Elements of Statistics and Probability

STA 201

S M Rajib Hossain

MNS, BRAC University

Lecture-5 &6

Measures of Dispersion

In statistics, the measures of dispersion help to interpret the variability of data i.e.to know how much homogenous or heterogeneous the data is. In simple terms, it shows how scattered the variable is.

Types of Measures of Dispersion

There are two main types of dispersion methods in statistics which are:

- ✓ Absolute Measure of Dispersion
- ✓ Relative Measure of Dispersion

Absolute Measure of Dispersion

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

- ✓ Range
- ✓ Variance
- ✓ Standard Deviation
- ✓ Quartile Deviation
- ✓ Mean Deviation

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

- ✓ Co-efficient of Range
- ✓ Co-efficient of Variation (C.V)
- ✓ Co-efficient of Standard Deviation
- ✓ Co-efficient of Quartile Deviation
- ✓ Co-efficient of Mean Deviation

Range

It is simply the difference between the maximum value and the minimum value given in a data set.

Example: 1, 3, 5, 6, 7

Range = 7 -1

= 6

Variance

The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean (average), and thus from every other number in the set. Variance is often depicted by this symbol: σ^2

In our study, we have two types of variances,

- Population variance: Let, $x_1, \ldots, x_N$ be N observations in a population and μ be the population mean. Then the population variance denoted by σ^2 is defined as,

$$\sigma^2 = \sum_{1}^{N} \frac{(x_i - \mu)^2}{N}$$

For grouped data $\sigma^2 = \sum_1^N \frac{f_i(x_i - \mu)^2}{\sum_1^n f_i}$

- Sample variance: Let, $x_1, \ldots, x_n$ be n observations in a sample and $\bar{x}$ be the sample mean. Then the sample variance denoted by s^2 is defined as,

$$s^2 = \sum_{1}^{n} \frac{(x_i - \bar{x})^2}{n - 1}$$

For grouped data $s^2 = \sum_1^n \frac{f_i(x_i - \bar{x})^2}{\sum_1^n f_i - 1}$

Standard deviation

The positive square root of variance is called standard deviation.

- Population standard deviation $\sigma = \sqrt{\sum_1^N \frac{(x_i-\mu)^2}{N}}$

For grouped data $\sigma = \sqrt{\sum_1^N \frac{f_i(x_i-\mu)^2}{\sum_1^n f_i}}$

- Sample standard deviation $s = \sqrt{\sum_1^n \frac{(x_i-\bar{x})^2}{n-1}}$

For grouped data $s = \sqrt{\sum_1^n \frac{f_i(x_i-\bar{x})^2}{\sum_1^n f_i-1}}$

Q1

Find the variance and standard deviation for the following data

Class	Mid value (x_i)	Freq (f_i)	$f_i x_i$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
2.735-3.135	2.935	6	17.61	1.26068	7.564079
3.135-3.535	3.335	7	23.345	0.52244	3.657079
3.535-3.935	3.735	12	44.82	0.1042	1.250398
3.935-4.335	4.135	12	49.62	0.00596	0.071518
4.335-4.735	4.535	10	45.35	0.22772	2.277198
4.735-5.135	4.935	7	34.545	0.76948	5.386359
5.135-5.535	5.335	3	16.005	1.63124	4.89372
Total		$\sum_1^n f_i$ =57	$\sum_1^n f_i x_i$ =231.295	4.521719	$\sum_1^n f_i(x_i - \bar{x})^2$ =25.10035

Sample mean $\bar{x} = \frac{\sum_1^n f_i x_i}{\sum_1^n f_i}$

$= \frac{231.295}{57}$

= 4.0578

Sample variance $s^2 = \sum_1^n \frac{f_i(x_i-\bar{x})^2}{\sum_1^n f_i-1}$

$$= \frac{25.10035}{57-1}$$

$$= 0.448221$$

Sample standard deviation $s = \sqrt{0.448221}$

$$= 0.669$$

Mean Deviation

Mean deviation is used to compute how far the values in a data set are from the center point. Mean, median, and mode all form center points of the data set. In other words, the mean deviation is used to calculate the average of the absolute deviations of the data from the central point.

In case of mean

$$\text{M.A.D } (\bar{x}) = \frac{\sum_1^n |x_i-\bar{x}|}{n}$$

For grouped data $\quad \text{M.A.D } (\bar{x}) = \frac{\sum_1^n f_i|x_i-\bar{x}|}{\sum_1^n f_i}$

In case of median

$$\text{M.A.D } (Me) = \frac{\sum_1^n |x_i-Me|}{n}$$

For grouped data $\quad \text{M.A.D } (Me) = \frac{\sum_1^n f_i|x_i-Me|}{\sum_1^n f_i}$

In case of mode

$$\text{M.A.D } (Mo) = \frac{\sum_1^n |x_i-Mo|}{n}$$

For grouped data $\quad \text{M.A.D } (Mo) = \frac{\sum_1^n f_i|x_i-Mo|}{\sum_1^n f_i}$

Q2

Find the mean deviation for the following data

Class	Mid value (x_i)	Freq (f_i)	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
15-25	20	25	500	13.68	342
25-35	30	54	1620	3.68	198.72
35-45	40	34	1360	6.32	214.88
45-55	50	20	1000	16.32	326.4
Total		$\sum_1^n f_i$ =133	$\sum_1^n f_i x_i$ =4480		$\sum_1^n f_i\lvert x_i - \bar{x}\rvert$ =1082

Sample mean $\bar{x} = \frac{\sum_1^n f_i x_i}{\sum_1^n f_i}$

$=\frac{4480}{133}$

$=33.68$

M.A.D ($\bar{x}$)$=\frac{\sum_1^n f_i|x_i-\bar{x}|}{\sum_1^n f_i}$

$=\frac{1082}{133}$

$= 8.14$

Coefficient of variation

The coefficient of variation is the ratio of the standard deviation to the mean.

It is usually expressed as percentage. Mathematically,

$$CV = \frac{\sigma}{\mu} \times 100; For\ population$$

$$CV = \frac{s}{\bar{x}} \times 100; For\ Sample$$

Q3

Find the coefficient of variation for the two plants of a factory for the given data and interpret the results.

Two plants C and D of a factory show the following results about the number of workers and the wages paid to them.

No. of workers	5000	6000
Average monthly wages	\$2,500	\$2,500
Standard deviation	9	10

Solution:

To Find: Which plant has greater variability.

For this, we need to find the coefficient of variation. The plant that has a higher coefficient of variation will have greater variability.

Using coefficient of variation formula, $CV = \frac{s}{\bar{x}} \times 100$

Coefficient of variation for plant C

CV = (9/2500) × 100

CV = 0.36%

Now, Coefficient of variation for plant D

CV = (10/2500) × 100

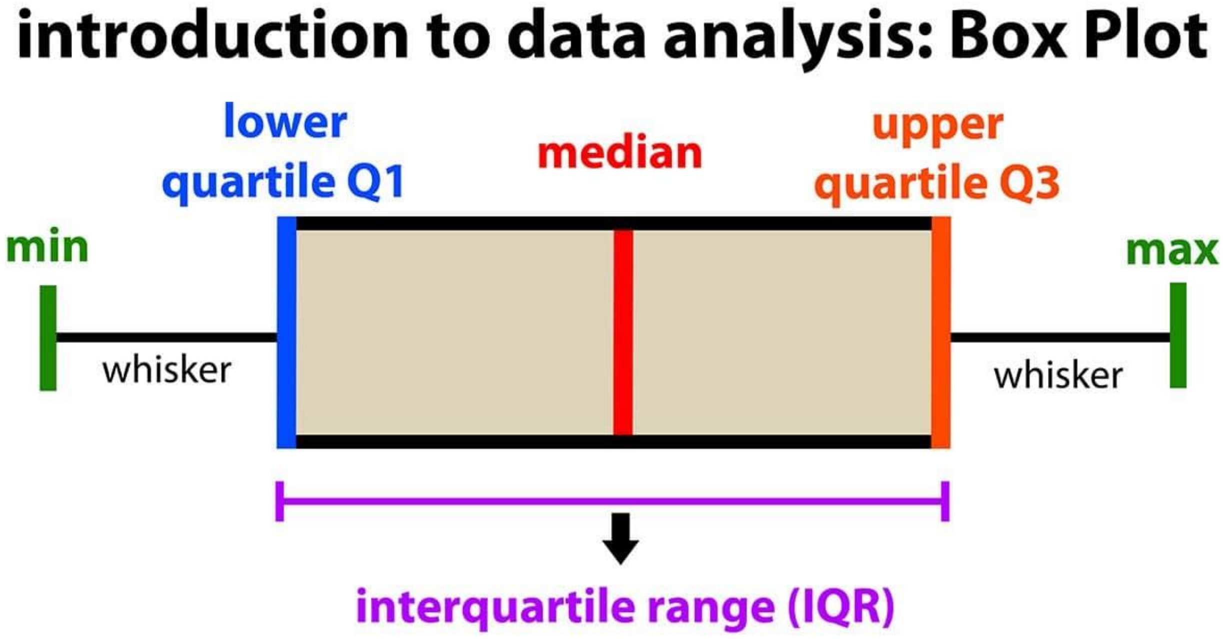
CV = 0.4%

Plant C has CV = 0.36 and plant D has CV = 0.4

Hence plant D has greater variability in individual wages.

Box plot

When we display the data distribution in a standardized way using 5 summary – minimum, Q_1 (First Quartile), Q_2 (median) Q_3 (third Quartile), and maximum, it is called a Box plot. It is also termed as box and whisker plot.


introduction to data analysis: Box Plot
lower quartile Q1
median
upper quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

Minimum Score

The lowest score, excluding outliers (shown at the end of the left whisker).

Lower Quartile

Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).

Median

The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value, and half are less.

Upper Quartile

Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.

Maximum Score

The highest score, excluding outliers (shown at the end of the right whisker).

Whiskers

The upper and lower whiskers represent scores outside the middle 50% (i.e., the lower 25% of scores and the upper 25% of scores).

The Interquartile Range (IQR)

The box plot shows the middle 50% of scores (i.e., the range between the 25th and 75th percentile).

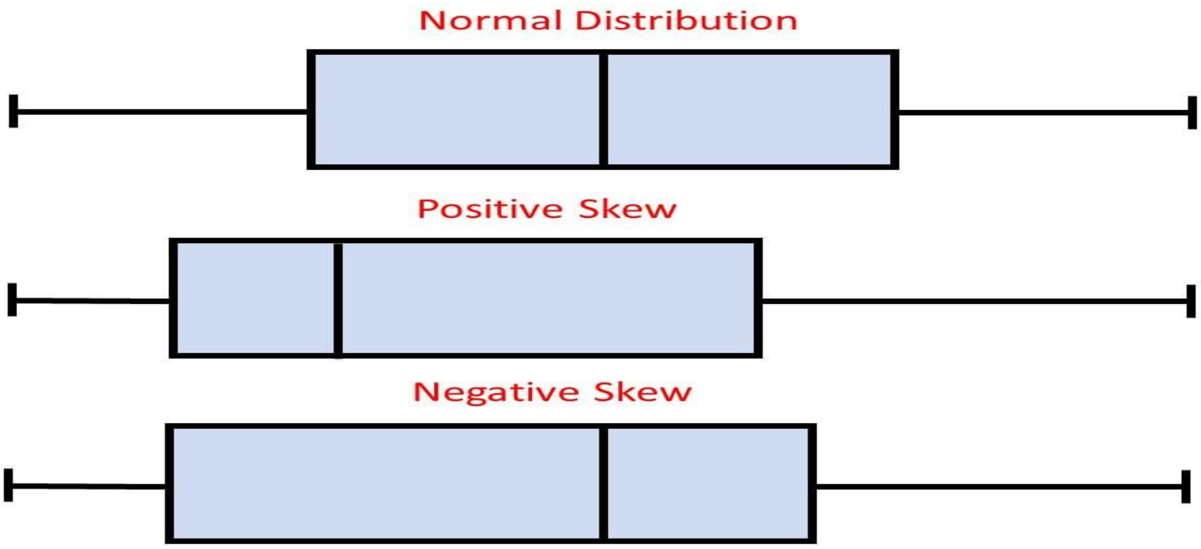
Why Are Box Plots Useful?

Box plots are useful as they show the average score of a data set:

The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value, and half are less.

Box plots are useful as they show the skewness of a data set:

The box plot shape will show if a statistical data set is normally distributed or skewed.



When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.

When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).

When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

Box plots are useful as they show outliers within a data set:

An outlier is an observation that is numerically distant from the rest of the data.

When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

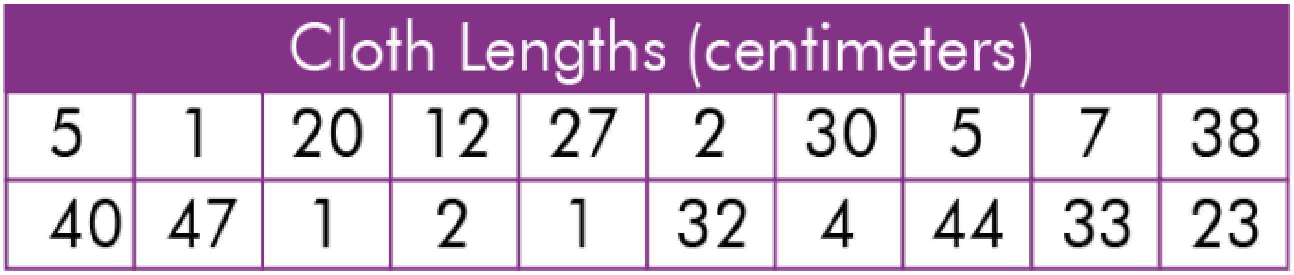
Stem and leaf plot

A stem and leaf plot, also known as a stem and leaf diagram, is a way to arrange and represent data so that it is simple to see how frequently various data values occur. It is a plot that displays ordered numerical data.

A stem and leaf plot is shown as a special table where the digits of a data value are divided into a stem (first few digits) and a leaf (usually the last digit). The symbol '|' is used to split and illustrate the stem and leaf values. For instance, 105 is written as 10 on the stem and 5 on the leaf. This can be written as 10 | 5. Here, 10 | 5 = 105 is called the key. The key depicts the data value a stem and leaf represent.

Problem 1

Construct a stem-and-leaf plot for the data in the table.



Cloth Lengths (centimeters)									
5	1	20	12	27	2	30	5	7	38
40	47	1	2	1	32	4	44	33	23

Solution:

Step 1: Sort the data values: 1, 1, 1, 2, 2, 4, 5, 5, 7, 12, 20, 23, 27, 30, 32, 33, 38, 40, 44, 47

Step 2: Choose the stems and the leaves. As the data values range from 1 to 47, use the tens digits for the stems and the ones digits for the leaves. Be sure to include the key.

Step 3: Write the stems to the left of the vertical line from the top to bottom.

Step 4: Write the leaf values corresponding to each stem to the right of the vertical line.

Stem	Leaf
0	1 1 1 2 2 4 5 5 7
1	2
2	0 3 7
3	0 2 3 8
4	0 4 7

Key: 0| 1 = 1 cm

Problem 2

The stem-and-leaf plot below shows the quiz scores of students.

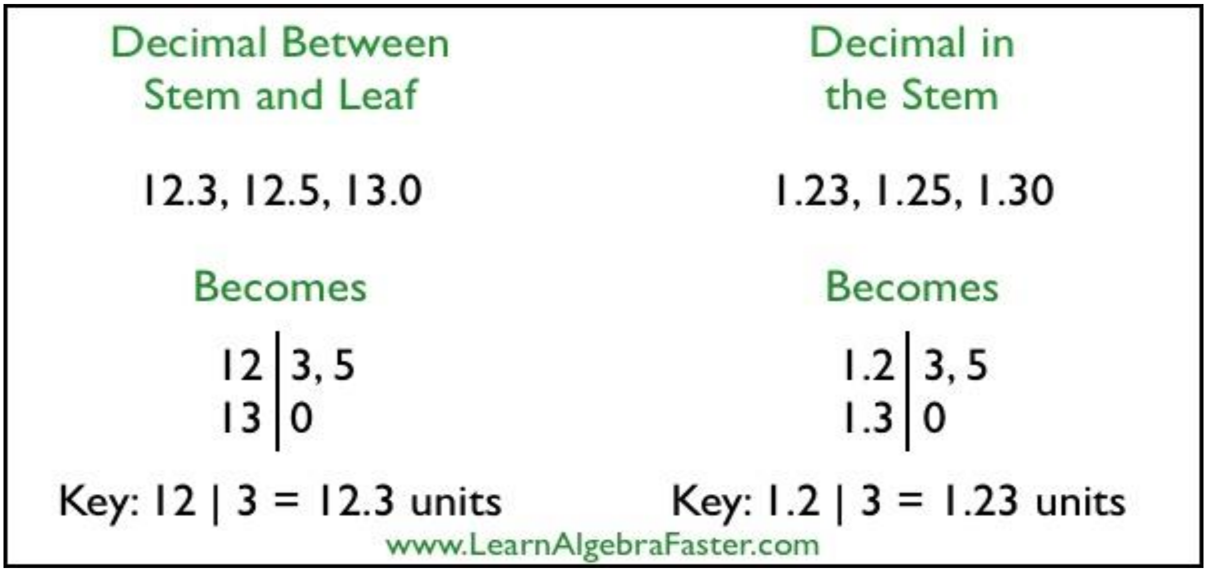
Stem	Leaf
6	6
7	0 5 7 8
8	1 1 3 4 4 6 8 8 9
9	0 2 9
10	0

Key: 9 | 2 = 9.2

(a) Find the number of students who scored less than 9 points?

(b) Find the number of students who scored a minimum of 9 points?

Decimal or fractional values can also be placed in a stem-and-leaf plot.



Decimal Between Stem and Leaf

12.3, 12.5, 13.0

Becomes

```
12 | 3, 5
13 | 0
```

Key: 12 | 3 = 12.3 units

Decimal in the Stem

1.23, 1.25, 1.30

Becomes

```
1.2 | 3, 5
1.3 | 0
```

Key: 1.2 | 3 = 1.23 units

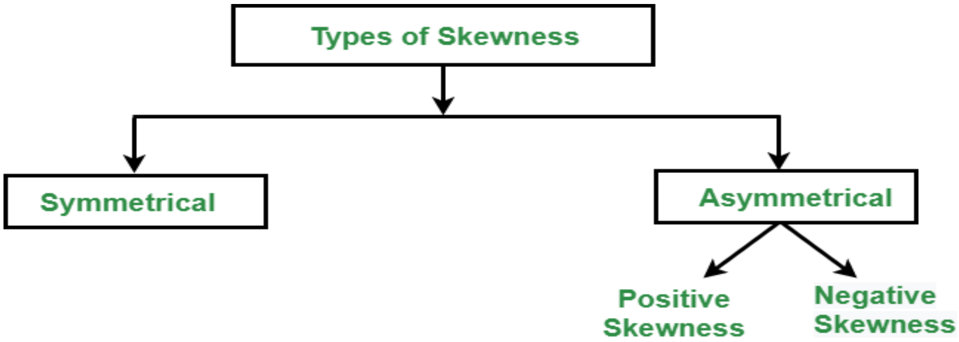
www.LearnAlgebraFaster.com

Skewness and Kurtosis

Skewness is a measure of the asymmetry of a distribution.

Skewness is an important statistical technique that helps to determine asymmetrical behavior than of the frequency distribution, or more precisely, the lack of symmetry of tails both left and right of the frequency curve. A distribution or dataset is symmetric if it looks the same to the left and right of the center point.

Types of skewness: The following figure describes the classification of skewness:

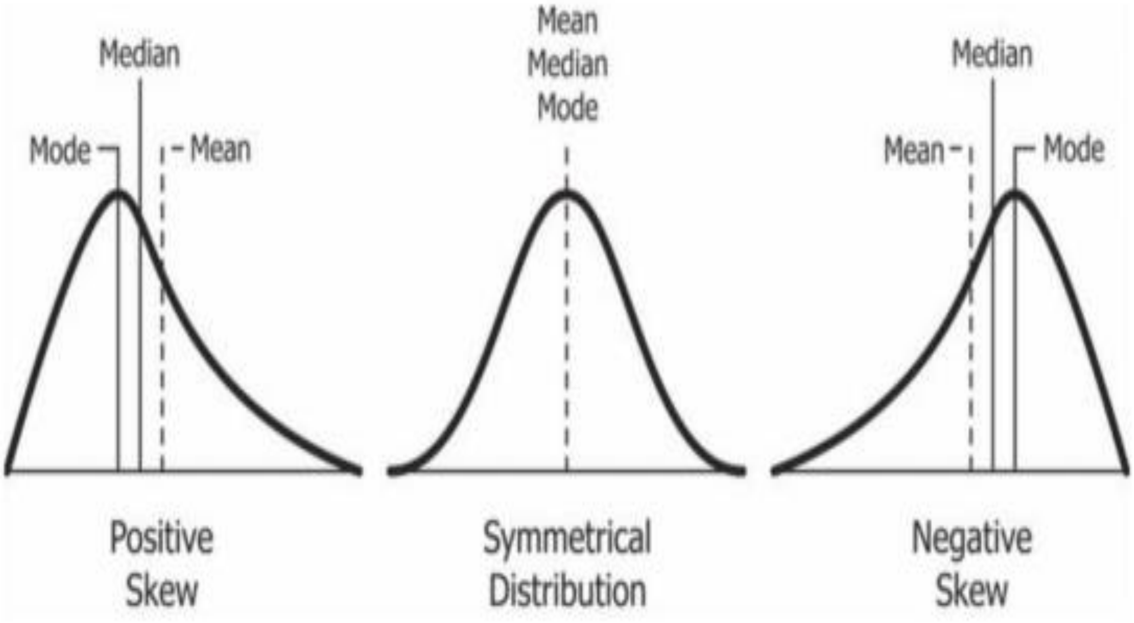


Symmetric Skewness: A perfect symmetric distribution is one in which frequency distribution is the same on the sides of the center point of the frequency curve. In this, Mean = Median = Mode. There is no skewness in a perfectly symmetrical distribution.

Asymmetric Skewness: A asymmetrical or skewed distribution is one in which the spread of the frequencies is different on both the sides of the center point or the frequency curve is more stretched towards one side or value of Mean. Median and Mode falls at different points.

Positive Skewness: In this, the concentration of frequencies is more towards higher values of the variable i.e. the right tail is longer than the left tail.

Negative Skewness: In this, the concentration of frequencies is more towards the lower values of the variable i.e. the left tail is longer than the right tail.



Measures of skewness

In studying skewness of a distribution, the first thing that we would like to know whether the distribution is positively skewed or negatively skewed. The second thing is to measure the degree of skewness. The simplest measure of skewness is the Pearson's coefficient of skewness defined as:

Pearson's coefficient of skewness = $\frac{mean-mode}{standard\ deviation}$

In many instances, mode cannot be uniquely defined, in which case, the above formula cannot be applied. It is observed that for a moderately skewed distribution, the following relationship holds:

mean-mode=3(mean-median)

Using this relation, the Pearson's coefficient of skewness assumes the following modified form

Pearson's coefficient of skewness = $\frac{3(mean-median)}{standard\ deviation}$

Interpretation

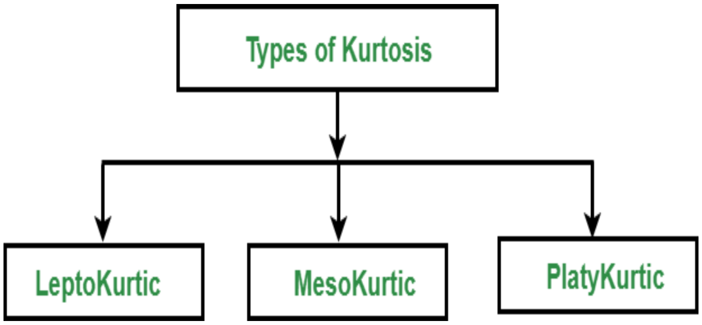
If the value of Pearson's coefficient of skewness is zero, the distribution is symmetric.

If the value of Pearson's coefficient of skewness is positive, the distribution is positively skewed.

If the value of Pearson's coefficient of skewness is negative, the distribution is negatively skewed.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

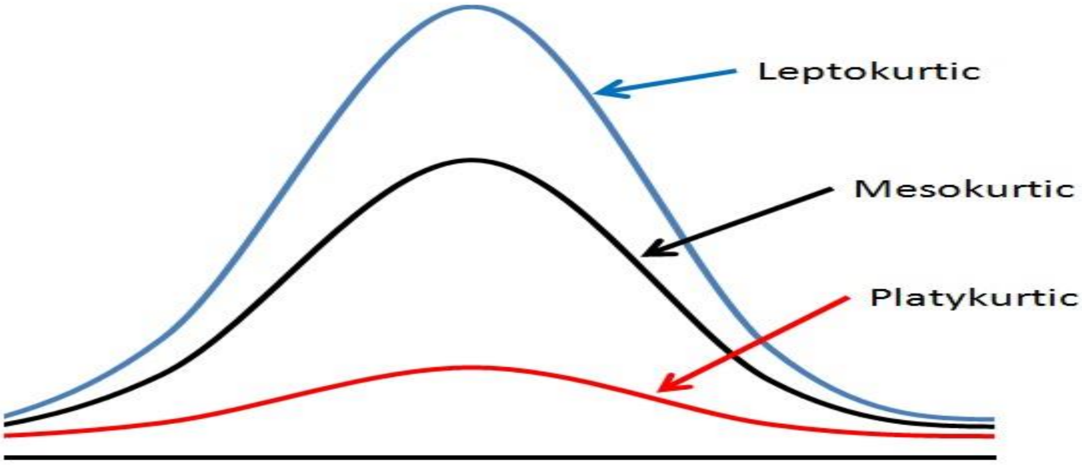
Types of kurtosis: The following figure describes the classification of kurtosis:



Leptokurtic: Leptokurtic is a curve having a high peak than the normal distribution. In this curve, there is too much concentration of items near the central value.

Mesokurtic: Mesokurtic is a curve having a normal peak than the normal curve. In this curve, there is equal distribution of items around the central value.

Platykurtic: Platykurtic is a curve having a low peak than the normal curve is called platykurtic. In this curve, there is less concentration of items around the central value.


Leptokurtic
Mesokurtic
Platykurtic