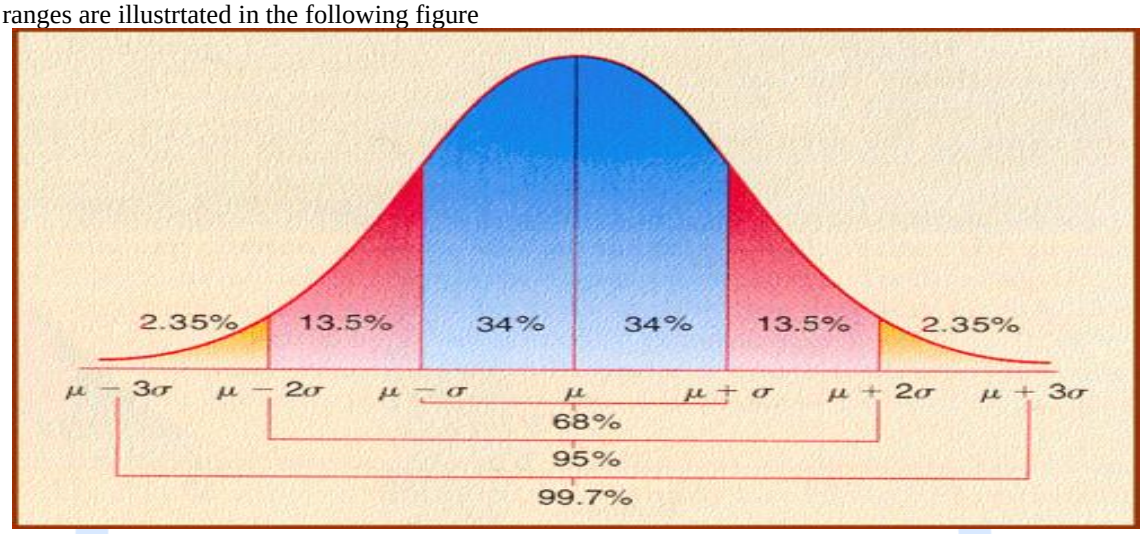
BRAC UNIVERSITY
Inspiring Excellence

Emperical Rule of Standard Deviation

For symmetrical, bell shaped frequency distribution (also called normal Curve), the range with in which a given percentage of values of the distribution are likely to fall within a specified number of standard deviation of the mean is determined as follows:

$\mu \pm \sigma$ covers approximately 68.27% of values in the data set
$\mu \pm 2\sigma$ covers approximately 95.45% of values in the data set
$\mu \pm 3\sigma$ covers approximately 99.73% of values in the data set

These ranges are illustrtated in the following figure



Problem:

The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate themean and standard deviation and determine the percentage of class that lie between $(i)\mu \pm \sigma$, $(ii)\mu \pm 2\sigma$ and $(iii)\ \mu \pm 3\sigma$. What percentage pf cases lie outside these limits?

Solution:

The calculation for mean and standard deviation are given in the following table

x	$x-\mu$	$(x-\mu)^2$
115		
122		
129		
113		
119		
124		
132		
120		
110		
116		

UNIVERSITY
Inspiring Excellence

$\mu = \frac{\sum x}{N} = \frac{????}{??} = 120$ and $\sigma^2 = \frac{\sum(x-\mu)^2}{N} = ??? = 43.6$

Therefore, $\sigma = \sqrt{\sigma^2} = \sqrt{43.3} = 6.60$

The percentage of cases that lie between a given limit are as follows:

Interval	Values within Interval	Percentage of population	Percentage falling Outside
$\mu \pm \sigma = 120 \pm 6.60$ = 113.4 and 126.6	115, 116, 119, 120, 122, 124	60%	40%
$\mu \pm 2\sigma$ =$120 \pm 2(6.60)$ = 106.80 and 133.20	110, 113, 115, 116, 119, 120, 122, 124, 129, 132	100%	Nil

Shape characteristics of a distribution

The study of shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions. 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 (that is asymmetry and peakedness) is accomplished through what is known as the measures of skewness and kurtosis.

We study two characteristics in the following section: Skewness and Kurtosis

Skewness:

The term skewness means the lack of symmetry. The skewness may be either positive or negative. When the skewness is positive the associated distribution is called positively skewed. When the skewness is negative the associated distribution is negatively skewed.

Now some very simple measures of skewness is shown here:

Method 1	If for a distribution $Mean > Median > Mode \Rightarrow$ The distribution is positively skewed $Mean < Median < Mode \Rightarrow$ The distribution is negatively skewed $Mean = Median = Mode \Rightarrow$ The distribution is symmetric
Method 2	Pearson's coefficient of skew ness $Sk_p = \frac{Mean - Mode}{SD} = \frac{3(Mean - Median)}{SD}$ Then if $Sk_p > 0 \Rightarrow$ The distribution is positively skewed $Sk_p < 0 \Rightarrow$ The distribution is negatively skewed $Sk_p = 0 \Rightarrow$ The distribution is symmetric.

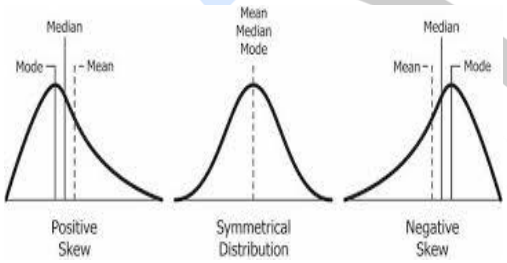


Figure: Skewness of Distribution

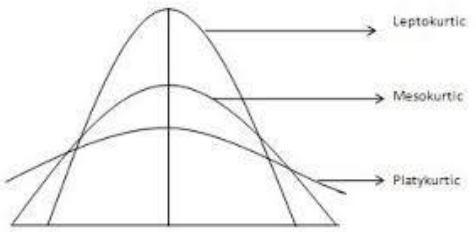


Figure: Kurtosis of distribution

Kurtosis:

There is considerable variation among symmetrical distributions. For instance, they can differ markedly in terms of peaked ness. This is what we call kurtosis. Kurtosis, as defined by Spiegel (Spiegel: Theory and Problems of Statistics) is the degree of peaked ness of a distribution, usually taken in relation to a normal distribution.

- A curve having relatively **higher peak** than the normal curve, is known as **leptokurtic**.
- A curve, which is **neither too peaked nor too flat topped**, is known as **mesokurtic**.
- A curve that is **more flat topped** than the normal curve is called **platykurtic**.

BRAC UNIVERSITY Inspiring Excellence

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

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. It helps to determine if the curve is more or less high as compared to the normal curve

Question:
If for a distribution Mean=18, Median=32 and Mode=36 $\Rightarrow$ the distribution is _ _ _ _ _ _ _ _ _ _ _ _ skewed.
a. Positively b. Symmetrically c. None d. Negatively

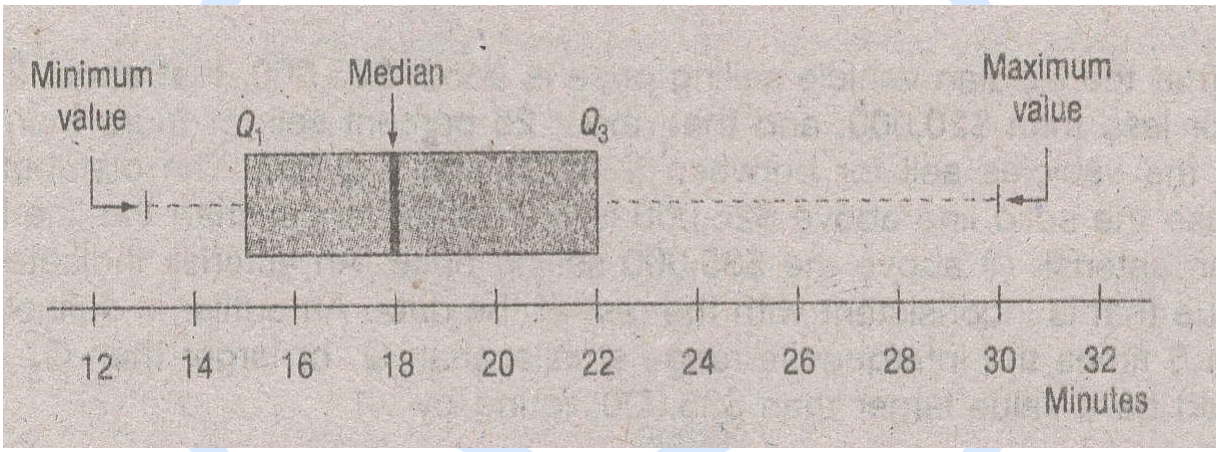
BRAC
UNIVERSITY
Inspiring Excellence

Box Plot:
A box plot is a graphic display that shows the general shape of a variable's distribution. It is based on five descriptive statistics: the minimum value, the first quartile (Q_1), Median, third quartile (Q_3) and the maximum value.

Example:
Pizza Hut offers free delivery of its pizza within 15 miles. Mr. Rahman the owner wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes
Q_1 = 15 minutes
Median = 18 minutes
Q_3 = 22 minutes
Maximum value = 30 minutes

Develop a boxes plot for the delivery times. What conclusions can you make about make about the delivery times?



Solution:
In order to draw box plot follow the steps mentioned below:

Step 1: Create an appropriate scale along the horizontal axis.

Step 2: Draw a box that starts at Q_1 (15 minutes) and ends at Q_3 (22 minutes)

Step 3: Place a vertical line to represent the median (18 minutes)

Step 4: Extend the horizontal lines[1] from the box out to the minimum value (13 minutes) and the maximum value (30 minutes)

Interpretation of the Box Plot:

- The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the inter quartile range[2]. That shows the spread or dispersion of the majority of deliveries.
- The box plot also reveals that the distribution of the delivery times is positively skewed. The guiding principle for such conclusion are
  - The dashed line to the right of the box from 22 minutes (Q_3) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes (Q_1) to the minimum value of 13 minutes.
  - The median is not in the middle in the center of the box. The distance from the first quartile to the median is smaller than the distances from the median to the third quartile.

Inspiring Excellence

[1] These horizontal lines outside of the box are sometimes called "whiskers" because the looks a bit like a cat's whiskers.
[2] The inter quartile range is the distance between the first and the third quartile.

Test yourself
Quartile and Box Plot

Question:

Construct a box plot for the data given below and hence comment on the skewness of the distribution:

99	75	84	33	45	66	97	69	55	61
72	91	74	93	54	76	62	91	77	68

Hints: Calculate Median, then the median of the 1st half, then the median of the 2nd half. Then proceed according to the instruction

Merits and Demerits of different Measures of Dispersion:

	Merits	Demerits	
Range	• Easy to understand and calculate. • It is based only on extreme observations and no detail in formations is required. • It gives us a quick idea of the variability of a set of data.	• It is not based on all observation. • Range does not give any indication of the character of the distribution with in the two extreme observations. • Range is subject of fluctuations from sample to sample. • Cannot be computed in case of open-end class.	**Range**
Quartile deviation	• It is superior to range as a measure of dispersion. • It is applicable in Open-end class. • Easy to understand and compute. • Not affected by extreme values.	• It ignores 50% of items that is the first 25% and last 25% of observations. • Very much affected by sampling fluctuations. • Not suited for further algebraic treatment.	**Quartile deviation**
Variance	• Rigidly defined. • Based upon all observation. • Easy to understand • Less affected by sampling fluctuations. • Suitable for further algebraic treatment.	• Difficult to calculate. • Affected by extreme values. • Difficult to calculate for open-end class.	**Variance**
	Merits	**Demerits**	

Inspiring Excellence

For any queries related to this presentation please contact
IFTEKHAR Mohammad Shafiqul Kalam
Assistant Professor
Department of Mathematics and Natural Sciences
Email: imskalam@bracu.ac.bd