TABULATION AND GRAPHICAL REPRESENTATION OF DATA & DESCRIPTIVE STATISTICS

By Dr. Bidisha Bhabani

Meaning of Data

- If you record the minimum and maximum temperature, or rainfall, or the time of sunrise and sunset, or attendance of children, or the body temperature of the patient, over a period of time, what you are recording is known as data.
- As an example, **Class-wise Attendance of Students**

Class	Number of Student Present
VI	42
VII	40
VIII	41
IX	35
X	36

- So, **data** refers to the set of observations, values, elements or objects under consideration.
- The complete set of all possible elements or objects is called a **population.** Each of the elements is called a **piece of data.**
- Data also refers to the known facts or things used as basis for inference or reckoning facts, material to be processed or stored.

Nature/Type of Data

- Cross-section, Time-series Data and Panel Data
- Qualitative and Quantitative Data
- Continuous and Discrete Data
- Primary and Secondary Data
- Structured and Unstructured Data

Cross-section and Time-series Data

- Data collected on a single point of time over different sections (may be classified on demographic, geographic or other considerations) are called cross-section data.
- Whereas data collected over a period of time are called time series data
- **Panel Data:** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).

Qualitative and Quantitative Data

- **Management-wise Number of Schools**

Management	No. of Schools
Government	4
Local Body	8
Private Aided	10
Private Unaided	2

- Such data is shown as **Categorical or Qualitative Data.**
- Here the category or the quality referred to is management. Thus categorical or qualitative data result from information which has been classified into categories.
- Such categories are listed alphabetically or in order of decreasing frequencies or in some other conventional way.
- Each piece of data clearly belongs to one classification or category.

- **Income wise number of Persons**

Income Range	No. of Persons
< 3 LPA	50
3 to 5 LPA	25
5 to 10 LPA	15
>10 LPA	5

- As the grouping is based on numbers, such data are called **Numerical or Quantitative Data.**
- Thus, numerical or quantitative data result from counting or measuring.
- We frequently come across numerical data in newspapers, advertisements etc. related to the temperature of the cities, cricket averages, incomes, expenditures and so on.

Continuous and Discrete Data

- **Heights of Students of a Class**

Height	No. of Students
4'8"- 4'10"	2
4'10" - 5'0"	2
5'0" - 5'2"	5
5'2" - 5'4"	8
5'4" - 5'6"	12
5'6" - 5'8"	10
5'8"- 5'10"	2

- Two students may vary by **almost** zero inch height. Even if we take two adjacent points, say 4' 8.00" and 4' 8.01" there may be several values between the two points. Such data are called **Continuous Data,** as the height is continuous.
- **Continuous Data** arise from the measurement of continuous attributes or variables, in which individual may differ by amounts just approaching zero. Weights and heights of children; temperature of a body; intelligence and achievement level of students, etc. are the examples of continuous data.

- **Number of books in a book shelf**

Book Shelf Number	No. of Books
1	10
2	20
3	15
4	12
5	25
6	18

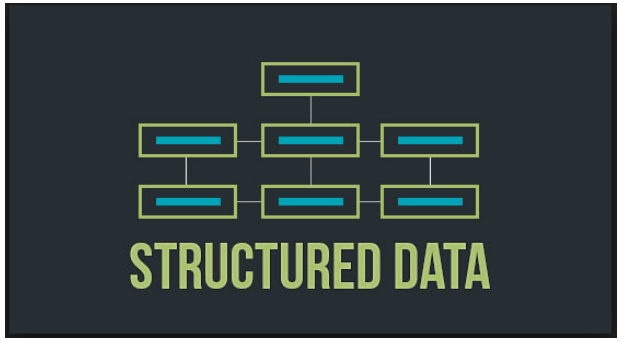
- Such data, where the elements being observed have gaps are called **Discrete Data.**
- **Discrete Data** are characterized by ,gaps in the scale, for which no real values may ever be found. Such data are usually expressed in whole numbers.
- The size of a family, enrolment of children, number of books etc. are the examples of discrete data.
- Generally data arising from measurement are continuous, while data arising from counting or arbitrary classification are discrete.

Primary and Secondary Data

- The data collected by or on behalf of the person or people who are going to make use of the data refers to **primary data.**
- For example, the attendance of children, the result of examinations conducted by you are primary data.
- If you contact the parents of the children and ask about their educational qualifications to relate them to the performance of the children, this also gives primary data.
- Actually, when an individual personally collects data or information pertaining to an event, a definite plan or design, it refers to primary data.
- Sometimes an investigator may use the data already collected by you, such as the school attendance of children, or performance of students in various subjects. etc., for his/her study, then the data are secondary data.
- The data used by a person or people other than the people by whom or for whom the data were collected refers to **secondary data.**
- For many reasons we may have to use secondary data, which should be used carefully, since the data could have been collected with a purpose different from that of the investigator and may lose some detail or may not be fully relevant. For using secondary data, it is always useful to know :
  - how the data have been collected and processed;
  - the accuracy of data;
  - how far the data have been summarized;
  - how comparable the data are with other tabulations; and
  - how to interpret the data, especially when figures collected for one purpose are **used** for another purpose.

Structured and Unstructured Data

- Structured data means that the data is described in a matrix form with labelled rows and columns.



- Any data that is not originally in the matrix form with rows and columns is an unstructured data.

Measurement Scales

- **Nominal Scale**
- **Ordinal Scale**
- **Interval Scale**
- **Ratio Scale**

- Let us take four different situations for a class of 30 students :
  - Assigning them roll nos. from 1 to 30 on random basis.
  - Asking the students to stand in a queue as per their heights and assigning them position numbers in queue from 1 to 30.
  - Administering a test of 50 marks to all students and awarding marks from 0 to 50, as per their performance.
  - Measuring the height and weight of students and making student-wise record.

Nominal Scale

Assigning them roll nos. from 1 to 30 on random basis.

- In the first situation, the numbers have been assigned purely on arbitrary basis.
- Any student could be assigned No. 1 while any one could be assigned No. 30.
- No two students can be compared on the basis of allotment of numbers, in any respect. The students have been labelled from 1 to 30 in order to give each an identity.
- This scale refers to **nominal scale.**
- Here the property of identity is applicable but the properties of order and additivity are not applicable

Ordinal Scale

Asking the students to stand in a queue as per their heights and assigning them position numbers in queue from 1 to 30.

- In the second situation, the students have been assigned their position numbers in queue from 1 to 30.
- Here the numbering is not on arbitrary basis. The numbers have been assigned according to the height of the students. So the students are comparable on the basis of their heights, as there is a sequence in this regard.
- Every subsequent child is taller than the previous one, and so on.
- This scale refers to **ordinal scale.**
- Here the object or event has got its identity, as well as order.
- As the difference in height of any two students is not known, so the property of addition of numbers is not applicable to the ordinal scale.

Interval Scale

Administering a test of 50 marks to all students and awarding marks from 0 to 50, as per their performance.

- In the third situation, the students have been awarded marks from 0 to 50 on the basis of their performance in the test administered on them.
- Consider the marks obtained by 3 students, which are 30,20 and 40 respectively. Here, it may be interpreted that the difference between the performance of the 1st and 2nd student is the same, as between the performance of the 1st and the 3rd student.
- This scale refers to **interval scale.** Here the properties of identity, order and additivity are applicable.

Ratio Scale

Measuring the height and weight of students and making student-wise record.

- In the fourth situation, the exact physical values pertaining to the heights and weights of all students have been obtained.
- Here the values are comparable in all respect.
- If two students have heights of 120 cm and 140 cm, then the difference in their heights is 20 cm and the heights are in the ratio 6:7.
- This scale refers to **ratio scale.**

Statistics

- Statistics can be described as the science of classifying and organizing data in order to draw inferences.
- Statistics refers to the methodology for the collection, presentation and analyses of data and for the uses of such data.
- Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of this analysis. It is concerned with the systematic collection of numerical data and its interpretation. This systematic collection of data distinguishes statistics from other kinds of information.
- Statistics is the science which helps us to extract useful information for numerical data. It does not restrict itself to the collection and presentation of data, but it also deals with the interpretation and drawing of inferences from the data.
- The term statistics is used both in its singular and plural sense.
- In the singular sense, it is a science which concerns itself with the collection, presentation and drawing of conclusions from numerical data.
- In the plural sense, it means numerical facts or observations collected with a definite object in view.
- Statistics are expressed quantitatively and not qualitatively.

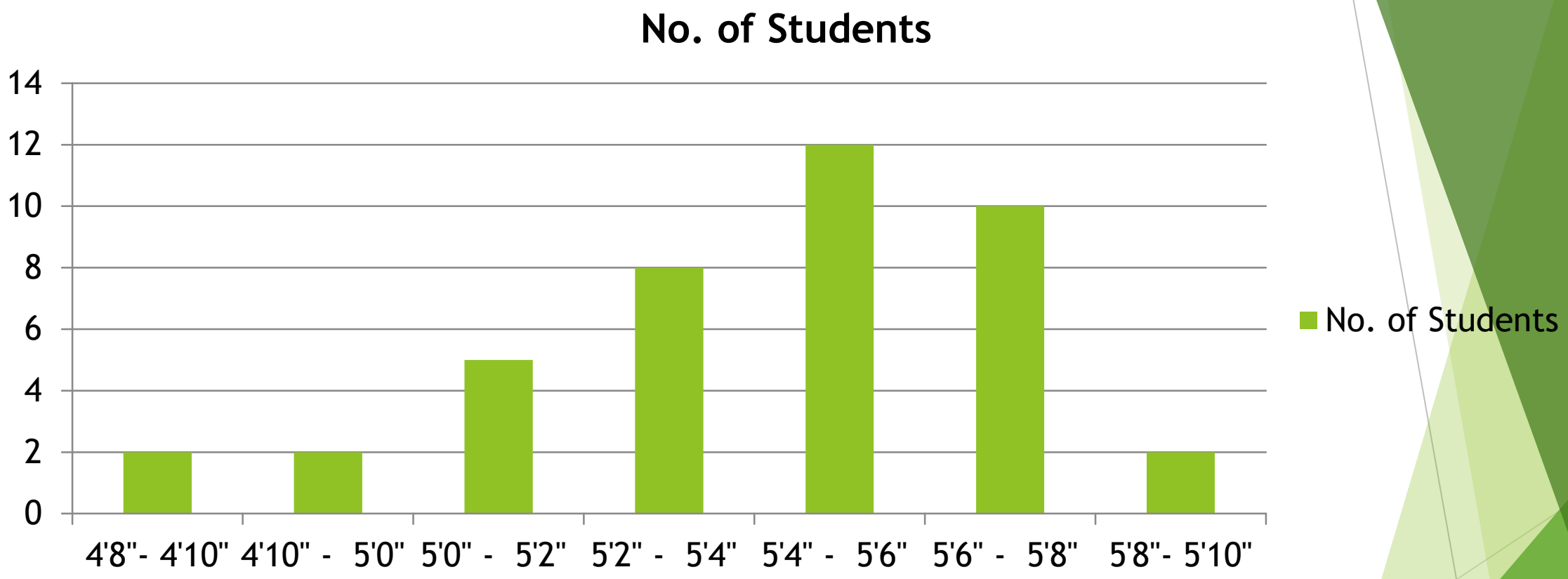
Presentation of Data

- In Sequence
  - Ascending Order
  - Descending Order
- Grouping and Tabulation of Data
  - **Frequency distribution.**
    - A grouped frequency distribution has a minimum of two columns - the first has the classes arranged in some meaningful order, and a second has the corresponding frequencies. The classes are also referred to as class intervals. The range of scores or values in each class interval is the same.
  - For the presentation of data in the form of a frequency distribution for grouped data, a number of steps are required. These steps are :
    1. Selection of non-overlapping classes.
    2. Enumeration of data values that fall in each class.
    3. Construction of the table.

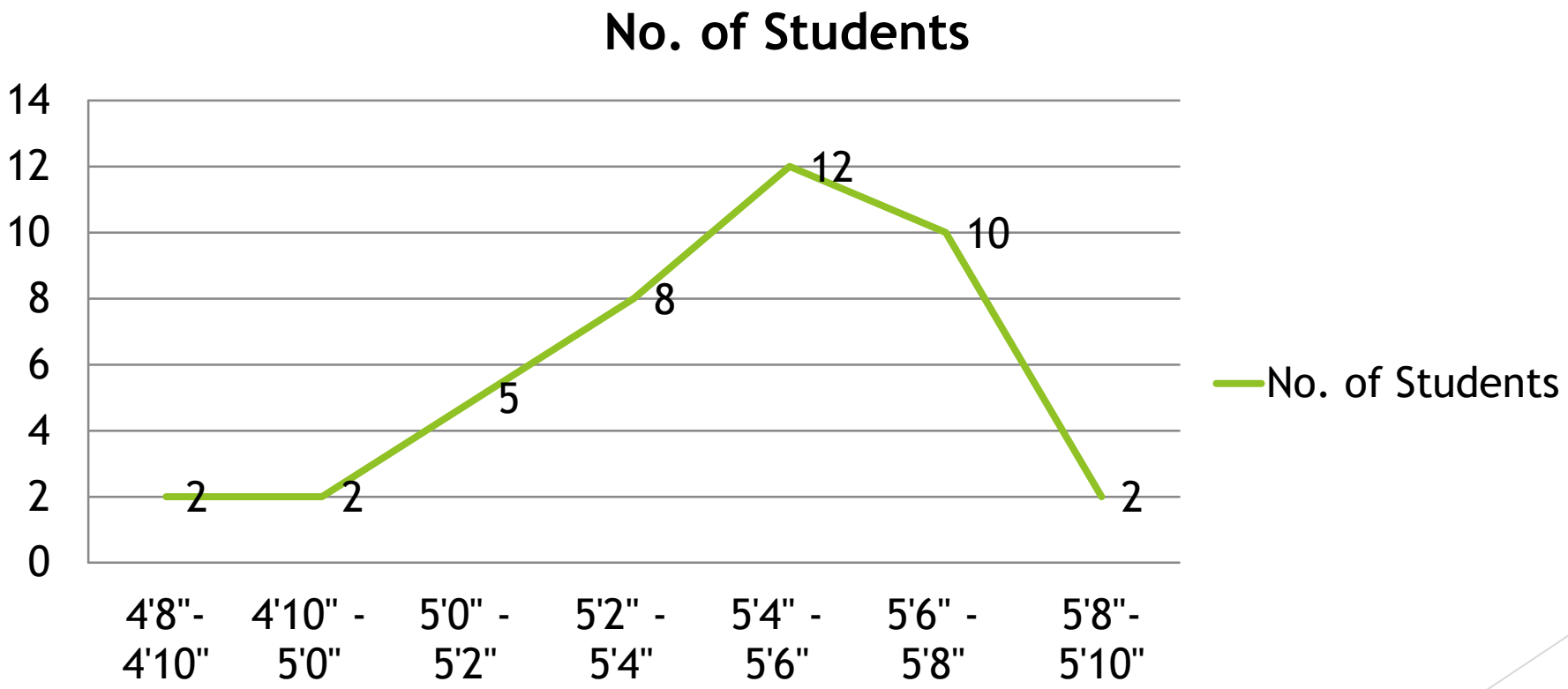
Graphical Representation of Data

- Histograms
- Line Diagrams
- Bar Diagram or Bar Graph
- Pie Chart
- Scatter Plot
- Frequency Polygons
- Cumulative Frequency Curve

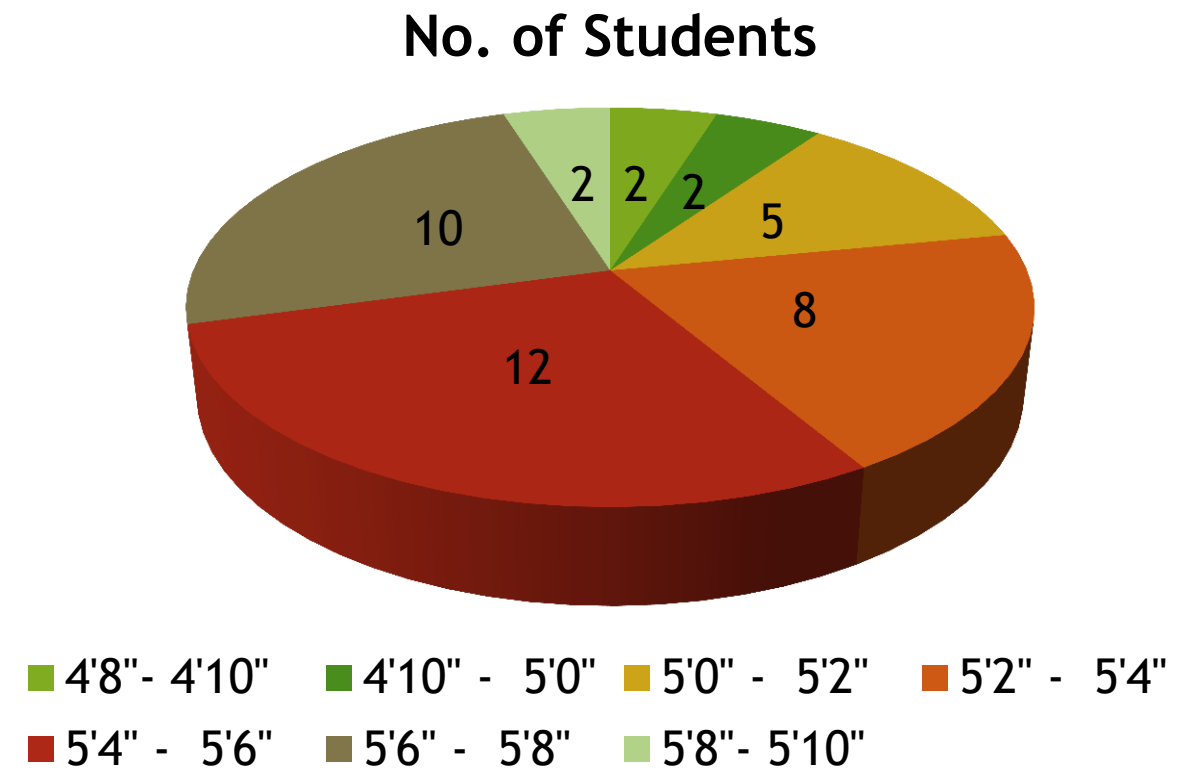
Histograms



Line Diagrams



Pie Diagrams

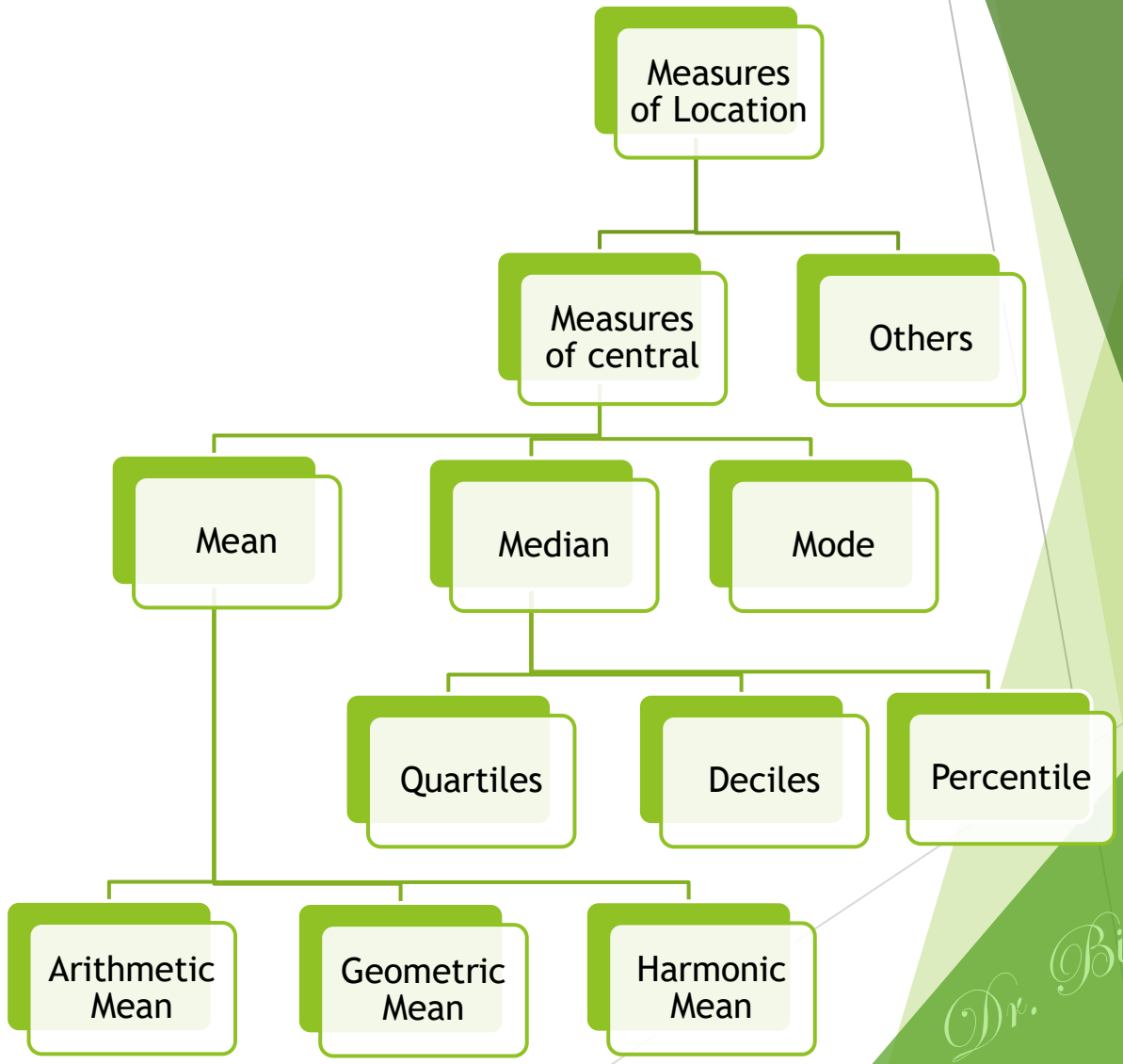


Descriptive Statistics

- The collection, organization and graphic presentation of numerical data help to describe and present these into a form suitable for deriving logical conclusions.
- Analysis of data is another way to simplify quantitative data by extracting relevant information from which summarized and comprehensive numerical measures can be calculated.
- Most important measures for this purpose are measures of location, dispersion and symmetry and skewness.

Measures of Location

- A single value can be derived for a set of data to describe the elements contained in it. Such a value is called a measure of location.
- Again, the central tendency is the property of data by virtue of which they tend to cluster around some central part of the distribution.
- Mean, median and mode are the measures of central tendency. There are other measures of location, namely, quartiles, deciles and percentiles.



Arithmetic mean

- Arithmetic mean of a set of realizations of a variable is defined as their sum divided by the number of observations.
- It is usually denoted by $\overline{x}$ (read as x bar) where x denotes the variable. Depending on whether the data are grouped or ungrouped arithmetic mean may be of two types.
- First, simple arithmetic mean for ungrouped data and second, weighted arithmetic mean for grouped (frequency type) data.
- If the realizations of the variable x are x_1, x_2…x_n then,
  - Simple Arithmetic Mean ($\overline{x}$) = (x_1 + x_2 +……. + x_n) / n
  - $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- If the variable x takes the values x_1, x_2…x_n with frequencies f_1, f_2…f_n then
  - Weighted arithmetic mean ($\overline{x}$) = (x_1 . f_1 + x_2 . f_2 +……. + x_n . f_n) / $\sum_{i=1}^{n} f_i$
  - $\overline{x} = \frac{\sum_{i=1}^{n} x_i . f_i}{\sum_{i=1}^{n} f_i}$
- Given two groups of observations, n_1 and n_2, and x_1 and x_2 being the number of observations and arithmetic mean of two groups respectively, we can calculate the composite mean using the following formula:
  - Composite Mean (x) = (n_1. x_1 + n_2. x_2) / n_1 + n_2

***Example**: Given the following data calculate the simple and weighted arithmetic average price per ton of iron purchased by an industry for six months.*

Month	Price per ton (in Rs.)	Iron purchased (in ton)
Jan.	42	25
Feb	51	35
Mar	50	31
Apr	40	47
May	60	48
June	54	50

Month	Price per ton (in Rs.) x	Iron purchased (in ton) f	x . f
Jan.	42	25	1050
Feb	51	35	1785
Mar	50	31	1550
Apr	40	47	1880
May	60	48	2880
June	54	50	2700
Total	297	236	11845

- $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = 297/6 = 49.5$
- $\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = 11845/236 = 50.19$

Geometric Mean & Harmonic Mean

- **Geometric Mean** of a set of observations is nth root of their product, where n is the number of observation.
- In case of non frequency type data, simple geometric mean

$$= \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \ldots \ xn}$$

- And in case of frequency type data weighted geometric mean

$$= \sqrt[n]{x_1 f_1 \ . \ x_2 f_2 \ . \ x_3 f_3 \ldots . \ xnf_n}$$

- **Harmonic Mean** is the reciprocal of the arithmetic mean and computed with the reciprocal of the observations. For data without frequency,

$$\text{Simple Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

In case of data with frequency,

$$\text{Harmonic Mean} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

Median

- Median of a set of observation is the middle most value when the observations are arranged in order of magnitude.
- The number of observations smaller than median is the same as the number of observations greater than it.
- Thus, median divides the observations into two equal parts and in a certain sense it is the true measure of central tendency, being the value of the most central observation.
- It is independent of the presence of extreme values and can be calculated from frequency distributions with open-ended classes.
- Note that in presence of open-ended process calculation of mean is not possible.

Calculation of Median

- For ungrouped data, the observations have to be arranged in order of magnitude to calculate median. If the number of observations is odd, the value all the middle most observation is the median . However, if the number is even, the arithmetic mean of the two middle most values is taken as median.
- For simple frequency distribution, to calculate median, we have to calculate the less than type cumulative frequency distribution. If the total frequency is **N**, the value of the variable corresponding to the cumulative frequency **(N+1)/2** gives the median.
- Median of a grouped frequency distribution is that value of the variable which corresponds to the cumulative frequency N/2. Median can be calculated using either the formula or graph. Both these methods are given below:
- To use the formula for median we have to calculate the cumulative frequency for each class. The class in which the cumulative frequency N/2 lies is called the median class. To compute median we apply the following formula:
  - Median = $l_1 + ((N/2 - F) / f_m) \times c$
  - where, l_1 : lower boundary of the median class
  - N : total frequency
  - F : cumulative frequency below l_1
  - f_m : frequency of the median class
  - c : difference between upper and lower class limits of the median intervals.

Calculation of Median

- An appropriate value of median can be calculated graphically from ogives or cumulative frequency polygon. We have to draw a horizontal line from the point N/2 on the vertical axis, which shows the cumulative frequency, until it meets the ogives (either less than type or greater than type).
- From this point of intersection, a perpendicular is dropped on the horizontal axis.
- The position of the foot of the perpendicular is read from the horizontal scale showing the values of the variable.
- The advantage of median is that it is easy to understand and calculate. It could be calculated even if all the observations are not known.
- Median could also be calculated from grouped frequency distributions with classes of unequal width. But there are many disadvantages also.
- For the calculation of median data must be arranged. Unlike mean, median cannot be treated algebraically.
- In median, it is not possible to give higher weights to smaller values and smaller weights to higher values.
- Calculation of median from grouped frequency distribution assumes that the observations in the median class are uniform which may not be true always.

- Imagine that a top running athlete in a typical 200-metre training session runs in the following times: 26.1 seconds, 25.6 seconds, 25.7 seconds, 25.2 seconds, 25.0 seconds, 27.8 seconds and 24.1 seconds. How would you calculate his median time?

Rank	Time
1	24.1
2	25.0
3	25.2
4	25.6
5	25.7
6	26.1
7	27.8

- There are n = 7 data points, which is an uneven number. The median will be the value of the data points of rank
- (n + 1) ÷ 2 = (7 + 1) ÷ 2 = 4.
- The median time is 25.6 seconds.

- Now suppose that the athlete runs his eighth 200-metre run with a time of 24.7 seconds. What is his median time now?

Rank	Time
1	24.1
2	24.7
3	25.0
4	25.2
5	25.6
6	25.7
7	26.1
8	27.8

- There are now n = 8 data points, an even number. The median is the mean between the data point of rank
- n ÷ 2 = 8 ÷ 2 = 4
- and the data point of rank
- (n ÷ 2) + 1 = (8 ÷ 2) +1 = 5
- Therefore, the median time is (25.2 + 25.6) ÷ 2 = 25.4 seconds.

Other Measures of Location

- Just as median divides the total number of observations into two equal parts, there are other measures which divide the observations into fixed number of parts, say, 4 or 10 or 100. These are collectively known as partition values or quartiles. Some of them are,
- 1) Quartiles 2) Deciles and 3) Percentiles.
- Median which falls into this group has already been discussed. Quartiles are such values which divide the total observations into four equal parts. To divide a set of observations into four equal parts three dividers are needed.
- These are first quartile, second quartile and third quartile. The number of observations smaller than Q1 is the same as the number of observations lying between Q1 and Q2, are between Q2 and Q3 or larger then Q3. One quarter of the observations is smaller than Q1, two quarter of the observations are smaller than Q2 and three quarter of the observations are smaller than Q3. This implies Q1, Q2, Q3 are values of the variable when the less than type cumulative frequencies is N/4, N/2 and 3N/4 respectively. Clearly, Q1 < Q2 < Q3; Q2 stands for median (as half of the observations are greater than the median and rest half are smaller than it. In other words, median divides the observations into two equal parts)
- Similarly, deciles divide the observations into ten equal parts and percentiles divide observations into 100 equal parts.

Mode

- Mode of a given set of observation is that value of the variable which occurs with the maximum frequency. Concept of mode is generally used in business as it is most likely to occur. Meteorological forecasts are based on mode.
- From a simple series, mode can be calculated by locating that value which occurs maximum number of times.
- From a simple frequency distribution mode can be determined by inspection only. It is that value of the variable which corresponds to the largest frequency.
- From the grouped frequency distribution mode can be determined. It is very difficult to find the mode accurately. However, if all classes are of equal width, mode is usually calculated using the following formula:

- Mode= $l_1+(l_2-l_1)(f_1-f_0)/(2f_1-f0-f2)$
  - Where l1= lower limit of modal class
  - L2=upper limit of modal class
  - f1= frequency of modal class
  - f0=frequency of pre-modal class
  - f2=frequency of post- modal class

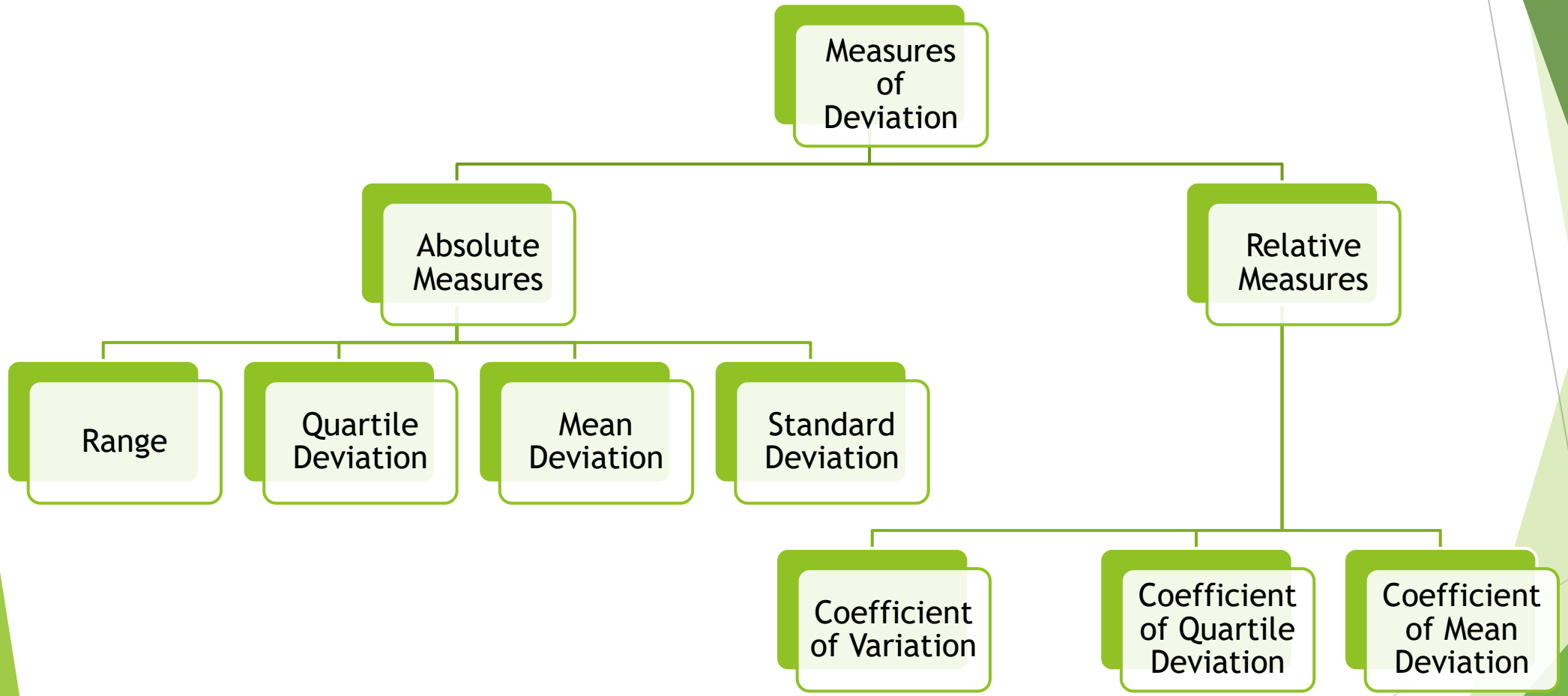
- Mode = l1 + {d1 / (d1 + d2)} × c
  - where, l1: lower boundary of the modal class (i.e., the class with the highest frequency)
  - d1: difference of the largest frequency and the frequency of the class just preceding the modal class
  - d2: difference of the largest frequency and the frequency of the class just following the modal class
  - c: common width of classes.

- The mode in statistics refers to a number in a set of numbers that appears the most often. For example, if a set of numbers contained the following digits, 1, 1, 3, 5, 6, 6, 7, 7, 7, 8, the mode would be 7, as it appears the most out of all the numbers in the set.

- For uni-modal distributions the following approximate relation holds:
  - Mean – Mode = 3 (Mean – Median)

Measures of Dispersion

- A measure of dispersion is defined as a numerical value explaining the extent to which individual observations vary among themselves. The measures are as following:
  1. Absolute Measures of Dispersion
  2. Relative Measures of Dispersion
- Absolute measures of dispersions measure numerically heterogeneity of data. These measures are not free from unit of measurement. Therefore, these measures cannot be used to measure the degree of heterogeneity between two data sets, which does not have the same unit of measurement.
- But relative measures are free from unit of measurement and therefore, they are useful in comparing the degree of variability between two sets of data with different unit of measurement.



Absolute Measures of Dispersion

- **Range:** The range of a set of observation is the difference between the maximum and minimum values.
- **Quartile Deviation:** Quartile deviation is defined as the half of the difference between the first and third quartiles. **Quartile Deviation :** = (Q3 - Q1) / 2.
- **Mean Deviation:** Mean deviation of a set of observations is the arithmetic mean of the absolute deviations, from mean or any other specified value. Here we take absolute deviations so that the positive and negative deviations do not cancel out each other.
  - Mean deviation about A = $\frac{1}{n}\sum |xi - A|$, where n is the number of observations
  - Mean deviation about mean = $\frac{1}{n}\sum |xi - \overline{x}|$
- **Standard Deviation:** Standard deviation of a set of observations is the square root of the arithmetic mean of squares of deviations from arithmetic mean. Here the deviation of each observation is taken to be the measure of the degree of heterogeneity of data from the central position and these deviations are squared to make all of them a positive number. In such a procedure the positive and negative values do not cancel out each other. After taking the mean of the square of the deviations we take square root of them to get the measure of standard deviation.
- **Standard deviation** is generally denoted by σ and is always is a non negative number.
- The square of standard deviation (S.D.) is called **variance** of a variable.
- S.D. and variance of a variable, say x, is denoted by σ_x and Var(x) or σ^2 respectively.

Absolute Measures of Dispersion

- For ungrouped frequency distribution, the S.D. is given by the following formula:
  - **Standard Deviation** = $\sqrt{\left(\frac{\sum_{i=1}^{n}(xi-\overline{x})^2}{n}\right)}$
- While for grouped frequency distribution it is given by
  - **Standard Deviation** = $\sqrt{\left(\frac{\sum_{i=1}^{n} f_i(xi-\overline{x})^2}{n}\right)}$
  - where f_i: frequency of the ith class.
  - x_i: mid value of the ith class.
- Some unique properties of S.D. make it superior to other measures of dispersion.
  - First, it is based on all observations. If the value of one observation changes, the S.D. changes but range and quartile deviation may remain unaffected due to such a change.
  - Secondly, it is least affected by the sampling fluctuation.
  - Thirdly, it is easy to be treated algebraically.
  - Fourthly, given the S.D. of two different groups S.D. along with the number of observations in each group and their mean (arithmetic mean) , the variance for the composite group can be easily calculated.

- **Variance** $= \frac{1}{n}\sum_{i=1}^{n}(xi - \overline{x})^2$
- The variance for the composite group can be easily calculated using the following formula:
  - $\sigma 2 = \{ (n_1\sigma_1^2 + n_2\sigma_2^2) + n_1 (\overline{x_1} - \overline{x})^2 + n_2 (\overline{x_2} - \overline{x})^2\} / (n_1 + n_2)$
  - where σ2 : composite variance
  - n_1 : number of observation in the first group
  - n_2 : number of observation in the second group
  - σ_1 : S.D. of the first group
  - σ_2 : S.D. of the second group
  - $\overline{x_1}$: mean of the first group
  - $\overline{x_2}$: mean of the second group
  - $\overline{x}$: composite mean

Relative Measures of Dispersion

- These measures are free of unit of measurement. Generally, the absolute measures are divided by measures of location to arrive at the relative measure of dispersion.
- There are three such measures, viz.,
  - Coefficient of Variation = S.D./ Mean × 100
  - Coefficient of Quartile Deviation = Quartile Deviation / Median × 100
  - Coefficient of Mean Deviation = Mean Deviation / Mean or Median × 100

Exercise

1. Calculate median of {12, 23, 7, 23, 5, 23, 3, 23, 13, 21, 14, 39, 40, 56, 29}.
2. Calculate mean of {5, 23, 3, 13, 21, 14, 39, 40, 56}.
3. Calculate mode of {12, 23, 7, 23, 5, 23, 3, 23, 13, 21, 14, 39, 40, 56, 29}.
4. Calculate variance & standard deviation of {50, 65, 72, 90, 87}.
5. Calculate the mean of {50, 65, 72, 90, 87}.
6. Evaluate mode of {12, 3, 7, 23, 5, 23, 3, 23, 12, 21, 14, 39, 40}.
7. Calculate the variance of {50, 65, 72, 90}.

1. Calculate mean median mode

Value	Frequency
1	3
2	5
3	2
4	4
5	3
6	3

2. Several seaside hotels were rated between "no stars" and "five stars" by the tourist board. The table below shows how many hotels got each number of stars. Calculate mean median mode.

Stars	Frequency
0	2
1	6
2	8
3	3
4	0
5	1

1. Find the Median

Monthly Sales	100 - 120	120 - 140	140 - 160	160 - 180	180 - 200	200 - 220
Frequency	15	35	50	60	30	10

2. Find Mode

Monthly Sales	100 - 200	200 - 300	300 - 400	400 - 500	500 - 600	600 - 700
Frequency	3	7	8	2	4	6

3. Find Median and Mode

Monthly Sales	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	13	7	10	8	4	8

4. Find S.D.

Monthly Sales	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
Frequency	10	30	40	20	20	20

Chebyshev's Theorem

- Chebyshev's theorem (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is that is $1-\frac{1}{k^2}$

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1 - \frac{1}{k^2}$$

- **Example:**

Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000?

- **Solution:**

$P(8000 \le X \le 16000)$

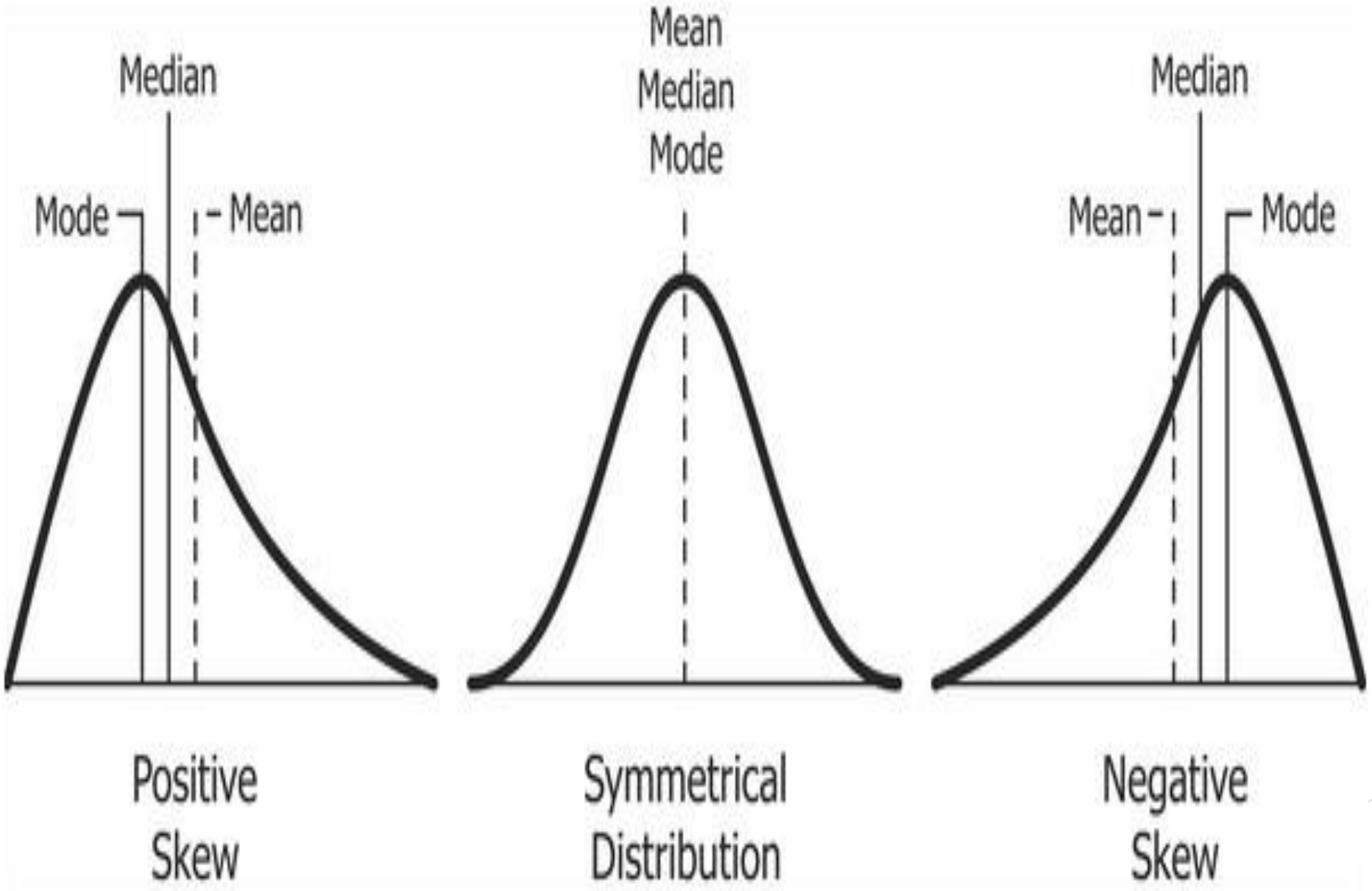
$= P(\mu - 2\sigma \le X \le \mu + 2\sigma) \ge 1 - \frac{1}{2^2}$

$= 0.75$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

Measures of Shape : Skewness

- Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.
- Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side. Skewness helps in understanding the shape and outliers in a dataset.
- Depending on the model, skewness in the values of a specific independent variable (feature) may violate model assumptions or diminish the interpretation of feature importance.
- In finance and investment analysis, skewness is used to measure the degree of asymmetry in returns on investment. Skewed returns can have an impact on portfolio management and risk management strategies, and understanding the skewness of a particular investment can help investors to make better-informed decisions.



Skewness

- **Pearson's first coefficient of skewness** : To calculate skewness values, subtract the mode from the mean, and then divide the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

- As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of -1 to +1. That accurately shows the range of the correlation values.
- Pearson's first coefficient of skewness is helping if the data present high mode. However, if the data exhibits low mode or multiple modes, it is preferable not to use Pearson's first coefficient, and instead, Pearson's second coefficient may be superior, as it does not depend on the mode.

Skewness

- **Pearson's second coefficient of skewness :** subtract the median from the mean, multiply the difference by 3, and divide the product by the standard deviation.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

- Rule of thumb :
- **Mean = Mode = Median, then the coefficient of skewness is zero for symmetrical distribution.**
- **Mean > Mode, then the coefficient of skewness will be positive.**
- **Mean < Mode, then the coefficient of skewness will be negative.**
- **For skewness values between -0.5 and 0.5, the data exhibit approximate symmetry.**
- **Skewness values within the range of -1 and -0.5 (negative skewed) or 0.5 and 1(positive skewed) indicate slightly skewed data distributions.**
- **Data with skewness values less than -1 (negative skewed) or greater than 1 (positive skewed) are considered highly skewed.**

Skewness

- Skewness is a measure of symmetry or lack of symmetry. A dataset is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between $\overline{x}$ and $\overline{x}$ - kσ is same as $\overline{x}$ and $\overline{x}$ + kσ , where k is some positive constant.

- Pearson's moment coefficient of skewness for a dataset with n observations is given by $$g1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^3/(n-1)}{\sigma^3}$$

- The value of g1 will be close to 0 when the data is symmetrical. A positive value of g1 indicates a positive skewness and a negative value indicates negative skewness.

Skewness

- The following formula is used usually for a sample with *n* observations (Joanes and Gill, 1998):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

- The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of *n* increases.

Skewness : Example 1

- **A boy collects some rupees in a week as follows (25,28,26,30,40,50,40) and finds the skewness of the given Data in question with the help of the skewness formula.**

Mean of Data = (25+28+26+30+40+50+40) / 7 = 239 / 7 = 34.14

Number of terms (n) =7 (odd)

Arrange Data in ascending order = 25,26 ,28,30,40,40,50

The median of data is = 30

Mode of Data = Highest Frequency term = 40 (frequency 2)

S.D = 9.3529

Skewness = 3(Mean - Median)/S.D.

By Applying Skewness Formula,

Skewness = 3(34.14 - 30)/9.3529 = 1.32

Skewness = 1.32

So skewness for these data is positive

Skewness : Example 2

- **Attendance of all classes of a school are as follows find their skewness? 1st (35), 2nd(32), 3rd(38), 4th(39), 5th(43)**

- *Mean of Data = (35 + 32 + 38 + 39 + 42)/5 = 186/5 = 37.2*
- *Number of terms (n) = 5 (odd)*
- *Arrange Data in ascending order = 32,35,38,39,42*
- *Median of Data = 38*
- *S.D. = 3.8341*
- ***Skewness = 3(mean - median)/s.d***
- *Skewness = 3(37.2 – 38)/3.8341 = -0.629*
- *So, the skewness of these data is negative.*

Measures of Shape : Kurtosis

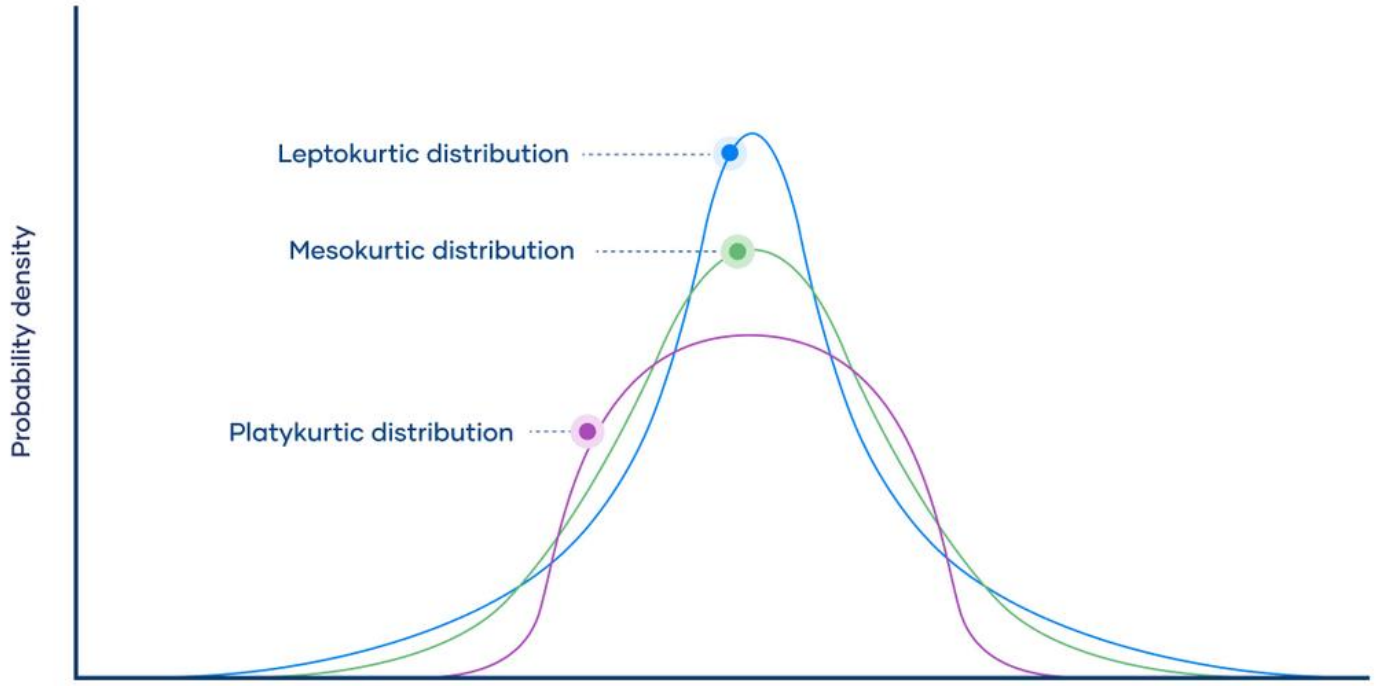
- Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution.
- Positive kurtosis indicates heavier tails and a more peaked distribution, while negative kurtosis suggests lighter tails and a flatter distribution. Kurtosis helps in analyzing the characteristics and outliers of a dataset.
- The measure of Kurtosis refers to the tailedness of a distribution. Tailedness refers to how often the outliers occur.
- Peakedness in a data distribution is **the degree to which data values are concentrated around the mean**. Datasets with high kurtosis tend to have a distinct peak near the mean, decline rapidly, and have heavy tails. Datasets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.
- In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

Kurtosis

- Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

- Kurtosis $= \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4/(n-1)}{\sigma^4}$

- Kurtosis value of less than 3 is called **platykurtic** distribution and greater than 3 is called **leptokurtic** distribution. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**)



Types of Kurtosis

- **Leptokurtic (Kurtosis > 3)**
- Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.
- **Platykurtic (Kurtosis < 3)**
- Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.
- **Mesokurtic (Kurtosis = 3)**
- Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Excess Kurtosis

- In statistics and probability theory, researchers use excess kurtosis to compare the kurtosis coefficient with that of a normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculated by subtracting kurtosis by 3.
- The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by:

$$\text{Excess Kurtosis} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4/(n-1)}{\sigma^4} - 3$$

Example

- **Calculate Sample Skewness, Sample Kurtosis from the following data 3,13,11,11,5,4,2**

- Kurtosis = 1.1231

- **Calculate Sample Skewness, Sample Kurtosis from the following data 85,96,76,108,85,80,100,85,70,95**

- Kurtosis = 1.9312

Conclusion

- Skewness measures the symmetry or asymmetry of data distribution, while kurtosis determines whether data exhibits a heavy-tailed or light-tailed distribution. Data may exhibit positive skewness (pushed towards the right side) or negative skewness (pushed towards the left side).

- Skewed data may cause the tail region to act as an outlier for the statistical model, and such outliers can adversely impact the performance of the model, particularly in regression-based models. Some statistical models are robust to outliers like Tree-based models, but it will limit the possibility of trying other models. So there is a necessity to transform the skewed data to be close enough to a Normal distribution.

Conclusion

- Skewness is a statistical measure of the asymmetry of a probability distribution. It characterizes the extent to which the distribution of a set of values deviates from a normal distribution.
- Skewness between -0.5 and 0.5 is symmetrical.
- Kurtosis determines whether the data exhibits a heavy-tailed or light-tailed distribution.
- Data sets with high kurtosis have heavy tails and more outliers, while data sets with low kurtosis tend to have light tails and fewer outliers.
- Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near zero (Mesokurtic distribution).