UNIT-2 SUMMARIZATION OF DATA (08 HOURS)

Statistics have majorly categorised into two types:

1. Descriptive statistics
2. Inferential statistics

Descriptive Statistics

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures.

For example, the collection of people in a city using the internet or using Television

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation.

Inferential Statistics

Inferential Statistics is a method which allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements which goes beyond the available data or information. For example, deriving estimates from hypothetical research.

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Data tabulation

Tabulation is a process of systematic arrangement of the classified data in rows and columns, in the form of table.

Example 1:

Number of oranges in the box	5	6	7	8	9	10	Total
Number of boxes	5	8	10	6	3	13	45

Example 2:

Height (cm)	140-150	150-160	160-170	170-180	Total
No. of students	6	24	18	2	50

The above two types example are **Frequency Distribution or Frequency table.**

Frequency distribution is a systematic presentation of the values taken by a variable along with their frequencies. **Frequency** refers to the number of times an observation is repeated.

The number of observations corresponding to a particular class is known as **class frequency**. Class frequency is a positive integer including zero.

From the above examples, we can explain the following terms:

Number of oranges per boxes, height are **variables** and number of boxes, number of students are **frequencies**.

While framing a frequency distribution, if class intervals are not considered, then it is called **Discrete frequency distribution (Ex.1).**

While framing a frequency distribution, if class intervals are considered, then it is called **Continuous frequency distribution (Ex.2).**

Formation of Discrete frequency distribution:

For formation of frequency distribution, three columns are formed

- Variable
- Tally bars
- Frequency.

In the first column, values of given variable are written without repetition in an order. For each value a tally/stroke is marked against that value in the second column. In this way tally scores are marked for all values. For easy counting the tallies are put as a group of 5 (~~IIII~~). Finally count the number of tally bars corresponding to each value of the variable in third column. It is known as frequency. The total frequency (N) is equal to the total number of observations.

Example 3.

In survey of 40 families in Haveri, the number of children per family was recorded and the following data were obtained.

1,0,3,2,1,5,6,2,2,1,0,3,4,2,1,6,3,2,1,5,3,3,2,4,2,2,3,0,2,1,4,5,3,3,4,4,1,2,4,5.

Represent the data in the form of a discrete frequency distribution.

Sol: Frequency distribution of the number of children.

Number of children (x)	Tally Marks	Frequency (f)
0	III	3
1	~~IIII~~ II	7
2	~~IIII~~ ~~IIII~~	10
3	~~IIII~~ III	8
4	~~IIII~~ II	6
5	IIII	4
6	II	2
	Total	40

Formation of continuous frequency distribution:

Suitable class intervals are formed on the basis of the magnitude of the data. For each value a tally mark is marked against the class in which it falls. This process is continued until all the values are exhausted. The tallies of each class are counted and written as frequency of that class.

To construct a continuous frequency distribution table, it essential to know the following factors:

Range: It is the difference highest and lowest value in the data
i.e., Range = H.V- L.V.

Class:
The sub range is called class.

Class limits:
The lowest and highest values which are taken to define the boundaries of a class are class limits. The lowest value is called lower limit (L.L) and the highest value is upper limit (U.L).
Example: (30 – 40), (40-50) … are class limits.
The lowest values of the class are 30-40 they are lower limits and the highest value of the class are 40-50 they are upper limits.

Inclusive class:
In a class, if lower as well as upper limits are included in the same class, such a class is called Inclusive class. Here, upper limit of a class is not equal to the lower limit of the next class.
Ex. 0-9, 10-19, 20-39…. Are inclusive classes.

Exclusive class:
In a class, If the lower limit is included in the same class and upper limit is excluded from that class but included in the next class, such a class is called Exclusive class. Here, upper limit of a class is equal to lower limit of the next class.
Ex: 30-40, 40-50, 50-60 are exclusive classes.

Correction factor:
It is half of the difference between lower limit of a class and upper limit of the preceding class. thus,

Correction factor (C.F) = $\frac{lower\ limit of\ class - upper\ limit\ of\ the\ preceding\ class}{2}$

To get exclusive class intervals from inclusive class intervals, add C.F from all lower limits.

Ex: Convert the flowing inclusive class intervals to exclusive class intervals.

C - I	10 -19	20 - 29	30 - 39	40-49

C.F = $\frac{20-19}{2} = \frac{1}{2} = 0.5$

By subtracting it from all lower limits we get lower limit as 9.5, 19.5,29.5, 39.5 and adding it to all upper limits we get upper limit as 19.5, 29.5, 39.5, 49.5.
Therefore, the exclusive class intervals are 9.5 -19.5, 19.5 -29.5, 29.5 - 39.5 and 39.5 - 49.5.

Open- end classes

In a class, if the lower or upper limit of the class is not specified such a class is called open-end class.

For example: less than (below) or more than (above) a particular class limit.

The frequency distribution based on open-end classes is called open end frequency distribution.

Ex:

Class interval	Frequency
Less than 20	8
20 – 30	15
30 - 40	23
40 - 50	12
50 - 60	9
More than 60	3

Mid-point (class mark)

The central value of a class is called mid-point or class mark. It is the average of class limits.

i.e., m or x = $\frac{lower\ limit + upper\ limit}{2}$

Ex: Midpoint of the class (10 – 20) is,

$$M = \frac{LL+UL}{2} = \frac{10+20}{2} = 15$$

Width (size) of the class:

The difference between the upper and lower limits of a class is called width of the class**.** It is denoted by C or I.

For example, the width of the class intervals (30 – 40) is 40 – 30 = 10.

Number of classes:

The number of classes can be obtained by using the Prof. Sturge's Rule

Number of classes (K) = 1 + 3.322 log N; where N: Number of observations.

The width of the class can also be obtained by:

Width of the class = C = $\frac{Range}{Number\ of\ classes(k)}$

Cumulative frequency

The added-up frequencies are called cumulative frequencies.

There are two types of cumulative frequencies:

1. Less than type
2. More than type

The number of observations (frequencies) below a certain limit is less than cumulative frequency (L.C.F). The frequency distribution formed for less than cumulative frequencies against upper class limits, is, less than cumulative frequency distribution.

Ex:

Frequency Distribution	
Weight (Kg)	Number of persons
30-40	10
40- 50	15
50-60	20
60-70	15

Less than cumulative frequency distribution	
Weight (kg)	Number of persons
Less than 40	10
Less than 50	(10+15)=25
Less than 60	(25+20)=45
Less than 70	(15+45)=60

The number of frequencies above a certain limit is more than cumulative frequency. The frequency distribution formed for more than cumulative frequencies against lower class limit is more than cumulative frequency distribution.

Frequency Distribution	
Weight (Kg)	Number of persons
30-40	10
40- 50	15
50-60	20
60-70	15

More than cumulative frequency distribution	
Weight (kg)	Number of persons
More than 30	(50+10)=60
More than 40	(45+15)=50
More than 50	(15+20)=45
More than 60	15

Frequency density:

The frequency per unit of class interval is the frequency density (f/c). or It is the ratio of the class frequency to the width of the class interval.

i.e., Frequency density = $\frac{Frequency\ of\ the\ class}{width\ of\ the\ class}$

It is used to compare the concentration of the frequencies of different classes for a given frequency distribution.

Weight (Kg)	Number of persons(f)	Width of the class	Frequency density(f/c)
0-10	10	10	$\frac{10}{10}=1$
10- 30	15	20	.75
30-50	40	20	2
50-60	45	10	4.5
60-65	20	5	5

Relative frequency (relative frequency table)

Relative frequency is the ratio of frequency of the value of the variable to the total frequency.

i.e. Relative frequency (R.f) = $\frac{frequency\ of\ the\ value\ of\ the\ variable}{total\ frequency}$

Ex:

No of apples Per box	No of boxes	Relative Frequency R.f=f/N
5	5	5/45 = 0.111
6	8	8/45 = 0.178
7	13	13/45 = 0.289
8	10	10/45 = 0.222
9	6	6/45 = 0.133
10	3	3/45 = 0.067
Total (N)	45	1

Grouped Data:

The data formed by aggregating individual observations of a variable into groups, so that a frequency distribution of these groups serves as a convenient means of summarizing or analysing the data is called as Grouped Data.

The following are the different tools or methods using for grouping Data.

1) **Bar Graph:** The pictorial representations of a grouped data, in the form of vertical or horizontal rectangular bars, where the lengths of the bars are equivalent to the measure of data, are known as bar graphs or bar charts. The bars drawn are of uniform width, and the variable quantity is represented on one of the axes. Also, the measure of the variable is depicted on the other axes. The heights or the lengths of the bars denote the value of the variable, and these graphs are also used to compare certain quantities. The frequency distribution tables can be easily represented using bar charts which simplify the calculations and understanding of data.

Types of Bar Charts: The bar graphs can be vertical or horizontal. The primary feature of any bar graph is its length or height. If the length of the bar graph is more, then the values are greater of any given data. Bar graphs normally show categorical and numeric variables arranged in class intervals. They consist of an axis and a series of labelled horizontal or vertical bars. The bars represent frequencies of distinctive values of a variable or commonly the distinct values themselves. The number of values on the x-axis of a bar graph or the y-axis of a column graph is called the scale.

The types of bar charts are as follows:

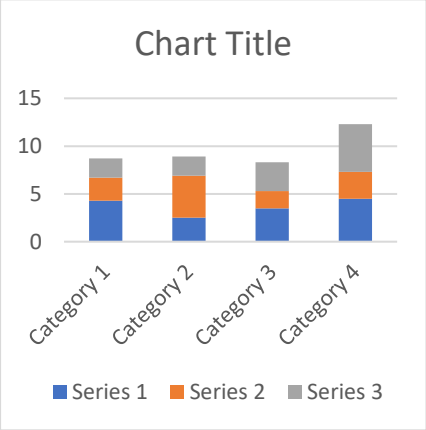
1. Vertical bar chart
2. Horizontal bar chart

Even though the graph can be plotted using horizontally or vertically, the most usual type of bar graph used is the vertical bar graph. The orientation of the x-axis and y-axis are changed depending on the type of vertical and horizontal bar chart. Apart from the vertical and horizontal bar graph, the two different types of bar charts are:

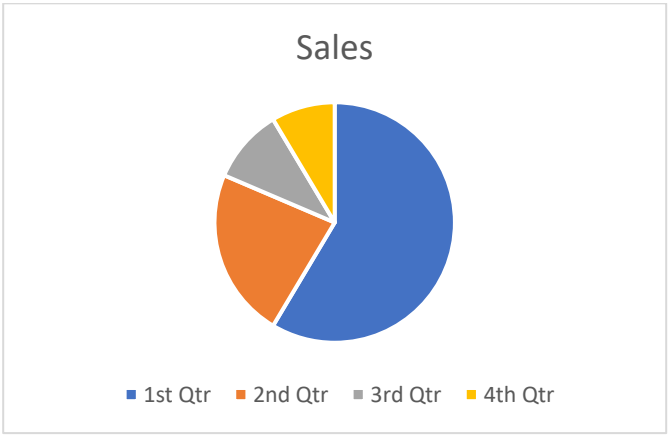
Uses of Bar Graphs:

Bar graphs are used to match things between different groups or to trace changes over time. Yet, when trying to estimate change over time, bar graphs are most suitable when the changes are bigger.

Bar charts possess a discrete domain of divisions and are normally scaled so that all the data can fit on the graph. When there is no regular order of the divisions being matched, bars on the chart may be organized in any order. Bar charts organized from the highest to the lowest number are called Pareto charts.

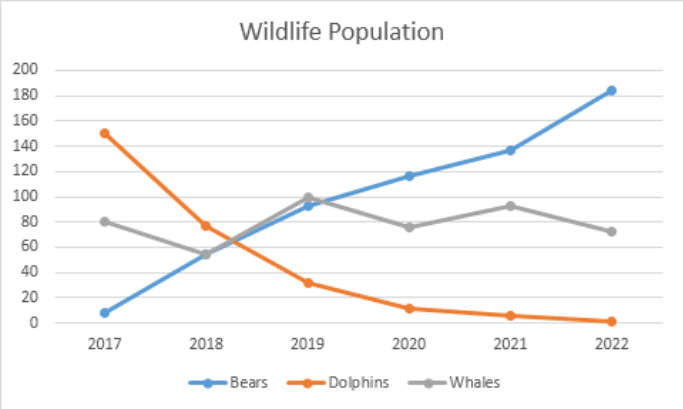


2) Pie Chart: The "pie chart" also is known as "circle chart", that divides the circular statistical graphic into sectors or slices in order to illustrate the numerical problems. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie-chart works the best at that time. In most of the cases, pie charts replace some other graphs like the bar graph, line plots, histograms etc. The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sectors of a pie chart forms a certain portion of the total(percentage). The total of all the data is equal to 360°.

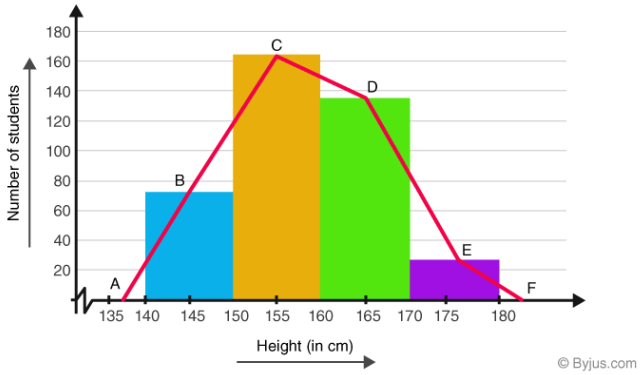


3) Line Graph: A line graph is a type of chart used to show information that changes over time. We plot line graphs using several points connected by straight lines. We also call it a line chart. The line graph comprises of two axes known as 'x' axis and 'y' axis.

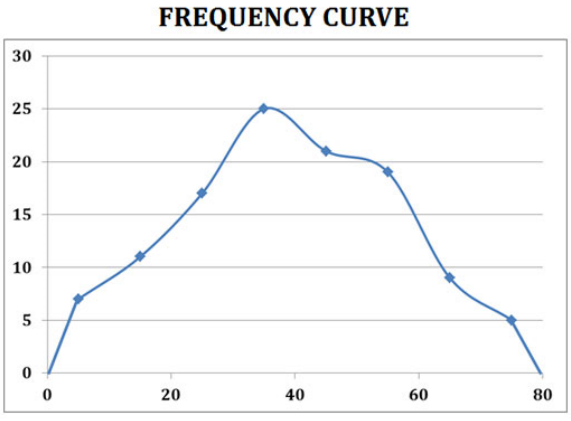
- The horizontal axis is known as the x-axis.
- The vertical axis is known as the y-axis.



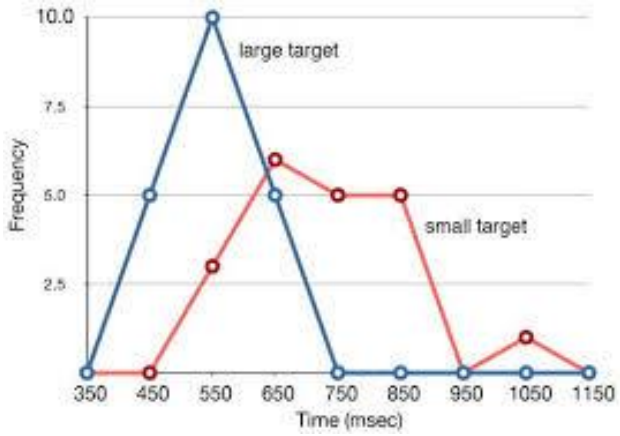
4) Frequency Polygon: A frequency polygon is a graph constructed by using lines to join the midpoints of each interval, or bin. A frequency polygon is almost identical to a histogram, which is used to compare sets of data or to display a cumulative frequency distribution. It uses a line graph to represent quantitative data.



5) Frequency Curve: Smooth free curve moving around the frequency polygon is known as frequency curve. A frequency-curve is a smooth curve for which the total area is taken to be unity. It is a limiting form of a histogram or frequency polygon. The frequency curve for a distribution can be obtained by drawing a smooth and free hand curve through the midpoints of the upper sides of the rectangles forming the histogram.



6) Relative Frequency Polygon: A frequency polygon is a type of chart that helps us visualize a distribution of values. For Relative frequency polygon refer above mentioned steps of relative frequency distribution, calculate the mid points of class intervals and frequency. Using these draw frequency polygons, it gives Relative frequency polygon.



7) Histogram: A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart, but there are important differences between them. A histogram is a common data analysis tool in the business world. It's a column chart that shows the frequency of the occurrence of a variable in the specified range.

Histogram is a graphical representation, similar to a bar chart in structure, that organizes a group of data points into user-specified ranges. The histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

A simple example of a histogram is the distribution of marks scored in a subject. You can easily create a histogram and see how many students scored less than 35, how many were between 35-50, how many between 50-60 and so on.

Here are some of the things you can do to customize this histogram chart:

1. **By Category:** This option is used when you have text categories. This could be useful when you have repetitions in categories and you want to know the sum or count of the categories. For example, if you have sales data for items such as Printer, Laptop, Mouse, and Scanner, and you want to know the total sales of each of these items, you can use the By Category option. It isn't helpful in our example as all our categories are different (Student 1, Student 2, Student3, and so on.)

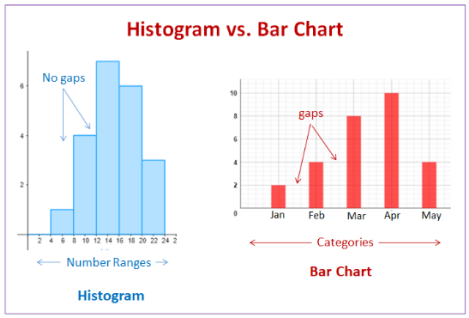
2. **Automatic:** This option automatically decides what bins to create in the Histogram. For example, in our chart, it decided that there should be four bins. You can change this by using the 'Bin Width/Number of Bins' options (covered below).

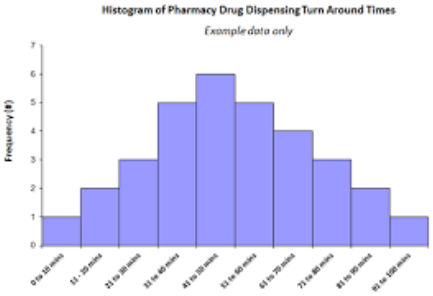
3. **Bin Width:** Here you can define how big the bin should be. If I enter 20 here, it will create bins such as 36-56, 56-76, 76-96, 96-116.

4. **Number of Bins:** Here you can specify how many bins you want. It will automatically create a chart with that many bins. For example, if I specify 7 here, it will create a chart as shown below. At a given point, you can either specify Bin Width or Number of Bins (not both).

5. **Overflow Bin:** Use this bin if you want all the values above a certain value clubbed together in the Histogram chart. For example, if I want to know the number of students that have scored more than 75, I can enter 75 as the Overflow Bin value. It will show me something as shown.

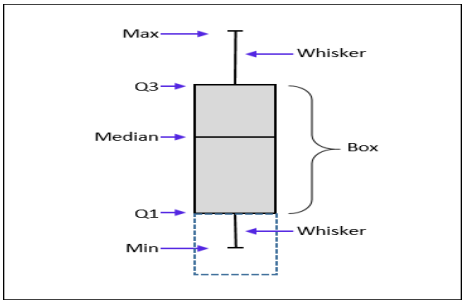
6. **Underflow Bin:** Similar to Overflow Bin, if I want to know the number of students that have scored less than 40, I can enter 4o as the value and show a chart as shown below.

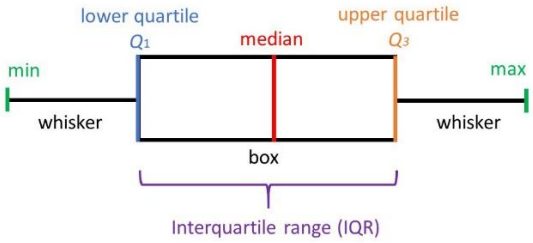




8) Box plot:

A **boxplot** (box plot, or whisker plot) is a compact, but efficient way to represent a dataset using descriptive stats. This "little diagram" combines informative, standard values such as the first and third **quartiles** (the bottom and top of the box, respectively), the **median** (the flat line inside the box) and sometimes the mean (a second flat line inside the box). The **whiskers** are often used to represent the minimum and maximum values, but some use other parameters such as: one standard deviation above and below the mean of the data OR the lowest and highest values contained in the range defined by the 1st quartile minus 1,5 times the interquartile range and the 3rd quartile plus 1,5 times the interquartile range (cf. "Tukey plot") OR the 5th and 95th percentiles, etc. Anyway, because the whiskers are defined by the user (and not by convention), it is important, when creating the boxplot, to mention what they represent in the legend of the chart.





9) Stem and leaf plot:

A stem-and-leaf display (also known as a stem plot) is a diagram designed to allow you to quickly assess the distribution of a given dataset. Basically, the plot splits two-digit numbers in half:

Stems – The first digit
Leaves – The second digit