**Notes – Ch 3 Descriptive Statistics: Numerical Measures**

**Measures of Location of data are Mean, Median, Mode, Percentile, Quartile**

**1. Mean -** Perhaps the most important measure of location is the mean, or average value, for a variable. The mean provides a measure of central location for the data. If the data are for a sample, the mean is denoted by if the data are for a population, the mean is denoted by the Greek letter μ. In statistical formulas, it is customary to denote the value of variable x for the first observation by x1, the value of variable x for the second observation by x2, and so on. In general, the value of variable x for the *ith* observation is denoted by *xi*. For a sample with n observations, the formula for the sample mean is as follows.

In the preceding formula, the numerator is the sum of the values of the n observations.

The Greek letter Σ is the summation sign.

**Advantages:**

* Familiar concept
* Can be used for comparison purpose as all data has a unique mean

**Disadvantages:**

* It is affected by extreme values in case of skewed data
* Tedious to calculate for large data set
* Not possible to calculate for open-ended classes at higher end or lower end of the scale

**2. Median -** The median is another measure of central location. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value). With an odd number of observations, the median is the middle value. An even number of observations has no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations. For convenience the definition of the median is restated as follows.

Arrange the data in ascending order (smallest value to largest value).

(a) For an odd number of observations, the median is the middle value.

(b) For an even number of observations, the median is the average of the two middle values.

**Advantages:**

* Not affected by extreme values. Mostly appropriate to use while finding central values for income or property prices
* Can be used for open-ended classes also
* Central location can be located for qualitative data for ordinal scale of measurement

eg good, very good, extremely good

* Can be derived graphically as intersection of 2 ogives

**Disadvantages:**

* Certain statistical procedure using median are more complex compared to mean
* Arranging in ascending or descending order is tedious if data set is large

**3. Mode -** The mode is the value that occurs with greatest frequency. Situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists. If the data contain exactly two modes, we say that the data are bimodal. If data contain more than two modes, we say that the data are multimodal. In multimodal cases the mode is almost never reported because listing three or more modes would not be particularly helpful in describing a location for the data.

**Uses of mode:**

When we are interested in knowing the consumers preferences for different brands of television sets or different kinds of advertising, the choice should go in favor of mode. The use of mean and median would not be proper. Let us take another example. Suppose we invite applications for a certain vacancy in a company. A large number of candidates apply for that post. We are now interested to know as to which age or age group has the largest concentration of applicants. Here, obviously the mode will be the most appropriate choice.

**Advantages:**

* Central location can be located for qualitative data for ordinal scale of measurement
* Not affected by extreme values. Mostly appropriate to use while finding central values for income or property prices
* Can be used for open-ended classes also

**Disadvantages:**

* At times data has no value occurring more than once or at times all values occur the same no of times
* Difficult to interpret and compare a data set with multiple modes

**Relation between mean, median and mode**

* Mode = 3 \* Median – 2 \* Mean

**4. Percentile -** A percentile provides information about how the data are spread over the interval from the smallest value to the largest value. For data that do not contain numerous repeated values, the *pth* percentile divides the data into two parts. Approximately p percent of the observations have values less than the *pth* percentile; approximately (100 - p) percent of the observations have values greater than the *pth* percentile. The *pth* percentile is formally defined as follows.

**Percentile:** The *pth* percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100 - p) percent of the observations are

greater than or equal to this value.

Colleges and universities frequently report admission test scores in terms of percentiles.

For instance, suppose an applicant obtains a raw score of 54 on the verbal portion of an admission test. How this student performed in relation to other students taking the same test may not be readily apparent. However, if the raw score of 54 corresponds to the 70th percentile, we know that approximately 70% of the students scored lower than this individual and approximately 30% of the students scored higher than this individual.

The following procedure can be used to compute the *pth* percentile.

Step 1. Arrange the data in ascending order (smallest value to largest value).

Step 2. Compute an index i

where p is the percentile of interest and n is the number of observations.

Step 3. (a) If *i* is not an integer, round up. The next integer greater than *i* denotes

the position of the *pth* percentile.

(b) If *i* is an integer, the *pth* percentile is the average of the values in positions

*i* and *i+1.*

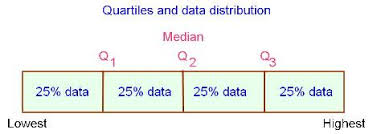
**5. Quartiles -** It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. Figure shows a data distribution divided into four parts. The division points are referred to as the quartiles and are defined as

Q1 : first quartile, or 25th percentile

Q2 : second quartile, or 50th percentile (also the median)

Q3 : third quartile, or 75th percentile.

Quartiles are just specific percentiles; thus, the steps for computing percentiles can be applied directly in the computation of quartiles



**6. Weighted Mean:** A mean computed by giving each observation a weight that reflects its importance is referred to as a weighted mean.

where xi is value of observation *i,* wi is weight for observation *i*

In weighted mean computations, quantities such as grades, pounds, dollars, or volume are frequently used as weights. In any case, when observations vary in importance, the analyst must choose the weight that best reflects the importance of each observation in the determination of the mean. The advantage of weighted mean is that it takes into account the importance of each value to the overall total.

**7. Grouped Data -** In most cases, measures of location are computed by using the individual data values. Sometimes, however, data are available only in a grouped or frequency distribution form. To compute the mean using only the grouped data, we treat the midpoint of each class as being representative of the items in the class. Let Mi denote the midpoint for

class *i* and let fi denote the frequency of class *i*. The weighted mean formula is then

used with the data values denoted as Mi and the weights given by the frequencies fi . In this case, the denominator of equation is the sum of the frequencies, which is the sample size *n*. That is, . Thus, the equation for the sample mean for grouped data is

Where Mi is the midpoint for class *i*, fi is the frequency for class *i* and *n* is the sample size

**8. Geometric mean** - The geometric mean is defined as the [nth root](https://en.wikipedia.org/wiki/Nth_root) of the [product](https://en.wikipedia.org/wiki/Product_(mathematics)) of

n numbers, i.e., for a set of numbers x1, x2, ..., xn, the geometric mean is defined as

Geometric Mean =

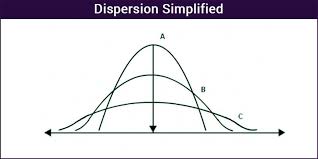
The geometric mean must be used when working with percentages, which are derived from values, while the standard [arithmetic mean](https://www.investopedia.com/terms/a/arithmeticmean.asp) works with the values themselves.

**Uses:**

* Is used to find the average per cent increase in sales, production, population or other economic or business series.
* Most suitable when large weights have to be given to small items and small weights have to be given to large items, situations which usually come across in social or economic fields.

**9. Harmonic Mean:** The harmonic mean can be expressed as the [reciprocal](https://en.wikipedia.org/wiki/Multiplicative_inverse) of the [arithmetic mean](https://en.wikipedia.org/wiki/Arithmetic_mean) of the reciprocals of the given set of observations. Typically, it is appropriate for situations when the average of [rates](https://en.wikipedia.org/wiki/Rate_(mathematics)) is desired. The harmonic mean H of the positive [real numbers](https://en.wikipedia.org/wiki/Real_number) x1, x2, …….. , xn is defined to be

{\displaystyle H={\frac {n}{{\frac {1}{x\_{1}}}+{\frac {1}{x\_{2}}}+\cdots +{\frac {1}{x\_{n}}}}}={\frac {n}{\sum \limits \_{i=1}^{n}{\frac {1}{x\_{i}}}}}=\left({\frac {\sum \limits \_{i=1}^{n}x\_{i}^{-1}}{n}}\right)^{-1}.}

**Measures of dispersion or variability** –

In the figure, the mean of all the curves is the same, but curve A has less spread (or variability) than curve B, and curve B has less variability than curve C. If we measure only the mean of these distributions, we will miss out an important difference among the three curves. Likewise for any data, the mean, median and the mode tell us only part of what we need to know about the characteristics of the data. To increase our understanding of the pattern of the data, we must also measure its dispersion – its spread, or variability.

* It gives us additional information that enables us to judge the reliability of our measure of central tendency. If data are widely dispersed, such as those in curve C, the central location is less representative of the data as a whole than it would be for data more closely centered around the mean, as in curve A
* There are problems peculiar to widely dispersed data, we must be able to recognize that data are widely dispersed before we can tackle those problems.
* We may wish to compare the dispersion of various samples. If a wide spread of values away from the center is undesirable or presents an unacceptable risk, we need to be able to recognize and avoid choosing the distribution with the greatest dispersion.

Some common measures of dispersion or variability are range, inter-quartile range, variance, standard deviation and co-efficient of variation.

**1. Range-** The simplest measure of variability is the range.

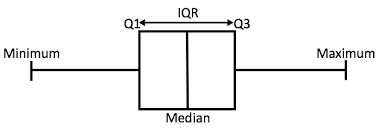
Range = Largest value – Smallest value

Advantage:

* Easy to calculate

Disadvantage:

* It considers only highest and lowest value and fails to take account of other observations. Heavily influenced by extreme values
* Changes drastically from one sample to another belonging to a same population
* Open ended distributions have no range

**2. Inter Quartile Range-** A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR). This measure of variability is the difference between the third quartile, Q3, and the first quartile, Q1. In other words, the interquartile range is the range for the middle 50% of the data. It is how far from the median we must go on either side before we can include one-half the values of the data set. 

**3. Variance -** The variance is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation (*xi*) and the mean. The difference between each *xi* and the mean (for a sample, μ for a population) is called a deviation about the mean.

For a sample, a deviation about the mean is written (); for a population, it is written

(*xi* - μ ). In the computation of the variance, the deviations about the mean are squared.

If the data are for a population, the average of the squared deviations is called the population variance. The population variance is denoted by the Greek symbol σ2. For a population of N observations and with μ denoting the population mean, the definition of the population variance is as follows.

In most statistical applications, the data being analyzed are for a sample. When we compute

a sample variance, we are often interested in using it to estimate the population variance

σ2. If the sum of the squared deviations about the sample mean is divided by n - 1, and

not n, the resulting sample variance provides an unbiased estimate of the population variance. For this reason, the sample variance, denoted by s2, is defined as follows.

Because the values being summed in the variance calculation are squared, the units associated with the sample variance are also squared.

The variance is useful in comparing the variability of two or more variables. In a comparison of the variables, the one with the largest variance shows the most variability.

**4. Standard Deviation -** The standard deviation is defined to be the positive square root of the variance. Following the notation, we adopted for a sample variance and a population variance, we use *s* to denote the sample standard deviation and σ to denote the population standard deviation. The standard deviation is derived from the variance.

and

The standard deviation is easier to interpret than the variance because the

standard deviation is measured in the same units as the data.

**Variance for grouped data** - With grouped data, we treat the class midpoint, Mi, as being representative of the xi values in the corresponding class. Just as we did with the sample

mean calculations for grouped data, we weight each value by the frequency of the class, fi.

The term n - 1 rather than n appears in the denominator in order to make the sample variance the estimate of the population variance. Thus, the following formula is

used to obtain the sample variance for grouped data.

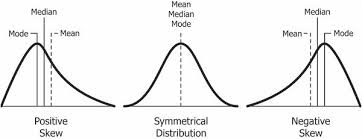
**5. Coefficient of variation** - The coefficient of variation is a relative measure of variability; it measures the standard deviation relative to the mean. It relates the standard deviation and the mean by expressing the standard deviation as a percentage of the mean. In general, the coefficient of variation is a useful statistic for comparing the variability of variables that have

different standard deviations and different means It gives us the magnitude of the deviation relative to the magnitude of the mean.

Eg, if we have a standard deviation of 10 and mean of 5, the values vary by an amount twice as large as the mean itself. Hence, we get a large coefficient of variation. If we have standard deviation of 10 and mean as 5,000 the variation relative to the mean is insignificant. Hence, we get a small coefficient of variation.

**Measures based on shape of distribution: Skewness and Kurtosis**

**1. Skewness:** An important numerical measure of the shape of a distribution is called skewness.

****

**a. Symmetric distribution or No Skewed distribution -** If the data are symmetric, the skewness is zero.

* The values of mean, median and mode coincide.
* Data when plotted on a graph give the normal bell-shaped form
* Sum of the positive deviations from the median is equal to the sum of the negative deviations.
* Quartiles are equidistant from the median.
* Frequencies are equally distributed at points of equal deviations from the mode.

**b. Positively skewed distribution(Right Skewed) -** For data skewed to the right, the skewness is positive. The frequencies are spread out over a greater range of values on the high-value end of the curve (the right-hand side) than they are on the low-value end. Mode is at the highest point of the distribution. The mean > median > mode.

Eg. If the distribution of the household incomes of a region is studied, most of the citizens fall in the group which is the lower side. However, a couple of individuals may have a very high income, in millions. This makes the tail of extreme values (high income) extend longer towards the positive, or right side.

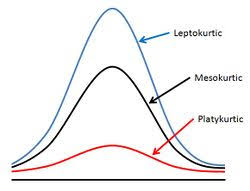
Eg. If an exam has a high difficulty level, then most of the students will have a poor-to-average performance in it. This bulk of students will form the maximum part of the distribution, towards the left side of the curve. The highest marks in the test will be obtained only by a couple of students, which forms the right tail of extreme values. The students with very high marks will shift the mean towards the right, making it a positively skewed distribution. In other words, there will be a higher frequency of low scores and a lower frequency of high scores.

**c. Negatively skewed distribution(Left Skewed) -** For data skewed to the left, the skewness is negative. The value of mode is maximum and that of mean least, the median lies in between the two. The mode> median> mean. The excess tail is on the left-hand side. This is because the mean is pulled down below the median by extremely low values.

Eg. If an exam is easy, then most of the students will perform well in it. This maximum bulk of students will take up the right side of the curve. A few students get very low marks in the exam. These low marks extends the tail in the negative or left direction from the distribution, making it a negatively skewed distribution. Here, there is a high frequency of high scores and low frequency of less scores.

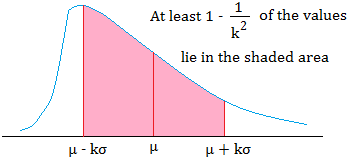
Eg. When comparison of human lifespans is done, most people live beyond their middle age, or even older. Thus, the maximum frequency is of such people, which takes up the right side of the distribution. Some individuals lose their lives at younger age. These individuals take up the lowest values, i.e., towards the left side of the distribution, making the tail longer.

Given the mean and median of a unimodal distribution, we can determine whether it is skewed to the right or left. When mean> median, it is skewed to the right; when median> mean, it is skewed to the left. It may be noted that the median is always in the middle between mean and mode. The median provides the preferred measure of location when the data are highly skewed.

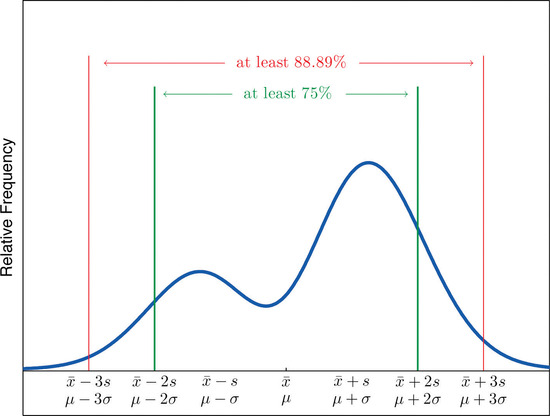
**2. Kurtosis:** While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution. Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are mesokurtic, leptokurtic and platykurtic. Mesokurtic curve is neither too much flattened nor too much peaked. In fact, this is the frequency curve of a normal distribution. Leptokurtic curve is a more peaked than the normal curve. In contrast, platykurtic is a relatively flat curve. The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate.

**Relative Location of values within the data set (z-Scores) -** In addition to measures of location, variability, and shape, we are also interested in the relative location of values within a data set. Measures of relative location help us determine how far a particular value is from the mean.By using both the mean and standard deviation, we can determine the relative locationof any observation. Suppose we have a sample of n observations, with the values denoted by x1 , x2 , . . . , xn . In addition, assume that the sample mean , and the sample standarddeviation *s,* are already computed. Associated with each value, xi , is another value calledits z-score. The z -score is often called the standardized value.

The z-score, zi, can be interpreted as the number of standard deviations xi is from the mean. A z-score greater than zero occurs for observations with a value greater than the mean, and a z-score less than zero occurs for observations with a value less than the mean. A z-score of zero indicates that the value of the observation is equal to the mean. The z-score for any observation can be interpreted as a measure of the relative location of the observation in a data set. Thus, observations in two different data sets with the same z-score can be said to have the same relative location in terms of being the same number of standard deviations from the mean.

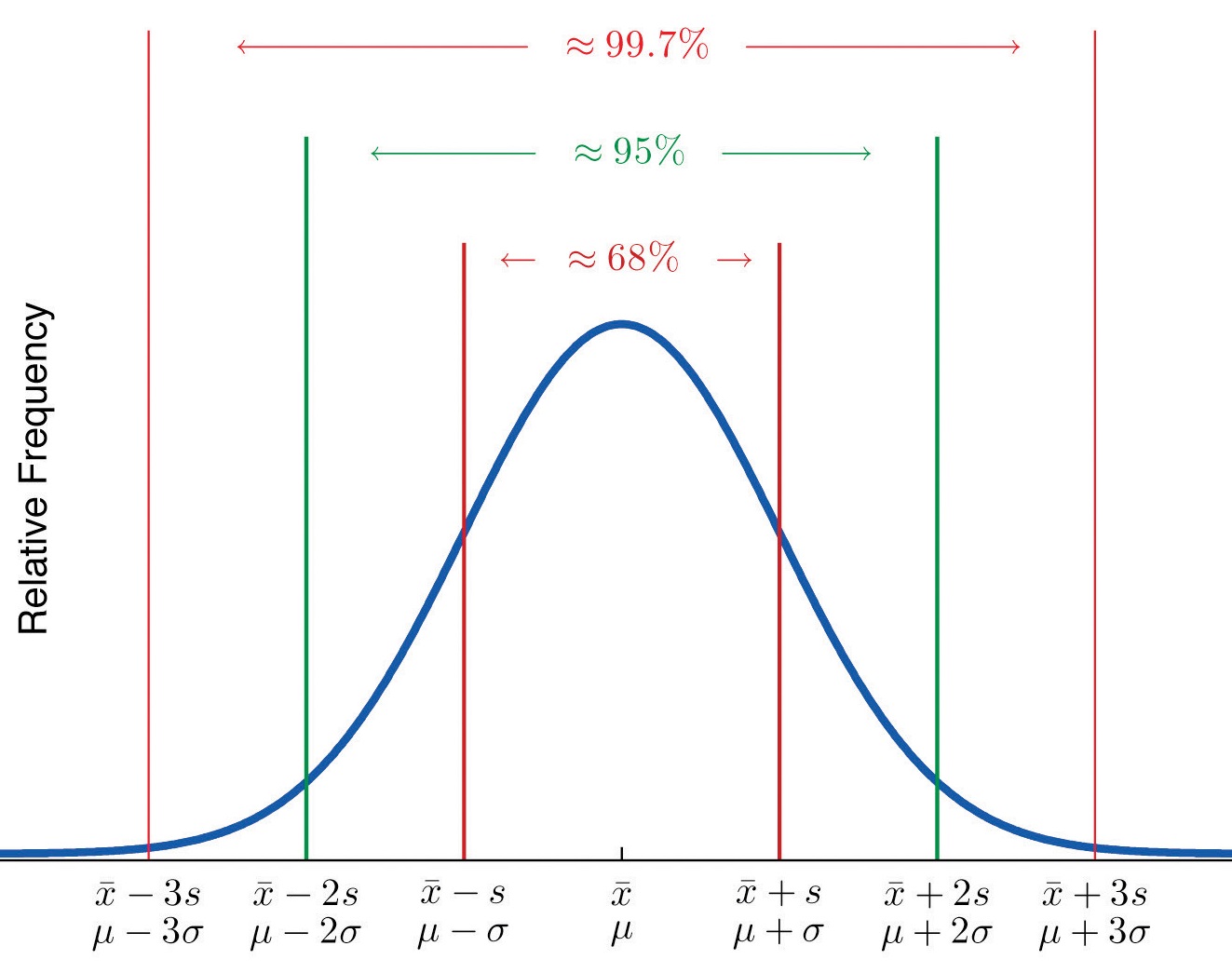
**Chebyshev’s Theorem -** At least of the data values must be within z standard deviations of the mean,where z is any value greater than 1.

Some of the implications of this theorem, with z = 2, 3, and 4 standard deviations, follow:

* ****At least 0.75, or 75%, of the data values must be within z = 2 standard deviations of the mean.
* At least 0.89, or 89%, of the data values must be within z = 3 standard deviations of the mean.
* At least 0.94, or 94%, of the data values must be within z = 4 standard deviations of the mean.

Chebyshev’s theorem enables us to make statements about the proportion of data values

that must be within a specified number of standard deviations of the mean. It applies to any data set regardless ofthe shape of the distribution of the data.

**Empirical Rule -** For data having a bell-shaped distribution:

* Approximately 68% of the data values will be within z = 1 standard deviation of the mean.
* Approximately 95% of the data values will be within z = 2 standard deviations of the mean.
* Almost all of the data values will be within z = 3 standard deviations of the mean.

In many practical applications, however, data sets exhibit a symmetric mound shaped or bell-shaped distribution. When the data are believed to approximate this distribution, the empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

**Outliers -** Sometimes a data set will have one or more observations with unusually large or unusuallysmall values. These extreme values are called outliers.

Experienced statisticians take stepsto identify outliers and then review each one carefully. An outlier may be a data value thathas been incorrectly recorded. If so, it can be corrected before further analysis. An outliermay also be from an observation that was incorrectly included in the data set; if so, it canbe removed. Finally, an outlier may be an unusual data value that has been recorded correctlyand belongs in the data set. In such cases it should remain.Standardized values (z-scores) can be used to identify outliers. The empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within three standard deviations of the mean.

Hence, any data value with a z-score or z-score ≥ 3 is an outlier. Such data values can then be reviewed for accuracy and to determine whether they belong in the data set.

**Five-number summary** - The following five numbers are used to summarize the data:

1. Smallest value

2. First quartile (Q1)

3. Median (Q2)

4. Third quartile (Q3)

5. Largest value

**Measures of association between two variables:**

**1. Covariance -** A measure of linear association between two variables. Positive values indicate

a positive relationship; negative values indicate a negative relationship.

For a sample of size n with the observations (x1, y1), (x2, y2), and so on, the sample covariance is defined as follows:

The formula for computing the covariance of a population (σxy) of size N is

μx is the population mean of the variable x and μy is the population mean of the variable y. One problem with using covariance as a measure of the strength of the linear relationship is that the value of the covariance depends on the units of measurement for x and y. For eg, suppose we are interested in the relationship between height x and weight y for individuals. The strength of the relationship should be the same whether we measure height in feet or inches. Measuring the height in inches, gives us much larger numerical values for than when we measure height in feet. With height measured in inches, we obtain a larger value for the numerator and hence a larger covariance, when in fact the relationship does not change.

**2. Karl Pearson Correlation coefficient** - A measure of linear association between two variables that takeson values between +1 and -1. Values near +1 indicate a strong positive linear relationship; values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.

For sample data, the Pearson product moment correlation coefficient(rxy) is defined as:

where

sxy = sample covariance

sx = sample standard deviation of x

sy = sample standard deviation of y

The correlation coefficient for a population, denoted by (ρxy )is:

where

xy = population covariance

x = population standard deviation of x

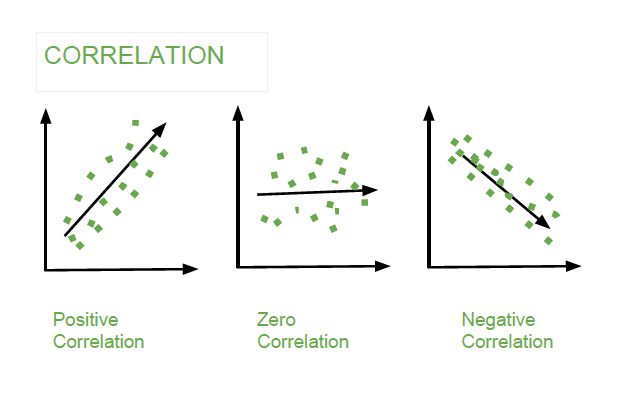
y = population standard deviation of y

The sample correlation coefficient rxy provides an estimate of the population correlation

coefficient .

A high correlation between two variables does not mean that changes in one variable will cause changes in the other variable. For example, we may find that the quality rating and the typical meal price of restaurants are positively correlated. However, simply increasing the meal price at a restaurant will not cause the quality rating to increase.

**Graphical Interpretation of covariance** **-** If the value of sxy is positive, as the value of x increases, the value of also y increases and as the value of x decreases, the value of also y decreases. Hence, it indicates a positive linear association between x and y. If the value of sxy is negative, as the value of x increases, the value of y decreases and as the value of x decreases, the value of y increases. Hence, it indicates a negative linear association between x and y. Finally, if the points are evenly distributed across the graph the value of sxy will be close to zero, indicating no linear association between x and y.



**Graphical Interpretation of correlation -** A sample correlation coefficient of +1 corresponds to a perfect positive linear relationship between x and y. A sample correlation coefficient

of -1 corresponds to a perfect negative linear relationship between x and y .

Let us now suppose that a certain data set indicates a positive linear relationship between

x and y but that the relationship is not perfect. The value of rxy will be less than 1, indicating that the points in the scatter diagram are not all on a straight line. As the points deviate more and more from a perfect positive linear relationship, the value of rxy becomes smaller and smaller. A value of rxy equal to zero indicates no linear relationship between x and y, and values of rxy near zero indicate a weak linear relationship.