Advanced Data Analysis using industry accepted and widely popular statistical package

- **Dr. Md. Abdus Salam Akanda**
- Professor
- Department of Statistics
- Faculty of Science
- University of Dhaka
- Dhaka, Bangladesh
- E-mail: akanda@du.ac.bd

email: akanda@du.ac.bd

email: akanda@du.ac.bd

Week#4: Measure of Dispersion

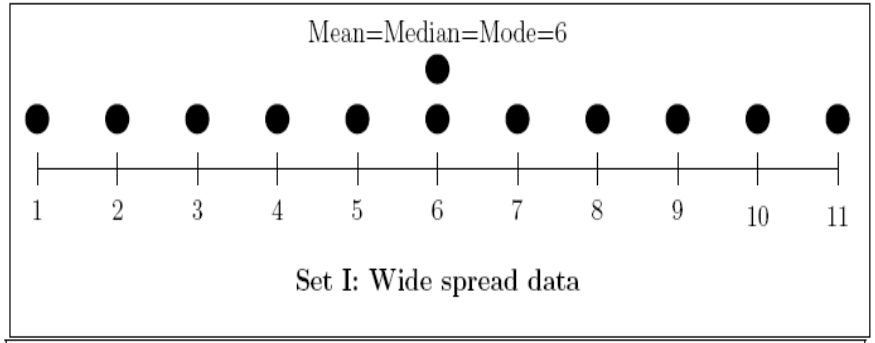
- **Course Outline**
- Concept of dispersion
- Definition of measure of dispersion
- Absolute and relative measures of dispersion
- Range
- Variance
- Standard deviation
- Skewness
- Kurtosis

- **TEXT BOOK: M. A. Salam Akanda (2018). *RESEARCH METHODOLOGY- A Complete Direction for Learners*, 2nd Edition, Akanda & Sons Publications, Dhaka, Bangladesh**

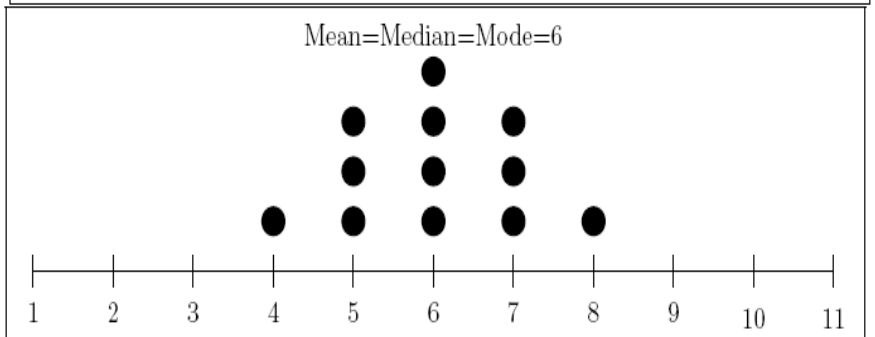
email: akanda@du.ac.bd

Concept of Dispersion

- In Set I, the observations are much more scattered (wide dispersion) from the center.
- In Set II, almost all the observations are concentrated (close dispersion) around the center.
- Certainly, the two sets differ even though they have the same mean, median and mode. Thus, there arises a need to differentiate between the sets. We need some other measures which concern with the measure of scatteredness (or spread).



Set I: Wide spread data



Set II: Data are clustered together

email: akanda@du.ac.bd

Definition of Dispersion

- Dispersion is the extent to which a distribution is stretched or squeezed. It is also called variability, scatter, or spread of a distribution. Common examples of measures of statistical dispersion are the range, variance, standard deviation, and the inter-quartile range.
- According to Bowley "**Dispersion is the measure of the variation of the item**".
- According to Conar "**Dispersion is a measure of the extent to which the individual items vary**".

email: akanda@du.ac.bd

Purpose of Measure of Dispersion

- Measures of variations or dispersion are needed for the following purposes:
  a) To compare two or more series with regard to their variability.
  b) To determine the reliability of an average.
  c) To serve as a basis for control of the variability.
  d) To facilitate the use of other statistical measures like correlation, regression, structural equation modeling, etc.

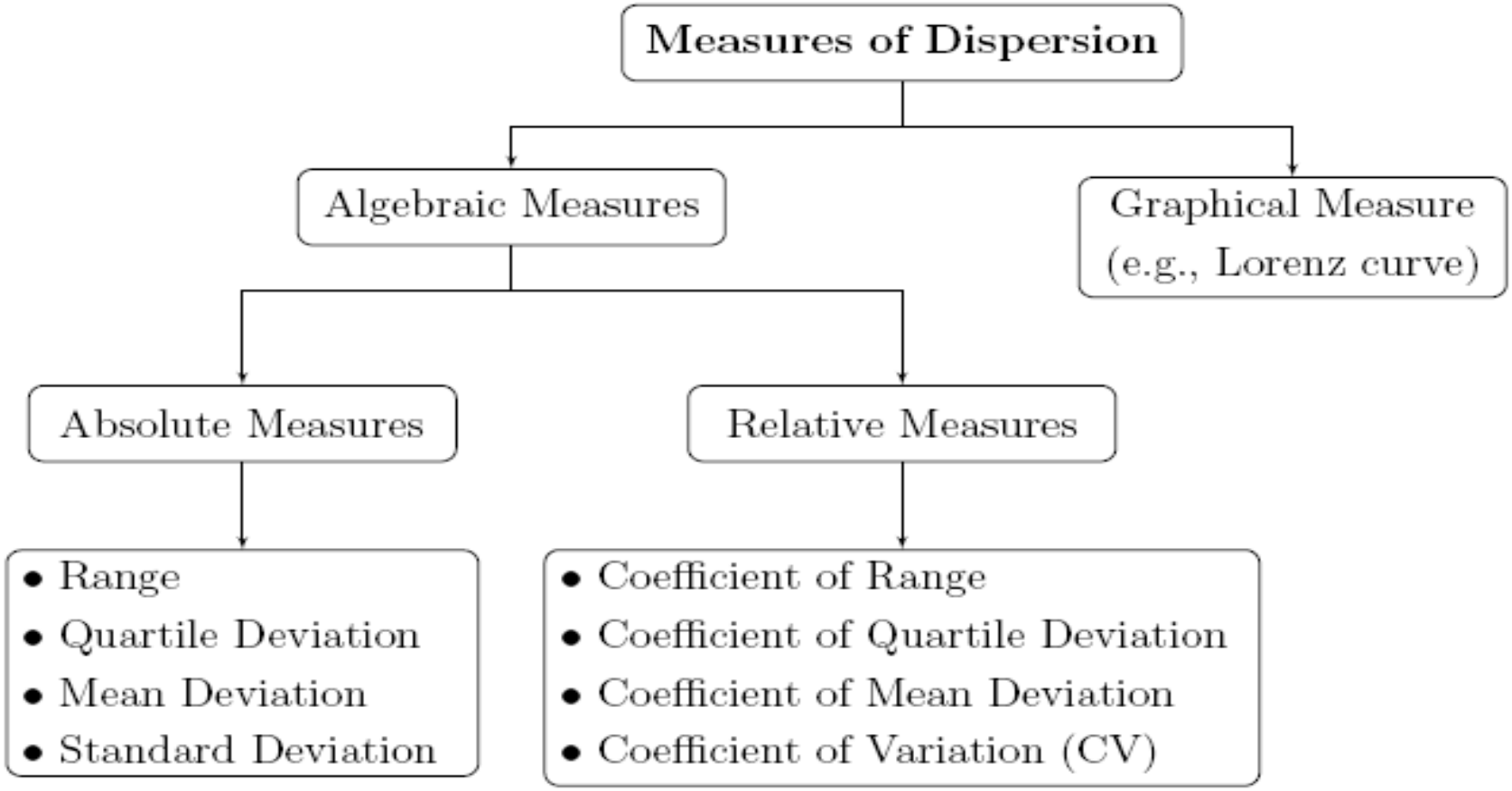
email: akanda@du.ac.bd

Properties of a Good Measure of Dispersion

- A good measure of dispersion should have the following properties:
  a) It should be easy to understand.
  b) It should be easy to compute.
  c) It should be rigidly defined.
  d) It should be based on all the items of the distribution.
  e) It should be amenable to further mathematical treatment.
  f) It should not be unduly affected by extreme values.
  g) It should be least affected from sampling fluctuations.

email: akanda@du.ac.bd

Methods of Measuring Dispersion



email: akanda@du.ac.bd

Absolute Measure of Dispersion

- The measures of dispersion which are expressed in terms of original units of a data are termed as absolute measures. For example, if computers measured by numbers, it shows dispersion by number.
- Absolute measure of dispersion gives an idea about the amount of variation in a set of observations.
- Absolute measures cannot be used to compare the variation of two or more data set.

email: akanda@du.ac.bd

Relative Measure of Dispersion

- The dispersion of the two distributions cannot be compared if they are expressed in two different units of measurements. In that case we need a relative measure of dispersion which is used to measure and **compare variations of data in different series expressed in different units of measurement**.
- A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is sometimes called a coefficient of dispersion because coefficient means a pure number that is independent of unit of measurement.

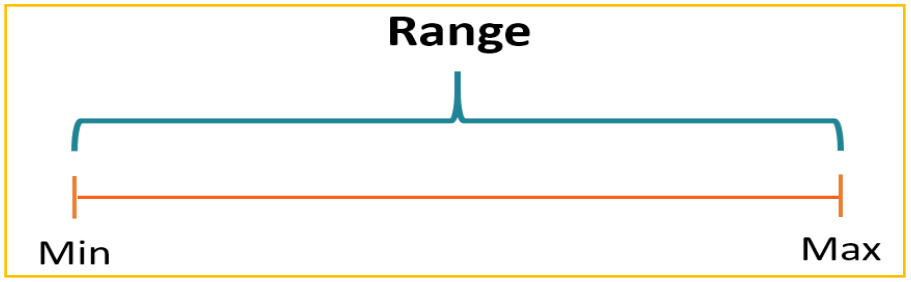
email: akanda@du.ac.bd

Range

- Range is measured just as the difference between the highest and the lowest values of a variable. The extent of dispersion increases as the divergence between the highest and the lowest values of the variable increases. Symbolically,

$$Range = H - L$$

  - where, H =the highest value and
  - L = the lowest value.



AGES OF STUDENTS

13,13,14,14,14,15,15,15,15,16,16,16

Range = highest - lowest

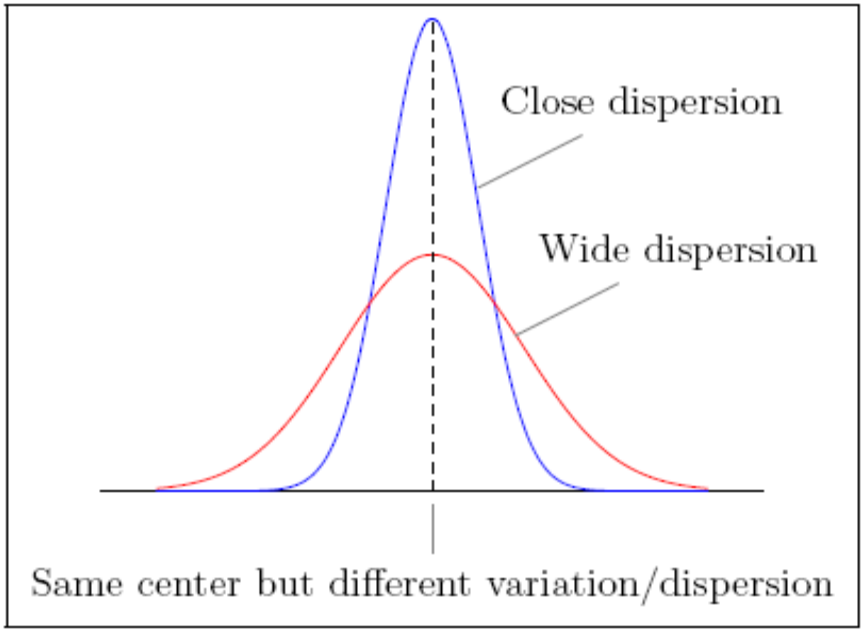
= 16 - 13

Range = 3

email: akanda@du.ac.bd

Interpretation of Range

- If the averages of the two distributions are almost same, the distribution with smaller range is said to have less dispersion and the distribution with larger range is said to have more dispersion.



email: akanda@du.ac.bd

Advantages of Range

- The advantages of range are as follows:
  a) It is easy to understand.
  b) It is easy to calculate.
  c) It does not require any special knowledge.
  d) It takes minimum time to calculate the value of range.

email: akanda@du.ac.bd

Disadvantages of Range

- The disadvantages of range are given below:
  a) It does not take into consideration all items of the distribution.
  b) Only two extreme values are taken into consideration.
  c) It is affected by extreme values.
  d) It is not possible to find out the range in open-end frequency distribution.
  e) It does not present very accurate picture of the series.
  f) It does not indicate the direction of the variability.

email: akanda@du.ac.bd

Finding Range Using SPSS

- Open the data file and follow the instructions given below:

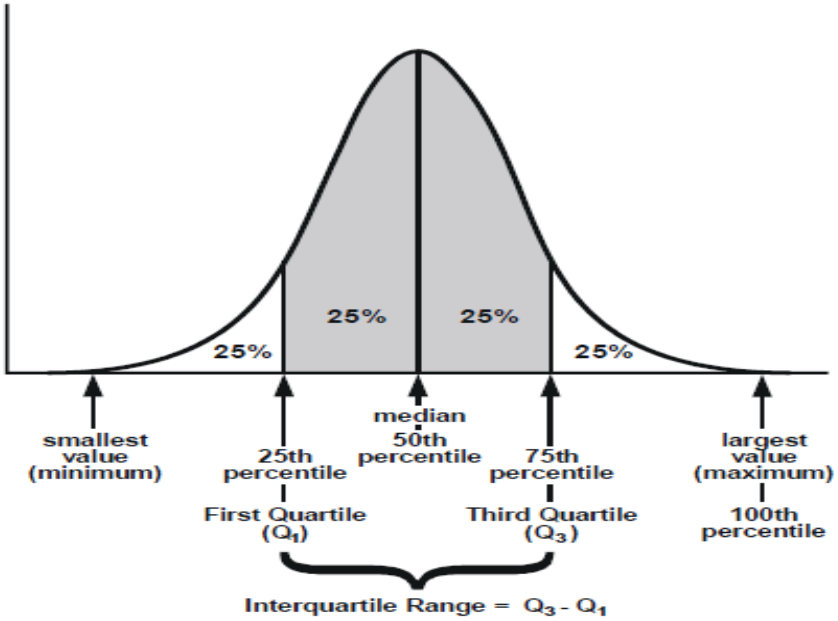
1. From the menu at the top of the screen click on **Analyze**, then click on **Descriptive Statistics**, then **Frequencies**…..
2. Choose and highlight the variables you are interested. Move these into the **Variable(s) box**. Then click **Statistics**…….
3. In the **Frequencies: Statistics dialog** box select **Range**, **Maximum** and **Minimum**.
4. Click **Continue** and then click **OK**.

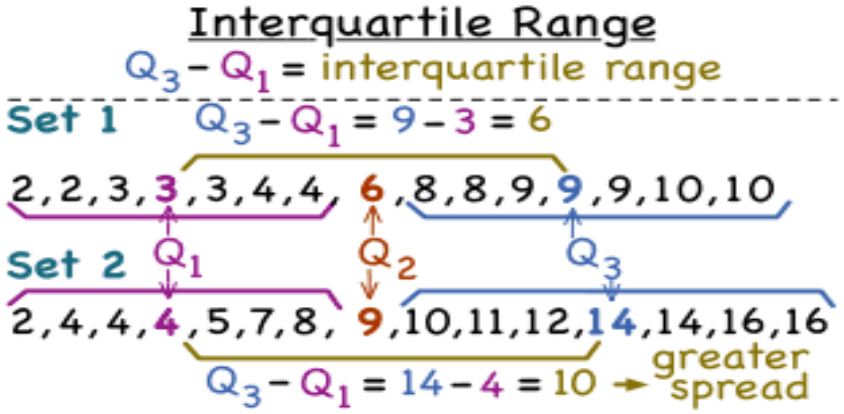
email: akanda@du.ac.bd

Quartile Deviation

- The average value of the difference between the third and the first quartiles, i.e., half of the Inter-Quartile Range is termed as the Quartile Deviation (QD). Symbolically, we can write:

$$Quartile\ Deviation = \frac{Q_3 - Q_1}{2}$$





email: akanda@du.ac.bd

Advantages of Quartile Deviation

- The advantages of quartile deviation are given below:
  a) Quartile deviation is easy to calculate numerically.
  b) Calculation for quartile deviation involves only the first and the third quartiles.
  c) It can be used safely as a suitable measure of dispersion at all situations.
  d) It remains unaffected from the extreme values of the variable.
  e) It can be calculated readily from frequency distributions with the open-end classes.
  f) QD at least is a better measure of dispersion compared to Range.

email: akanda@du.ac.bd

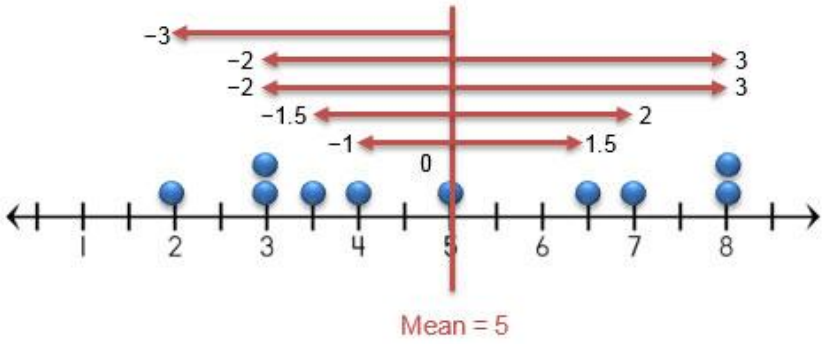
Disadvantages of Quartile Deviation

- The disadvantages of quartile deviation are given below:
  a) Quartile deviation as a measure of dispersion is not much popularly prescribed by the statisticians.
  b) It is used only in rare cases.
  c) It is not a reliable measure of dispersion as it ignores almost 50% of the data.

email: akanda@du.ac.bd

Mean Deviation

- The mean deviation (MD) is actually the mean of absolute deviations of the mean values of the variable from some average (normally the mean, the median or the mode).
- Graphically, the deviations can be represented on a number line from a dot plot (see figure in the right side).
- Numerically, the absolute deviations can be represented using the absolute value. The mean is calculated by adding the absolute deviations together, and then dividing by the number of values in the data set.



$$|-3|+|-2|+|-2|+|-1.5|+|-1|$$
$$+|0|+|1.5|+|2|+|3|+|3|=19$$
$$\therefore \text{MD}=19\div 10=1.9$$

email: akanda@du.ac.bd

Advantages of the Mean Deviation

- The advantages of mean deviation are given below:

a) On many occasions it gives fairly good results to represent the degree of variability or the extent of dispersion of the given values of a variable as it takes separately all the observations given into account.

b) It can also be calculated about the median value of those observations as their central value and then it gives us the minimum value for the mean deviation.

c) In usual situations, it is calculated taking deviations from the easily computable arithmetic mean of the given observations on the variable.

d) It is easy to calculate numerically and simple to understand.

email: akanda@du.ac.bd

Disadvantages of the Mean Deviation

- The disadvantages of mean deviation are given below:
  a) The main complaint against this measure is that it ignores the algebraic signs of the deviations.
  b) It is not generally computed taking deviations from the mode value and thereby disregards it as another important average value of the variable.
  c) It is not suitable for further mathematical treatments.
  d) It is rarely used in practical purposes.

email: akanda@du.ac.bd

Standard Deviation and Variance

- **Standard deviation** of a set of observations on a variable is defined as the square root of the arithmetic mean of the squares of deviations from their arithmetic mean. Again, it is often denoted as the positive square root of the variance of a group of observations on a variable.

- **Variance** is the arithmetic mean of the squares of the deviations of all values in a set of numbers from their arithmetic mean. In other words, **variance** is the square of the standard deviation.

- The **variance** is a positive quantity that measures the spread of the distribution of the random variable about its mean value. Larger values of the variance indicate that the distribution is more spread out.

email: akanda@du.ac.bd

Calculation of Standard Deviation and Variance

For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

Calculating Formula

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

The marks scored in a test by seven students are 3,4,6,2,8,8,5. Calculate the variance and standard deviation.

x	x^2
3	9
4	16
6	36
2	4
8	64
8	64
5	25

$\Sigma x = 36 \quad \Sigma x^2 = 218$

$$\text{Variance} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

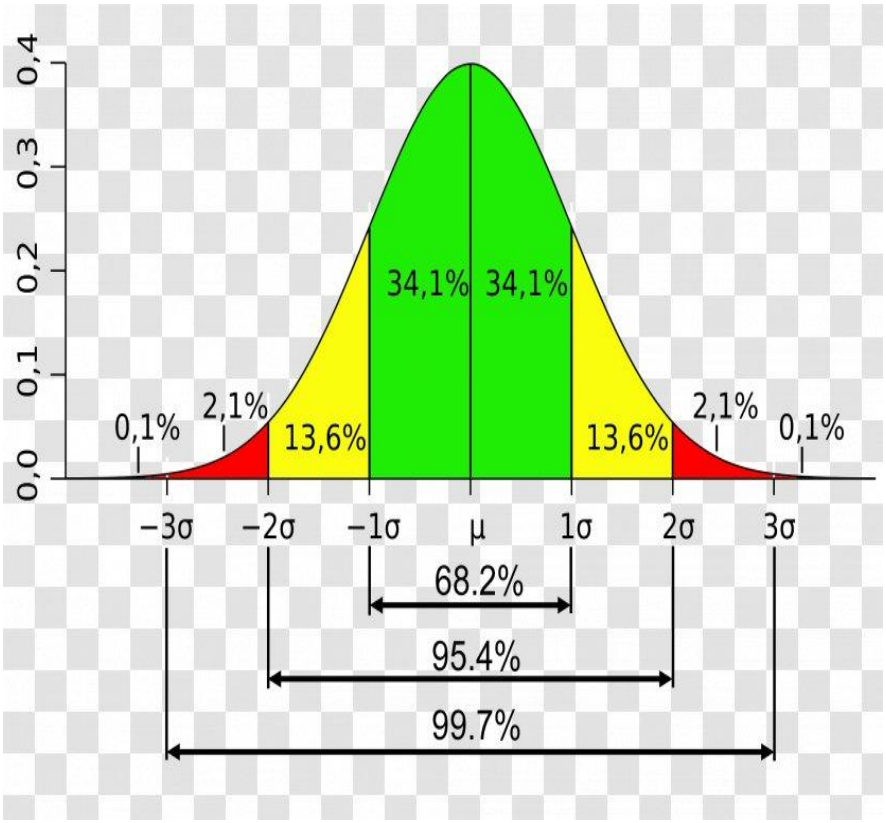
$$\sigma^2 = \frac{218}{7} - \left(\frac{36}{7}\right)^2 = 4.69$$

$$\sigma = \sqrt{4.69} = 2.17$$

email: akanda@du.ac.bd

Interpreting Standard Deviation

- Standard deviation is a measure of the typical amount an entry deviates from the mean. The more the entries are spread out, the greater the standard deviation.
- A small standard deviation means that the values in a statistical data set are close to the mean of the data set, on average, and a large standard deviation means that the values in the data set are farther away from the mean, on average.



email: akanda@du.ac.bd

Advantages of Standard Deviation

- The advantages of standard deviation are listed below:
  a) It is well defined.
  b) Calculation of standard deviation involves all the values of the given variable.
  c) It uses arithmetic mean of the given data as an important component which is simply computable.
  d) It is least affected by sampling fluctuations.
  e) It is easily usable and capable of further mathematical treatments.
  f) It is taken as the most reliable and dependable device for measuring dispersion of the given values of a variable.
  g) Statisticians very often prescribe standard deviation as the true measure of dispersion of a series of information.
  h) It can tactfully avoid the complication of considering negative algebraic sign while calculating deviations.

email: akanda@du.ac.bd

Disadvantages of Standard Deviation

- The disadvantages of standard deviation are given below:
  a) It involves complicated and laborious numerical calculations specially when the information are large enough.
  b) The concept of standard deviation is neither easy to take up, nor much simple to calculate.
  c) It is considerably affected by the extreme values of the given variable.
  d) To compute standard deviation correctly, the method claims much moments, money and manpower.

email: akanda@du.ac.bd

Best Measure of Dispersion

- ***Standard deviation is the best measure of dispersion*** as it satisfies the most essentials of the good measure of dispersion. The following points make standard deviation the best measure of dispersion:
  a) Most of the statistical theory is based on the standard deviation.
  b) It is based on the values of all the observations.
  c) It has a precise value and is a well-defined and definite measure of dispersion.
  d) It is amenable to algebraic treatment and is less of affected by fluctuations of sampling than most other measures of dispersion.
  e) It is independent of the origin and calculated from original data.
  f) It is a key note in sampling and provides a unit of measurement of the normal distribution.

email: akanda@du.ac.bd

Finding Variance and Standard Deviation in SPSS

- Open the data file and follow the instructions given below:
  1. From the menu at the top of the screen click on **Analyze**, then click on **Descriptive Statistics**, then **Frequencies…..**
  2. Choose and highlight the variables you are interested. Move these into the **Variable(s) box**. Then click **Statistics**…….
  3. In the Frequencies: Statistics dialog box select **Variance** and **Std. deviation**.
  4. Click **Continue** and then click **OK**.

email: akanda@du.ac.bd

Relative Measures of Dispersion

(i) Coefficient of Range

$$= \frac{\text{Range}}{\text{Highest value} + \text{Lowest value}} \times 100$$

(ii) Coefficient of Variation

$$= \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

(iii) Coefficient of Quartile Deviation

$$= \frac{\text{Quartile Deviation}}{\text{Median}} \times 100$$

(iv) Coefficient of Mean Deviation

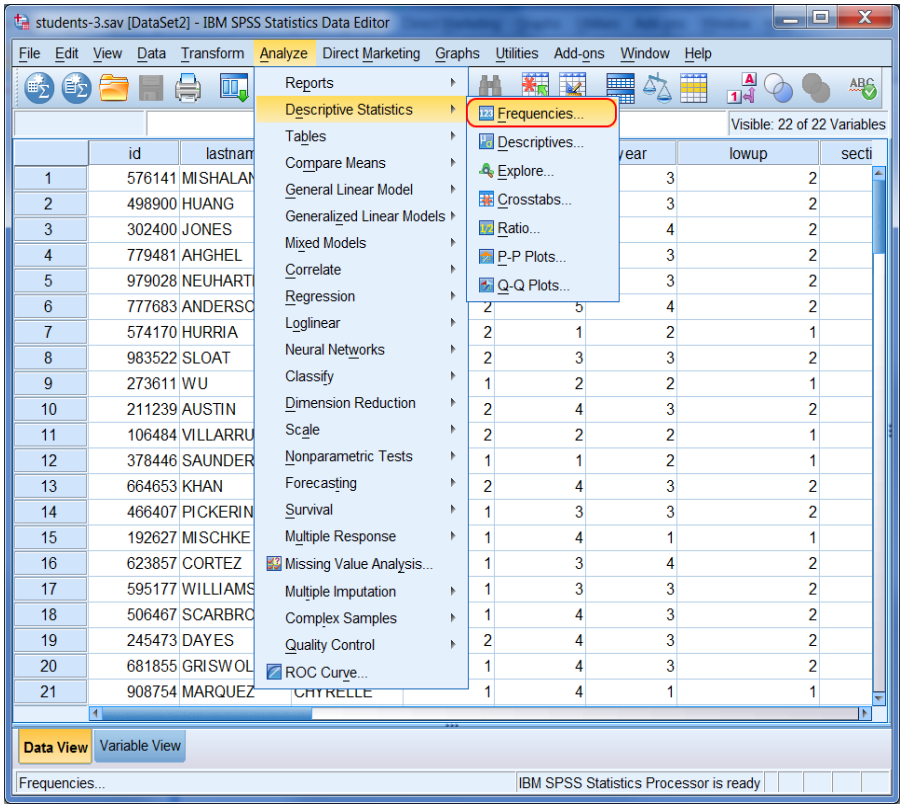
$$= \frac{\text{Mean Deviation}}{\text{Mean or Median}} \times 100$$

email: akanda@du.ac.bd

Measures of Dispersion in SPSS

- Open the data file (***students.sav***) and follow the instructions given below:

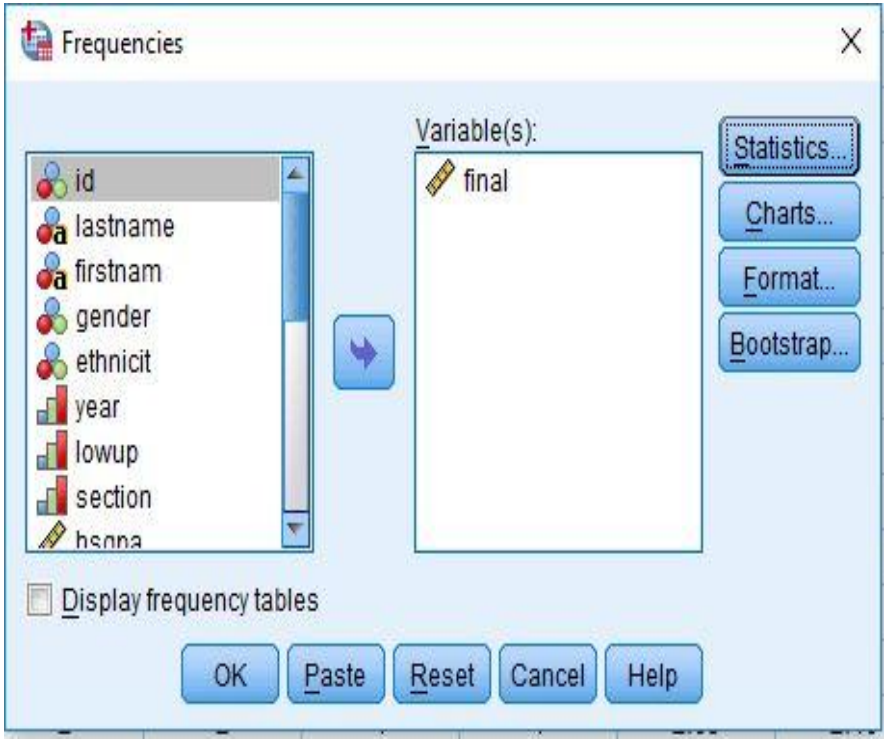
- Go to the **Statistics** menu, select the **Analyze** submenu, and then the **Descriptive Statistics** submenu, and then the **Frequencies** option.



email: akanda@du.ac.bd

Window: Frequencies

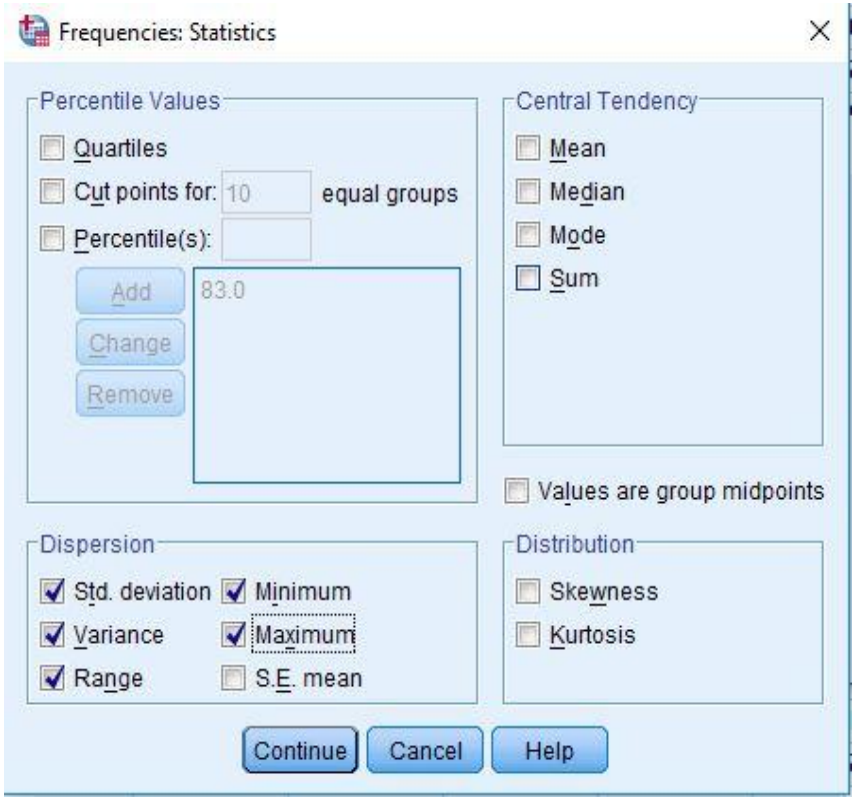
- This should open a window that looks like this:
- Select ***final*** as your variable. And then click on the **Statistics** button.
- This will open another window.



email: akanda@du.ac.bd

Window: Frequencies: Statistics

- In this window select **Std. deviation, Variance, Range, Minimum** and **Maximum**.
- Then click **Continue**.
- This will take you back to the previous window. Now click **OK**.



email: akanda@du.ac.bd

SPSS Output

- Now SPSS should open up an output window that includes a table that looks like this:

Statistics

final

N	Valid	105
	Missing	0
Std. Deviation		7.943
Variance		63.098
Range		35
Minimum		40
Maximum		75

email: akanda@du.ac.bd

Why do we study shape characteristics?

- The study of central tendency and dispersion provides us with valuable information relating to the central value as well as the variability of the distribution. Unfortunately, these measures fail to demonstrate how the observations are arranged and accumulated about the central value of the distribution. The arrangement and accumulation of the observations determine the characteristics of the distribution with respect to its shape and pattern. The study of these shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions.

email: akanda@du.ac.bd

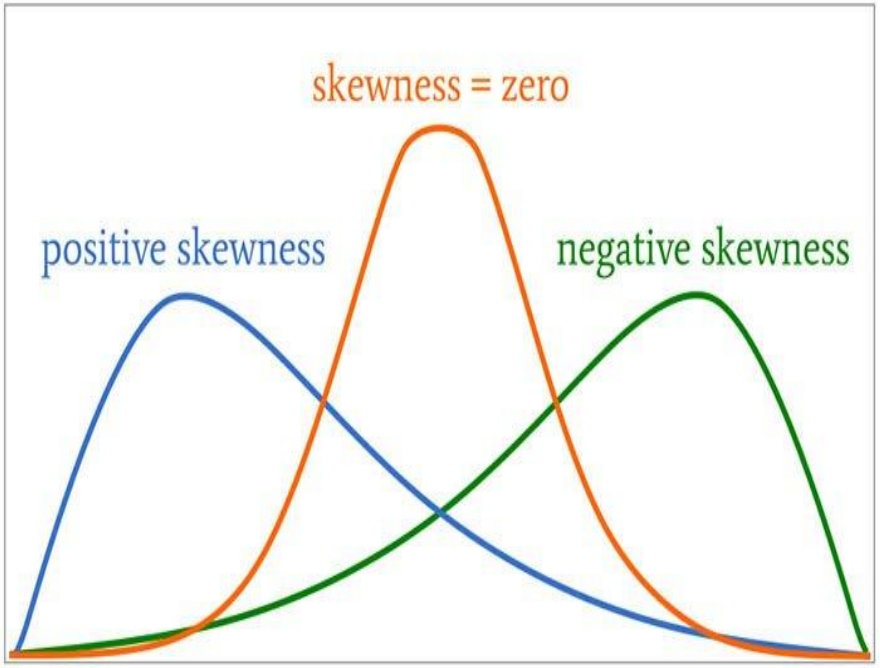
Shape Characteristics of a Distribution

- By shape characteristic of a distribution, we refer to the extent of its asymmetry and peakedness relative to an agreed upon standard and the study of these two characteristics is accomplished through what is known as the measures of skewness and kurtosis respectively.

email: akanda@du.ac.bd

Skewness

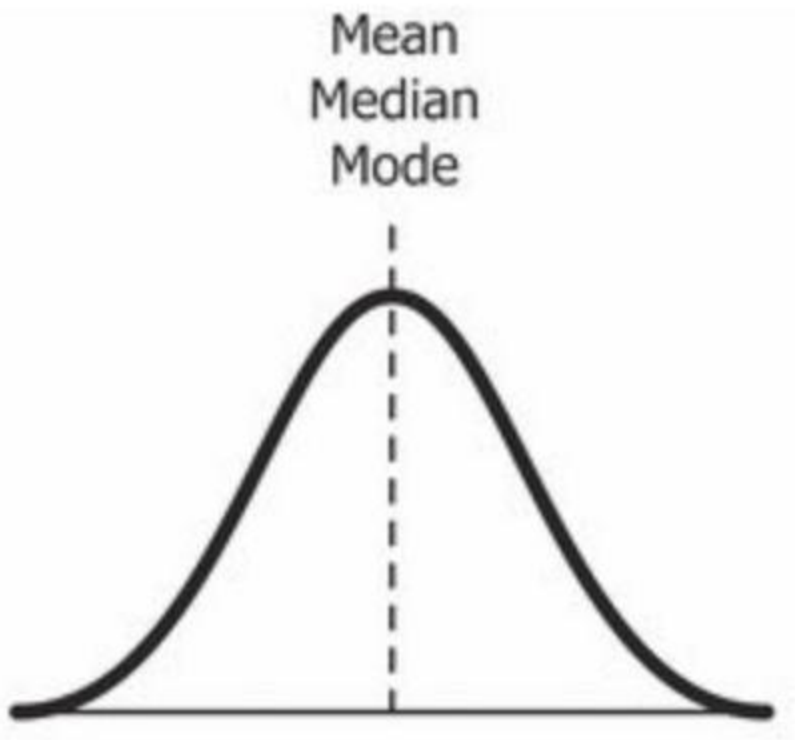
- Skewness is a measure of shape characteristics of a distribution. It is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or a series of data, is symmetric if it looks the same to the left and right of the center point. If the series is not symmetrical, it is said to be asymmetrical or skewed (i.e., departure from symmetry). As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.
- The departure from symmetry can be figured out by using the following diagram.



email: akanda@du.ac.bd

Symmetrical Distribution

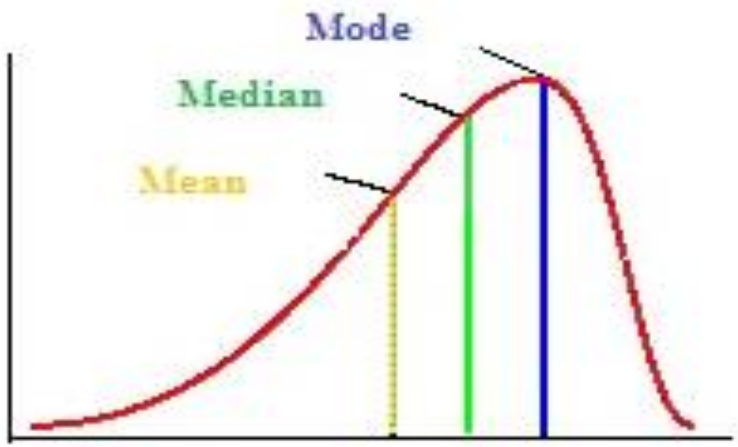
- A symmetric distribution is a type of distribution where the left side of the distribution mirrors the right side. In a symmetric distribution, the measures of its central tendency, i.e., arithmetic mean (AM), median (Me), and mode (Mo) are equal.
- There are many things, such as intelligence, height, weight, and blood pressure, that naturally follow a symmetric normal distribution.



email: akanda@du.ac.bd

Negatively Skewed Distribution

- A negatively skewed distribution is one in which the tail of the distribution shifts towards the left side, i.e., towards the negative side of the peak.
- In a negatively skewed distribution, Mean<Median<Mode.
- When the retirement age of employees is compared, it is found that most retire in their mid-sixties, or older. Thus, the distribution of most people will be near the higher extreme or the right side.

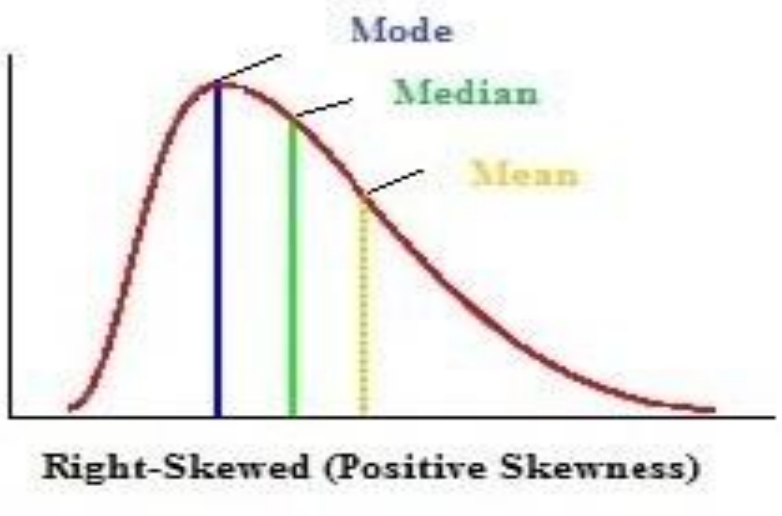


Left-Skewed (Negative Skewness)

email: akanda@du.ac.bd

Positively Skewed Distribution

- A positively skewed distribution is one in which the tail of the distribution shifts towards the right, i.e., it has a tail on the positive direction of the curve.
- In a positively skewed distribution, Mean>Median>Mode.
- For example, the frequency distribution of the number of school-going children in Bangladesh is a positively skewed distribution. Common pedagogical examples for positive skewness include people's incomes; mileage on used cars for sale; reaction times in a psychology experiment; house prices; number of accident claims by an insurance customer; number of children in a family etc.



email: akanda@du.ac.bd

Measure of Skewness

- A measure of skewness is defined by

$$\beta_1 = \frac{m_3^2}{m_2^3}$$

- where m_2 and m_3 are the second and the third moments about the mean of the distribution. In a symmetric distribution, β_1 will be zero. The greater the value of β_1 the more skewed the distribution.
- R. A. Fisher proposed a coefficient of skewness γ_1 (pronounced as Gamma one) which is defined as the square root of β_1. The coefficients of skewness (γ_1), we can have the following conclusions:
  a) If $\gamma_1 = 0$, the distribution is symmetrical.
  b) If $\gamma_1 > 0$, the distribution is positively skewed.
  c) If $\gamma_1 < 0$, the distribution is negatively skewed.

email: akanda@du.ac.bd

Interpretation of Skewness Output at SPSS

- A rule of thumb says:
  a) If the skewness is between -0.5 and 0.5, the data are fairly symmetrical (**normal distribution**).
  b) If the skewness is between -1 and -0.5 (**negatively skewed**) or between 0.5 and 1(**positively skewed**), the data are **moderately skewed**.
  c) If the skewness is less than -1(**negatively skewed**) or greater than 1(**positively skewed**), the data are **highly skewed**.

email: akanda@du.ac.bd

How to deal with skewed data

- Many statistical tests and machine learning models depend on normality assumptions. So, significant skewness means that data is not normal and that may affect your statistical tests or machine learning prediction power. In such cases, we need to transform the data to make it normal. Some of the common techniques used for treating skewed data:
  a) Log transformation
  b) Square root transformation
  c) Power transformation
  d) Exponential transformation
  e) Box-Cox transformation, etc.

email: akanda@du.ac.bd

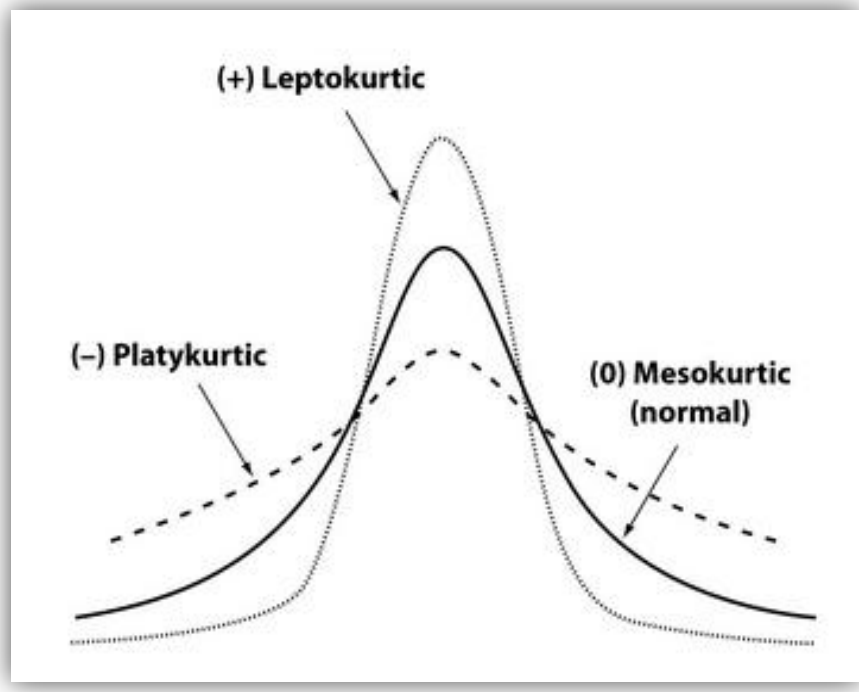
Kurtosis

- Kurtosis is another measure of the shape of a distribution. It is a Greek word which means bulginess. While skewness signifies the extent of asymmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve.

email: akanda@du.ac.bd

Types of Kurtosis

- Various frequency curves can be divided into three categories depending upon the shape of their peakness. These are:
  a) Mesokurtic
  b) Platykurtic
  c) Leptokurtic
- From the view point of kurtosis, the normal curve is **mesokurtic**, which means intermediate peakedness. Flat-topped curves on the other hand, are called **platykurtic**, and pronouncedly peaked curves are called leptokurtic. These three types are shown in the right diagrams.



email: akanda@du.ac.bd

Measure of Kurtosis

- A measure of kurtosis is defined by

$$\beta_2 = \frac{m_4}{m_2^2}$$

- where m_2 and m_4 are the second and the fourth moments about the mean of the distribution. This measure is a pure number and is always positive. This measure is also known as Karl Pearson's measure of kurtosis. From the coefficient of kurtosis, we can make the following conclusions:
  a) If the value of $\beta_2=3$ or $\gamma_2 = \beta_2-3=0$ then the distribution is mesokurtic. This value is taken as a standard against which the kurtosis of other distributions is judged.
  b) When $\beta_2>3$ or $\gamma_2=\beta_2-3>0$, the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. In the investment world, a leptokurtic distribution means that it is a high-risk investment.
  c) When $\beta_2<3$ or $\gamma_2=\beta_2-3<0$, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve. In the investment world, a platykurtic distribution means that it is a low-risk investment.
- A large value of kurtosis indicates a more serious outlier issue and hence may lead the researcher to choose alternative statistical methods.

email: akanda@du.ac.bd

Obtaining Skewness & Kurtosis in SPSS

- Open the data file and follow the instructions given below:
  - From the menu at the top of the screen click on **Analyze**, then click on **Descriptive Statistics**, then **Frequencies…..**
  - Choose and highlight the variables that you are interested in. Move these into the **Variable(s)** box. Then click **Statistics…….**
  - In the **Frequencies: Statistics** dialog box select **Skewness** and **Kurtosis.**
  - Click **Continue** and then click **OK**.

email: akanda@du.ac.bd

Interpretation of Kurtosis Output at SPSS

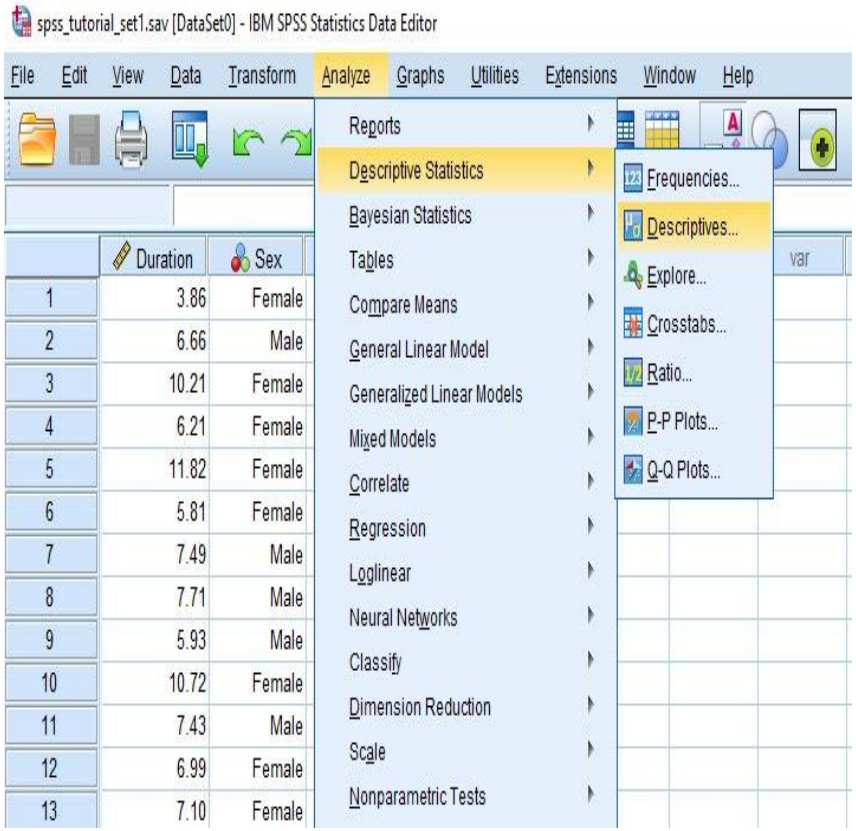
- If the value of kurtosis is less than -1.0, then the distribution is **Platykurtic**.
- If the value of kurtosis is greater than +1.0, then the distribution is **Leptokurtic**.
- If the value of kurtosis is between (-1,1), then the distribution is **Mesokurtic**.

email: akanda@du.ac.bd

Calculate Skewness and Kurtosis

- There are a number of different ways to calculate skewness and kurtosis in SPSS. We are going to use the Descriptives menu option.

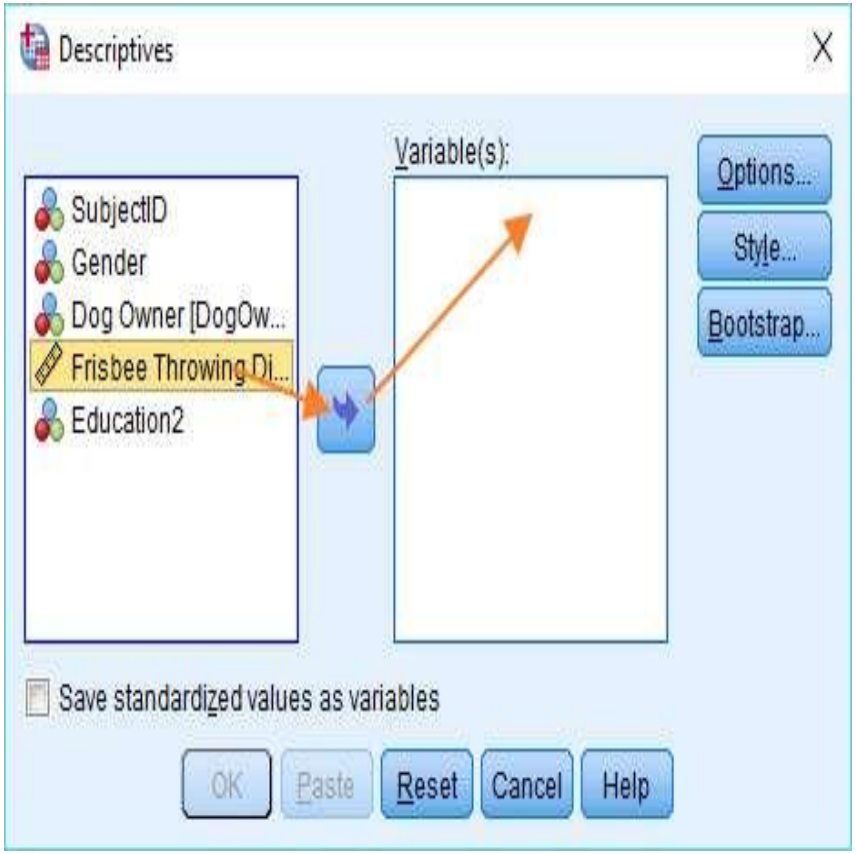
- To begin the calculation, click on Analyze -> Descriptive Statistics -> Descriptives...



email: akanda@du.ac.bd

Descriptives Dialog Box

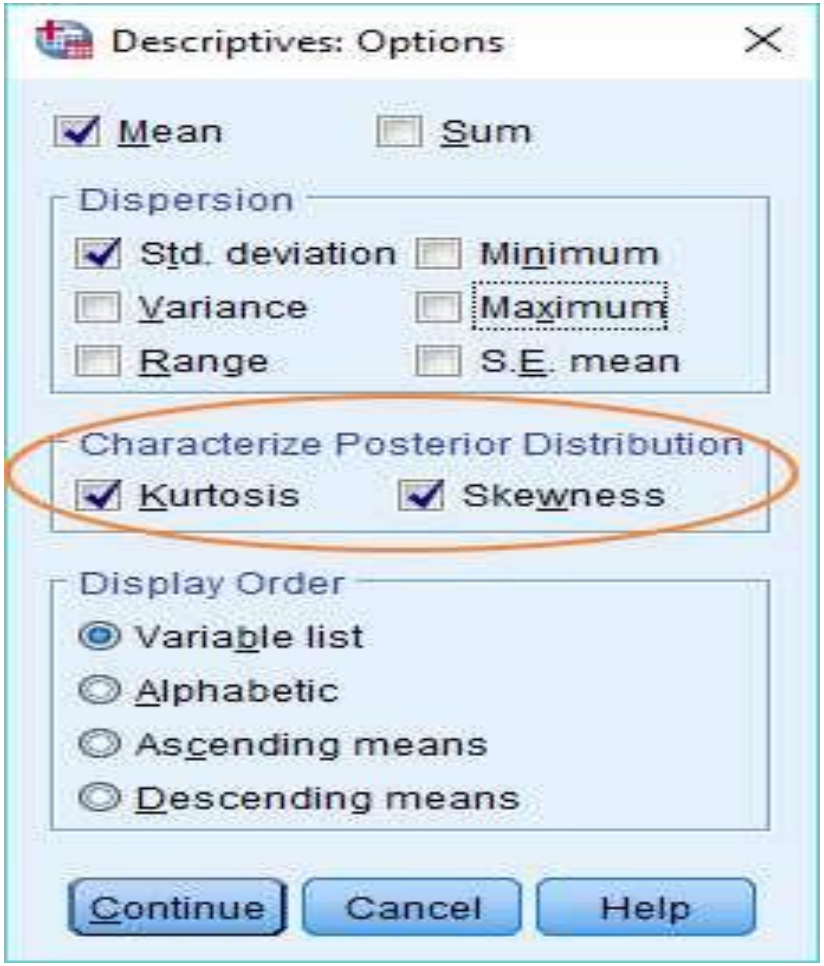
- This will bring up the **Descriptives dialog box**. You need to get the variable for which you wish to calculate skewness and kurtosis into the box on the right. You can drag and drop, or use the arrow button, as shown below.



email: akanda@du.ac.bd

Descriptives: Options Dialog Box

- Once you have got your variable into the right hand column, click on the Options button. This will bring up the **Descriptives: Options dialog box**, within which it is possible to choose a number of descriptive measures.
- To calculate skewness and kurtosis, just select the options (as right). You will notice that we have also instructed SPSS to calculate the mean and standard deviation.
- Once you have made your selections, click on **Continue**, and then on **OK** in the Descriptives dialog to tell SPSS to do the calculation.

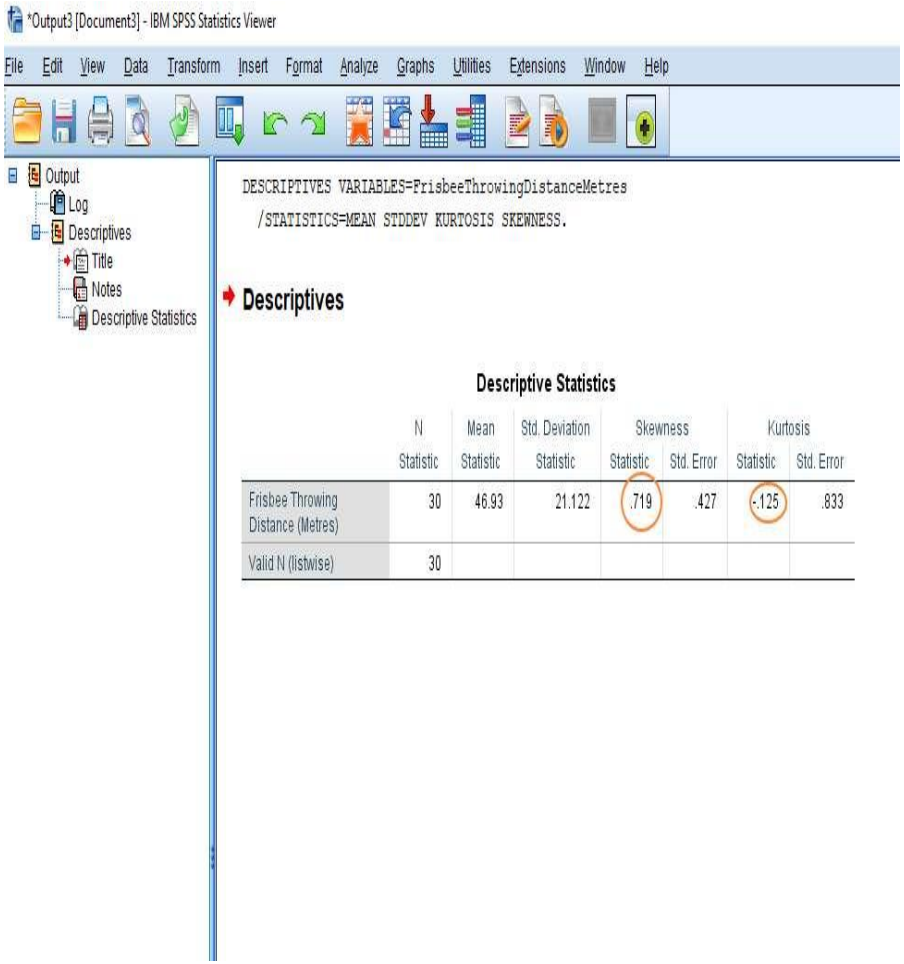


email: akanda@du.ac.bd

The Result

- The result will pop up in the SPSS output viewer. It will look something like this.

- This is fairly self-explanatory. The skewness statistic is .719 and kurtosis is -.125 (see right). **These indicate that the distribution is moderately skewed and mesokurtic.** You can also see that SPSS has calculated the mean (46.93 metres) and the standard deviation (21.122 metres). N represents the number of observations.



email: akanda@du.ac.bd

Thanks for your patience hearing…