | | | | |
| 14 | 93 | | | | | |
| 15 | 99 | | | | | |
| 16 | 99 | | | | | |

# MEASURES OF DISPERSION

21. The measure of central tendency describes that the values in the data set tend to cluster around a central value called, average. However, these measures do not reveal how the values are spread (dispersed or scattered) on each side of the central value. ***Dispersion is the amount of variation, scatter or spread in data***. A measure of dispersion is a value which indicates the degree of variability of data. Knowledge of the variability may be of interest in itself but more often is required to decide how precisely the sample mean – and an estimator of the mean - reflects the population mean. A low degree of variation implies greater uniformity.

**An example;** the mean score of two units in a firing exercise may be the same i.e. 74. But in one, on closer scrutiny, we may find an overwhelming majority have scored around 74 whereas in the second unit, we find a wide difference in scores, with a sizeable number scoring very low or very high. Describing the state of performance in these two units through the mean alone would conceal the wide difference in their states of training.

22. Some important Measures of Dispersion are:-

(a) **Range**. The range of a set of data is the difference between the largest and lowest observed values in a data set or the interval between these values. It is a measure of the spread of the data. For example, the range of 2, 3, 3, 5, 7, and 10 is 8, i.e 10-2.

(b) **Inter Quartile Range (IQR)**. The inter-quartile range also called the mid spread is the difference between the third and first quartile in a set of data. This measure considers the spread in the middle 50% of the data; therefore, it is not influenced by extreme values. ***IQR = 3$^{rd}$ Quartile – 1$^{st}$ Quartile (Q3-Q1).***

(c) **Variance**. Variance is a measure of variability based on the squared deviation of the observed values in the data set from its mean. Since variance is the average of the squared deviation from the mean, it is also called the ***Mean Square Average***.

(d) **Standard Deviation**. Or ***Root Mean Square Deviation*** is a kind of average of the deviations from the mean. The Standard Deviation (SD) is the most important and mathematically precise measure of variability. It is defined as the root mean square value of the differences of individual elements from the arithmetic mean. ***SD is nothing but the positive square root of the variance***.

23. **Measure of Relative Dispersion – Coefficient of Variation (CV)**. The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ***ratio of the standard deviation to the mean***, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. CV can also be regarded as SD expressed as a percentage of the mean of any distribution. The distribution for which the CV is greater is said to be more variable or less consistent (less uniform or less homogeneous).

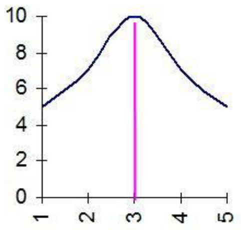
# MEASURES OF SYMMETRY AND SHAPE

24. The average and measure of dispersion can describe the distribution but they are not sufficient to describe the nature of the distribution. For this purpose, we use other concepts known as Skewness and Kurtosis.

(a) **Skewness**. Skewness means lack of symmetry. A distribution is said to be symmetrical when the values are uniformly distributed around the mean. In a symmetrical distribution the mean, median and mode coincide, that is, ***mean = median = mode***. Several measures are used to express the direction and extent of skewness of a dispersion. The first one is the Coefficient of Skewness:

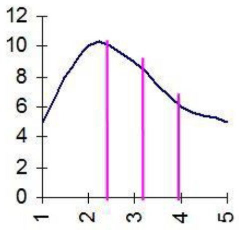
$$\mathbf{S_k} = \frac{\mathbf{3(mean - meadian)}}{\mathbf{Standard\ Deviation}}$$

(b) For a symmetric distribution Sk = 0. If the distribution is negatively skewed then Sk is negative and if it is positively skewed then Sk is positive. The range for Sk is from -3 to 3.



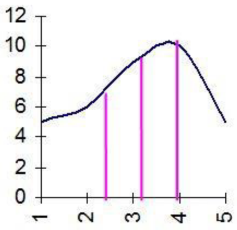
Mean = Median = Mode

**Symmetrical**



Mode >Med> Mean

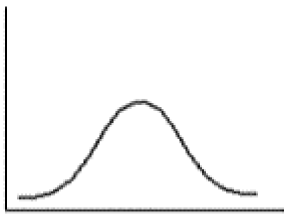
**Positively Skewed**



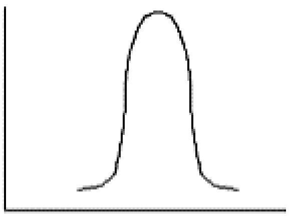
Mean< Med<Mode

**Negatively Skewed**

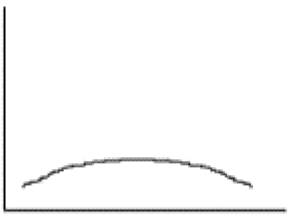
(c) **Kurtosis.** A measure of the peakedness or convexity of a curve is known as Kurtosis. It is the degree of flatness or peakedness in the region around the mode of the frequency curve. A positive value tells you that you have heavy tails (i.e. a lot of data in your tails) and a negative value means that you have light-tails (i.e. little data in your tails).



Mesokurtic Curve



Leptokurtic Curve



Platykurtic Curve

(i) Distribution in which value of observations clusters heavily in the centre is peaked or ***leptokurtic***. A distribution with kurtosis >3 (excess kurtosis >0).

(ii) Flat distribution, with values of observations more evenly distributed and tails flatter than the normal distribution is called ***platykurtic***. A distribution with kurtosis <3 (excess kurtosis <0).

(iii) A distribution that is almost normal, neither too peaked not too flat, is called ***mesokurtic***. A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0).

25. **<u>Calculation using MS Excel</u>**. Measures of Dispersion, Symmetry and shape can be calculated individually using MS Excel as illustrated below:-

**<u>Example</u>**; Marks obtained by fifteen HDMC participants in Statistics final exam is given below. Calculate dispersion, skewness and kurtosis.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 72 | 74 | 81 | 78 | 65 | 93 | 71 | 65 | 87 | 86 | 71 | 71 | 93 | 99 | 99 |

1 Enter the marks in cells A2 to A16

2 Calculate the following using the excel functions:

Minimum > **=MIN(A2:A16)**
Maximum > **=MAX(A2:A16)**

Calculate the **Range**, which is **Maximum – Minimum.** In this case, it is Cell D3-D2.

3 Calculate the following using the excel functions:

Quartile 1 > **=QUARTILE(A2:A16, 1)**
Quartile 3 > **=QUARTILE(A2:A16, 3)**

Calculate the **Inter Quartile Range (IQR)**, which is **Q3-Q1.** In this case, it is Cell D7-D6.

4 Calculate the **Variance** and **Standard Deviation** using the excel built-in functions as under:

Variance > **=VAR(A2:A16)**
SD > **=STDEV(A2:A16)**

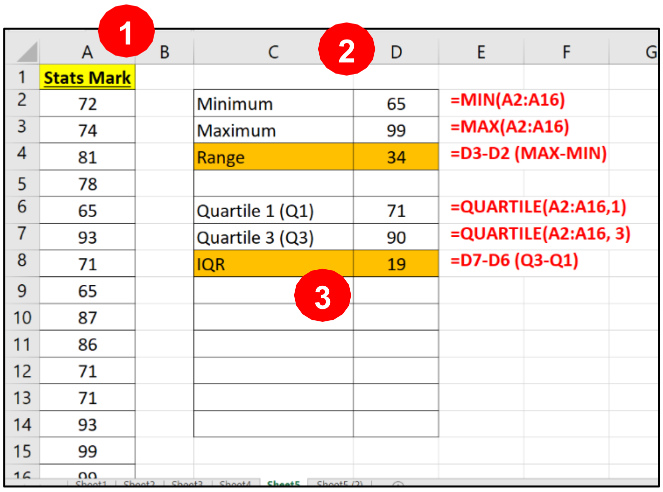
5 Calculate the Mean and Coefficient of Variance using the following formula:

SD > **=STDEV(A2:A16)**
Mean > **=AVERAGE(A2:A16)**

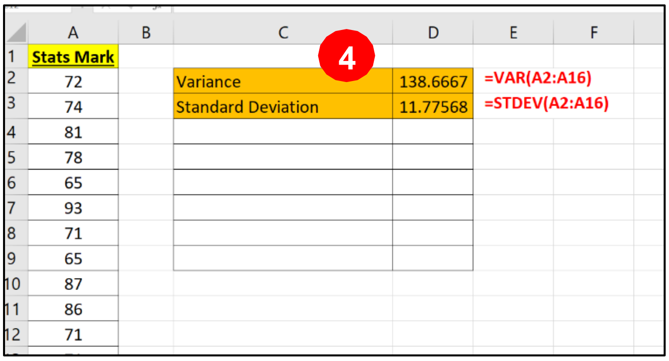
Calculate the **Coefficient of Variation (CV)** as [(SD/Mean) x 100]. In this case the formula is **D2/D3*100**.

6 Calculate the Skewness and Kurtosis using excel built-in functions as under:

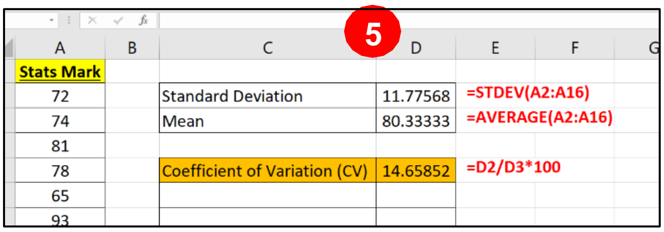
Skewness > **=SKEW(A2:A16)**
Kurtosis > **=KURT(A2:A16)**



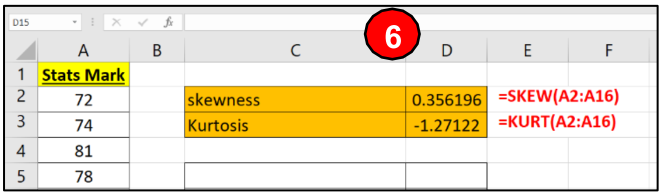
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Stats Mark | | | | |
| 2 | 72 | | Minimum | 65 | =MIN(A2:A16) |
| 3 | 74 | | Maximum | 99 | =MAX(A2:A16) |
| 4 | 81 | | Range | 34 | =D3-D2 (MAX-MIN) |
| 5 | 78 | | | | |
| 6 | 65 | | Quartile 1 (Q1) | 71 | =QUARTILE(A2:A16,1) |
| 7 | 93 | | Quartile 3 (Q3) | 90 | =QUARTILE(A2:A16, 3) |
| 8 | 71 | | IQR | 19 | =D7-D6 (Q3-Q1) |
| 9 | 65 | | | | |
| 10 | 87 | | | | |
| 11 | 86 | | | | |
| 12 | 71 | | | | |
| 13 | 71 | | | | |
| 14 | 93 | | | | |
| 15 | 99 | | | | |
| 16 | 99 | | | | |



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Stats Mark | | | | |
| 2 | 72 | | Variance | 138.6667 | =VAR(A2:A16) |
| 3 | 74 | | Standard Deviation | 11.77568 | =STDEV(A2:A16) |
| 4 | 81 | | | | |
| 5 | 78 | | | | |
| 6 | 65 | | | | |
| 7 | 93 | | | | |
| 8 | 71 | | | | |
| 9 | 65 | | | | |
| 10 | 87 | | | | |
| 11 | 86 | | | | |
| 12 | 71 | | | | |



| A | B | C | D | E |
|---|---|---|---|---|
| Stats Mark | | | | |
| 72 | | Standard Deviation | 11.77568 | =STDEV(A2:A16) |
| 74 | | Mean | 80.33333 | =AVERAGE(A2:A16) |
| 81 | | | | |
| 78 | | Coefficient of Variation (CV) | 14.65852 | =D2/D3*100 |
| 65 | | | | |
| 93 | | | | |



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Stats Mark | | | | |
| 2 | 72 | | skewness | 0.356196 | =SKEW(A2:A16) |
| 3 | 74 | | Kurtosis | -1.27122 | =KURT(A2:A16) |
| 4 | 81 | | | | |
| 5 | 78 | | | | |

## Descriptive Statistics Using MS Excel Data Analysis Tool

26. MS Excel provides an easy to use Data Analysis Toolpak to develop complex statistical or engineering analysis. You provide the data and parameters for each analysis, and the tool uses the appropriate statistical or engineering macro functions to calculate and display the results in an output table. Some tools generate charts in addition to output tables. The Descriptive Statistics analysis tool generates a report of univariate statistics for data in the input range, providing information about the central tendency and variability of your data. ***The following step will enable Data Analysis ToolPak in your Excel***.

1 Go to **File > Options**

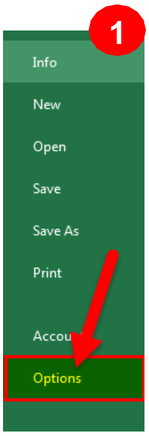
2 Go to **Add-ins**

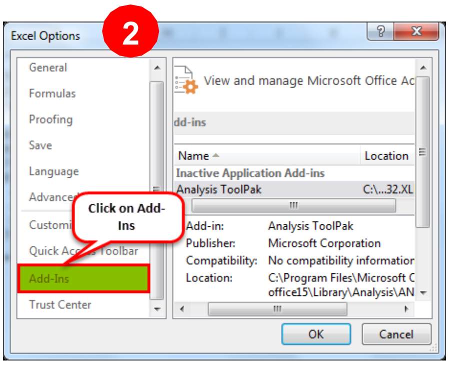
3 Under **Add-ins** on the right-hand side you will see all the inactive Applications**. Select Analysis Toolpak and click on GO**.

4 Now you will have the add-ins available for your excel. **Select Analysis Toolpak and click on OK**.

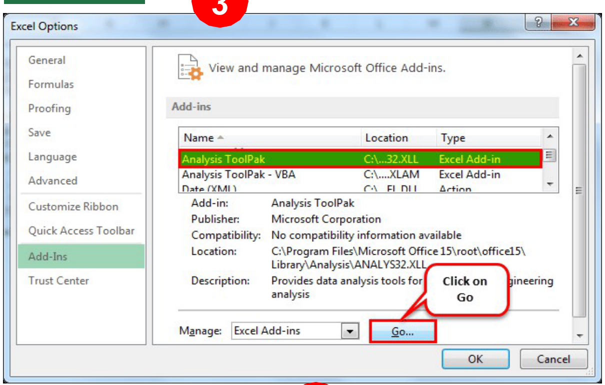
5 Now you must see the **Data Analysis option under the Data tab**.

6 Click on **Data Analysis you will see all the available analysis techniques** including Descriptive Statistics and many more under this tool.

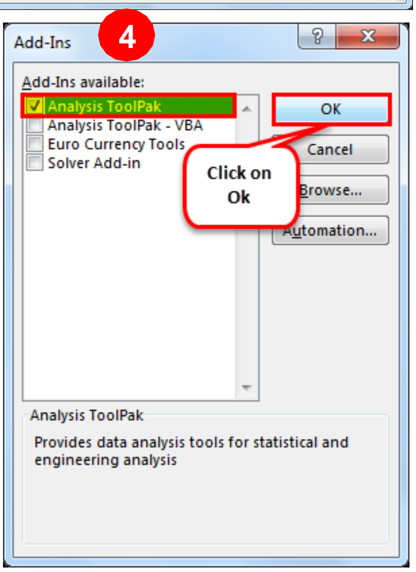



Excel Options
2
General
Formulas
Proofing
Save
Language
Advance
Customi
Quick Ac Toolbar
Add-Ins
Trust Center
View and manage Microsoft Office Ac
dd-ins
Name
Location
Inactive Application Add-ins
Analysis ToolPak
C:\...32.XL
Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information
Location: C:\Program Files\Microsoft C office15\Library\Analysis\AN
OK
Cancel

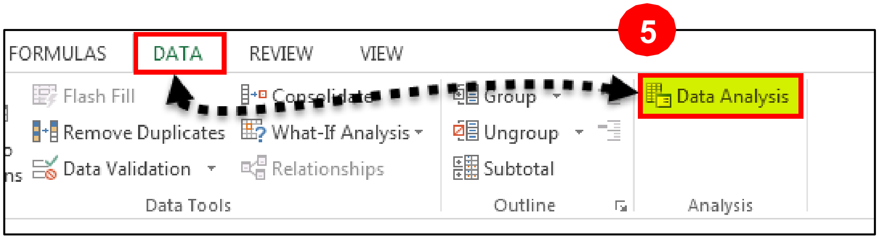
Click on Add-Ins

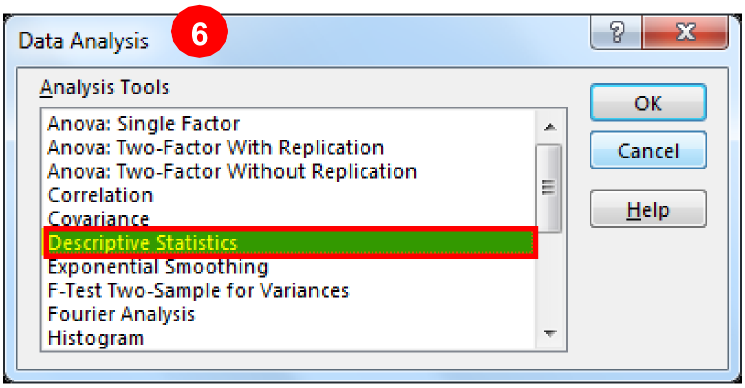


Click on Go



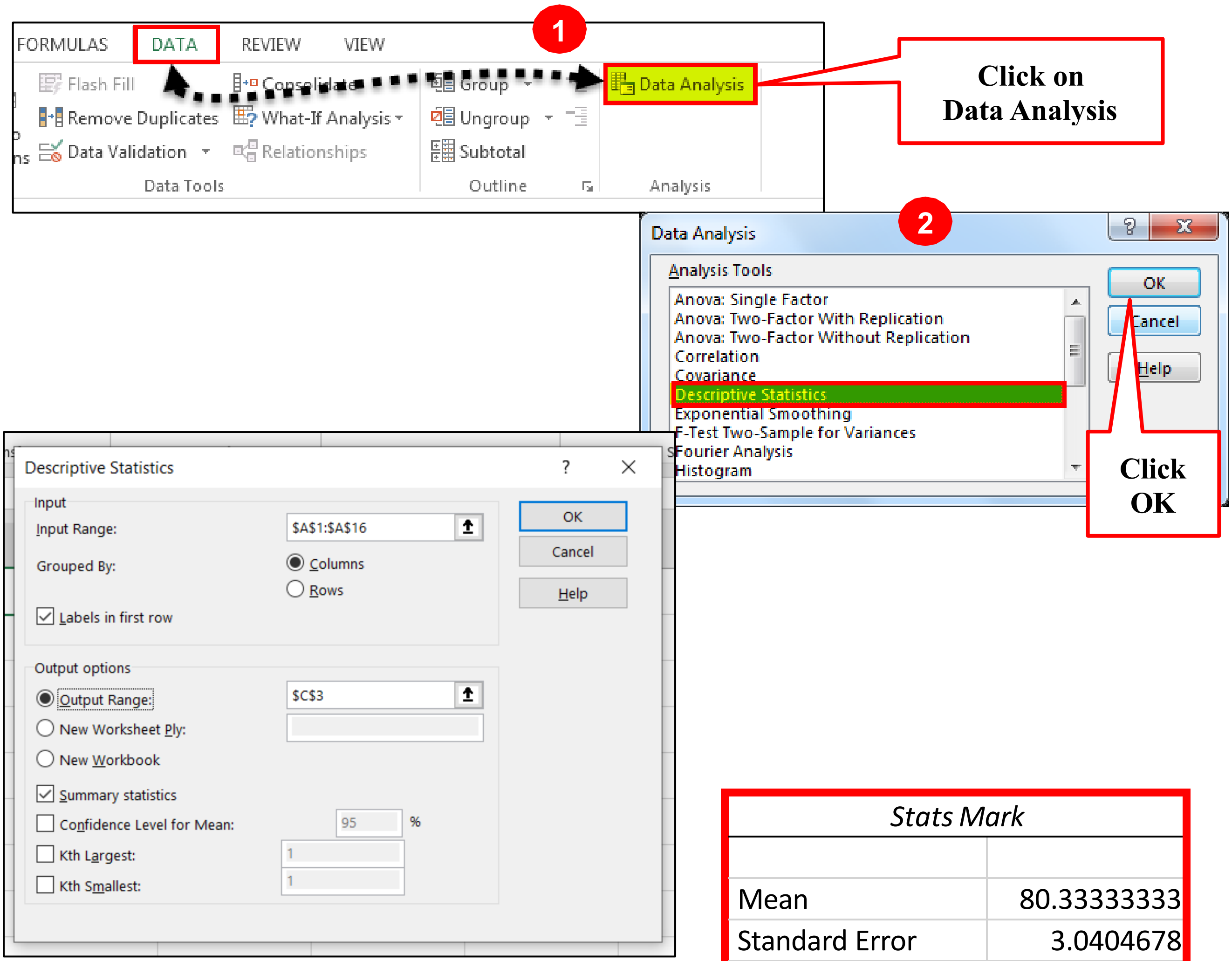
Click on Ok





**Example**; Marks obtained by fifteen HDMC participants in Statistics final exam is given below. Carry out Descriptive analysis and report the findings.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 72 | 74 | 81 | 78 | 65 | 93 | 71 | 65 | 87 | 86 | 71 | 71 | 93 | 99 | 99 |



**Click on Data Analysis**

**Click OK**

Descriptive Statistics

Input

Input Range: $A$1:$A$16

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $C$3

○ New Worksheet Ply:

○ New Workbook

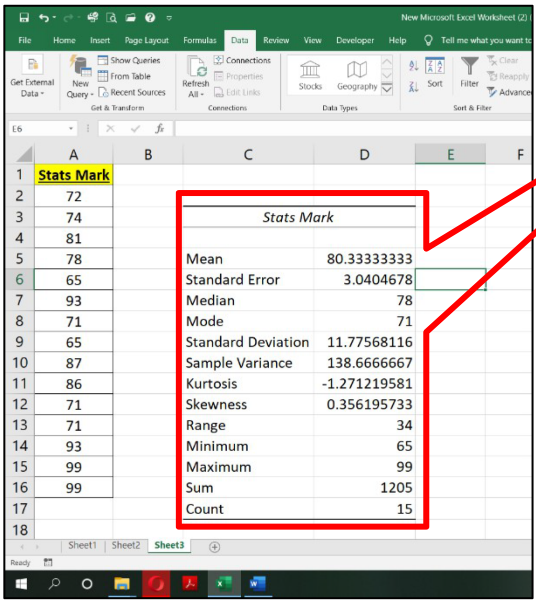
☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

OK | Cancel | Help



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Stats Mark | | | |
| 2 | 72 | | | |
| 3 | 74 | | Stats Mark | |
| 4 | 81 | | | |
| 5 | 78 | | Mean | 80.33333333 |
| 6 | 65 | | Standard Error | 3.0404678 |
| 7 | 93 | | Median | 78 |
| 8 | 71 | | Mode | 71 |
| 9 | 65 | | Standard Deviation | 11.77568116 |
| 10 | 87 | | Sample Variance | 138.6666667 |
| 11 | 86 | | Kurtosis | -1.271219581 |
| 12 | 71 | | Skewness | 0.356195733 |
| 13 | 71 | | Range | 34 |
| 14 | 93 | | Minimum | 65 |
| 15 | 99 | | Maximum | 99 |
| 16 | 99 | | Sum | 1205 |
| 17 | | | Count | 15 |
| 18 | | | | |

| *Stats Mark* | |
|---|---|
| | |
| Mean | 80.33333333 |
| Standard Error | 3.0404678 |
| Median | 78 |
| Mode | 71 |
| Standard Deviation | 11.77568116 |
| Sample Variance | 138.6666667 |
| Kurtosis | -1.271219581 |
| Skewness | 0.356195733 |
| Range | 34 |
| Minimum | 65 |
| Maximum | 99 |
| Sum | 1205 |
| Count | 15 |

## References

1. Levin, Rubin, Rastogi, Siddique. Statistics for Management, 8th Edition, 2017 Pearson India. (ISBN No.9789332581180).

2. JK Sharma. Business Statistics, 4th Edition, 2018 Vikas Publishing House Pvt Ltd (ISBN No. 9789325980815)

3. McClave, James T.; Benson, P. George; Sincich, Terry. Statistics for Business and Economics 13th Edition, 20147 Pearson India (ISBN No. 9780134506593)

4. https://www.youtube.com/watch?v=h8EYEJ32oQ8&list=PLU5aQXLWR3_yYS0ZYRA-5g5YSSYLNZ6Mc

5. https://www.youtube.com/watch?v=E4HAYd0QnRc

6. https://www.youtube.com/watch?v=QoQbR4lVLrs

7. https://www.youtube.com/watch?v=5MFjwM6K5Sg

## Self-Assessment Exercise

The OPD sick report records of MH Secunderabad for past one month is appended below. You are to provide:-

(a) Classification of data for further analysis and make deductions.
(b) A complete descriptive analysis of the data and make deductions.
(c) Effective presentation of the data using Tables and Charts of appropriate design.

**NUMBER OF OPD SICK REPORT FOR THE MONTH OF APR 2024**

| *Sunday* | *Monday* | *Tuesday* | *Wednesday* | *Thursday* | *Friday* | *Saturday* |
|---|---|---|---|---|---|---|
| | | | 180 | 291 | 346 | 120 |
| 43 | 321 | 221 | 142 | 201 | 289 | 132 |
| 31 | 286 | 201 | 136 | 245 | 240 | 168 |
| 22 | 240 | 239 | 182 | 262 | 231 | 112 |
| 16 | 221 | 264 | 199 | 240 | | |