Another important value defined in probability distribution is the mean or expected value, which is computed using the following equation (9)for random variable X:

$$\mu = \sum_{i=0}^{n} x_i \times p_i \quad (9)$$

Thus, the mean or expected number of heads in three trials would be:

$$\mu = x_0 \times p_0 + x_1 \times p_1 + x_2 \times p_2 + x_3 \times p_3$$
$$\mu = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5$$

Therefore, in a trail of 3 tosses of coins, the mean number of heads is 1.5.

2.3.1 Binomial Distribution

Binomial distribution is a discrete distribution. It shows the probability distribution of a discrete random variable. The Binomial distribution involves an experiment involving Bernoulli trails, which has the following characteristics:

- A number of trails are conducted, say *n*.
- There can be only two possible outcomes of a trail – Success(say *s*) or Failure (say *f*).
- Each trail is independent of all the other trails.
- The probability of the outcome Success (*s*), as well as failure (*f*), is same in each and every independent trial.

For example, in the experiment of tossing three coins, the outcome success is getting a head in a trial. One possible outcome for this experiment is THT, which is one of outcome of the sample space shown in Figure 2.

You may please note that in case of *n*=3, the for the random variable X, which represents the number of heads, the success is getting a Heads, while failure is getting a Tails. Thus, THT is actually Failure, Success, Failure. The probability for such cases, thus, can be computed as shown earlier. In general, in Binomial distribution, the probability of *r* successes is represented as:

$$P(X = r) \; or p_r = {}^{n}C_r \times s^{\,r} \times f^{n-r} \quad (10)$$

Where *s* is the probability of success and *f* is the probability of failure in each trail. The value of ${}^{n}C_r$ is computed using the combination formula:

$${}^{n}C_r = \frac{n!}{r!(n-r)!} \quad (11)$$

For the case of three tosses of the coins, where X is represented as number of heads in the three tosses of coins *n* = 3 and both *s* and *f* are 1/2, the probability as per Binomial Distribution would be:

$$P(X = 0) \; or \; p_0 = {}^{3}C_0 \times s^{\,0} \times f^{3-0} = \frac{3!}{0!\,(3-0)!} \times \left(\frac{1}{2}\right)^{0} \times \left(\frac{1}{2}\right)^{3} = \frac{1}{8}$$

$$P(X = 1) \; or \; p_1 = {}^{3}C_1 \times s^{\,1} \times f^{3-1} = \frac{3!}{1!\,(3-1)!} \times \left(\frac{1}{2}\right)^{1} \times \left(\frac{1}{2}\right)^{2} = \frac{3}{8}$$

$$P(X = 2) \; or \; p_2 = {}^{3}C_2 \times s^{\,2} \times f^{3-2} = \frac{3!}{2!\,(3-2)!} \times \left(\frac{1}{2}\right)^{2} \times \left(\frac{1}{2}\right)^{1} = \frac{3}{8}$$

$$P(X = 3) \; or \; p_3 = {}^{3}C_3 \times s^{\,3} \times f^{3-3} = \frac{3!}{3!\,(3-3)!} \times \left(\frac{1}{2}\right)^{3} \times \left(\frac{1}{2}\right)^{0} = \frac{1}{8}$$

Which is same as Figure 2 and Figure 3.

Finally, the mean and standard deviation of Binomial distribution for *n* trails, each having a probability of success as *s*, can be defined using the following formulas:

$$\mu = n \times s \quad (12a)$$

$$\sigma = \sqrt{n \times s \times (1-s)} \quad (12b)$$

Therefore, for the variable X which represents number of heads in three tosses of coin, the mean and standard deviation are:

$$\mu = n \times s = 3 \times \frac{1}{2} = 1.5$$

$$\sigma = \sqrt{n \times s \times (1-s)} = \sqrt{3 \times \frac{1}{2} \times (1-\frac{1}{2})} = \frac{\sqrt{3}}{2}$$

Distribution of a discrete random variable, thus, is able to compute the probability of occurrence of specific number successes, as well as the mean or expected value of a random probability experiment.

2.3.2 Probability Distribution of Continuous Random Variable

A continuous variable is measured using scale or interval measures. For example, height of the students of a class can be measured using an interval measure. You can study the probability distribution of a continuous random variable also, however, it is quite different from the discrete variable distribution. Figure 5 shows a sample histogram of the height of 100 students of a class. You may please notice it is typically a grouped frequency distribution.

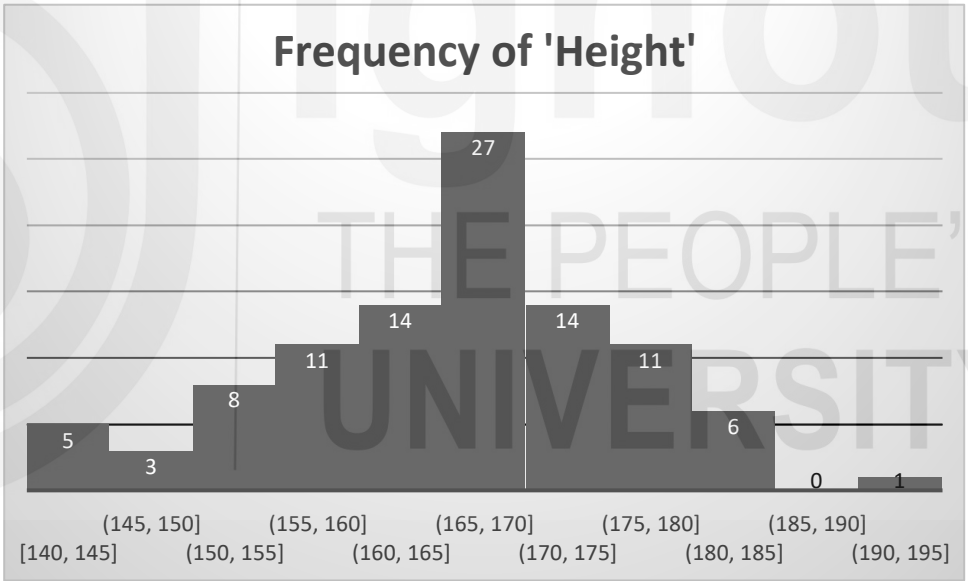

Figure 5: Histogram of Height of 100 students of a Class

The mean of the height was 166 and the standard distribution was about 10.The probability for a student height is in between 165 to 170 interval is 0.27.

In general, for large data set continuous random variable distribution is represented as a smooth curve, which has the following characteristics:

- The probability in each interval would be between 0 and 1. To compute the probability in an interval you need to compute the area of the curve between the starting and end points of that interval.
- The total area of the curve would be 1.

2.3.3 The Normal Distribution

An interesting frequency distribution of continuous random variable is the Normal Distribution, which was first demonstrated by a German Scientist C.F.

Gauss. Therefore, it is sometime also called the Gaussian distribution. The Normal distribution has the following properties:

- The normal distribution can occur in many real life situations, such as height distribution of people, marks of students, intelligence quotient of people etc.
- The curve looks like a bell shaped curve.
- The curve is symmetric about the mean value (μ).Therefore, about half of the probability distribution curve would lie towards the left of the mean and other half would lie towards the right of the mean.
- If the standard deviation of the curve is σ, then about 68% of the data values would be in the range (μ-σ) to (μ+σ) (Refer to Figure 6)
- About 95% of the data values would be in the range (μ-2σ) to (μ+2σ) (Refer to Figure 6)
- About 99.7% of the data values would be in the range (μ-3σ) to (μ+3σ) (Refer to Figure 6).
- Skewness and Kurtosis of normal distribution is closer to zero.
- The probability density of standard normal distribution is represented using a mathematical equation using parameters μ and σ. You may refer to the equation in the further readings.

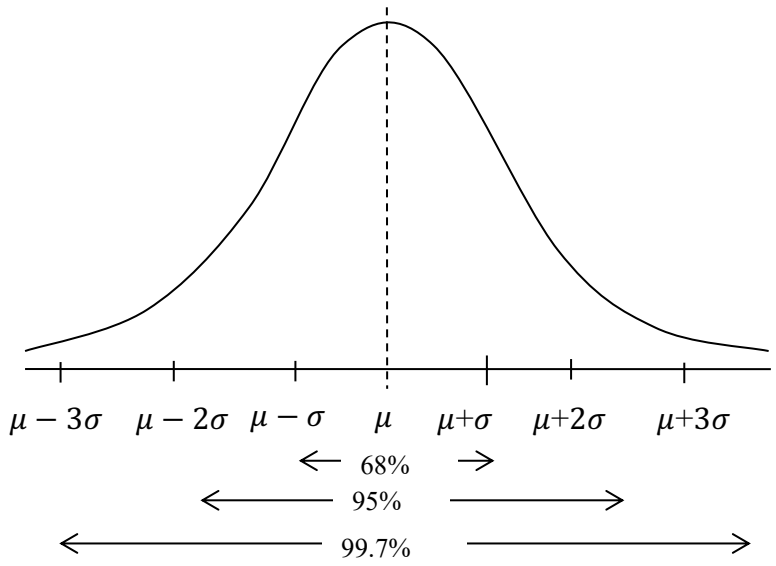

Figure 6: Normal Distribution of Data

Computing probability using Normal Distribution:

The Normal distribution can be used to compute the *z*-score, which computes the distance of a value *x* from its mean in terms of its standard deviation.

For a given continuous random variable *X* and its value *x*; and normal probability distribution with parameters μ and σ; the *z*-score would be computed as:

$$z = \frac{(x-\mu)}{\sigma} \quad (13)$$

You can find the cumulative probabilities at a particular *z*-value using Normal distribution, for example, the shaded portion of the Figure 7 shows the cumulative probabilities at *z*= 1.3, the probability of the shaded portion at this point is 0.9032

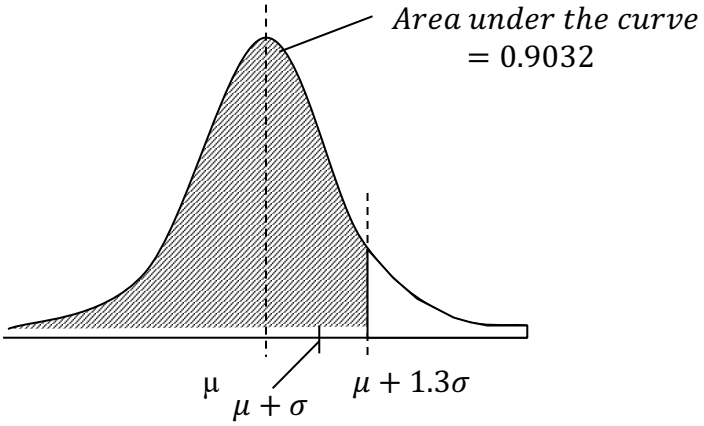

Figure 7: Computing Probability using Normal Distribution

Standard Normal Distribution is a standardized form of normal distribution, which allows comparison of various different normal curves. A standard normal curve would have the value of mean (μ) as zero and standard deviation (σ) as 1. The *z*-score for standard normal distribution would be:

$$z = \frac{(x-0)}{1} = x$$

Therefore, for standard normal distribution the *z*-score is same as value of *x*. This means that $z = \pm 2$contains the 95% area under the standard normal curve.

In addition to Normal distribution a large number of probability distributions have been studied. Some of these distributions are – Poisson distribution, Uniform Distribution, Chi-square distribution etc. Each of these distribution is represented by a characteristics equation involving a set of parameters. A detailed discussion on these distributions is beyond the scope of this Unit. You may refer to Further Reading for more details on these distributions.

2.4 SAMPLING DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

With the basic introduction, as above, next we discuss one of the important aspect of sample and population called sampling distribution. A typical statistical experiment may be based on a specific sample of data that may be collected by the researcher. Such data is termed as the primary data. The question is – Does the statistical results obtained by you using the primary data can be applied to the population? If yes, what may be the accuracy of such a collection? To answer this question, you must study the sampling distribution. Sampling distribution is also a probability distribution, however, this distribution shows the probability of choosing a specific sample from the population. In other words, a sampling distribution is the probability distribution of means of the random samples of the population. The probability in this distribution defines the likelihood of the occurrence of the specific mean of the sample collected by the researcher. Sampling distribution determines whether the statistics of the sample falls closer to population parameters or not. The following example explains the concept of sampling distribution in the context of a categorical variable.

Example 5: Consider a small population of just 5 person, who vote for a question "Data Science be made the Core Course in Computer Science? (Yes/No)". The following table shows the population:

P1	P2	P3	P4	P5	Population Parameter (proportion) (p)
Yes	Yes	No	No	No	0.4

Figure 8: A hypothetical population

Suppose, you take a sample size (n) = 3, and collects random sample. The following are the possible set of random samples:

Sample	Sample Proportion ($\hat{p}$)
P1, P2, P3	0.67
P1, P2, P4	0.67
P1, P2, P5	0.67
P1, P3, P4	0.33
P1, P3, P5	0.33
P1, P4, P5	0.33
P2, P3, P4	0.33
P2, P3, P5	0.33
P2, P4, P5	0.33
P3, P4, P5	0.00

Frequency of all the sample proportions is:

$\hat{p}$	Frequency
0	1
0.33	6
0.67	3

Figure 9: Sampling proportions

The mean of all these sample proportions = $(0\times1 + 0.33\times6+0.67\times3)/10$
= 0.4 (ignoring round off errors)

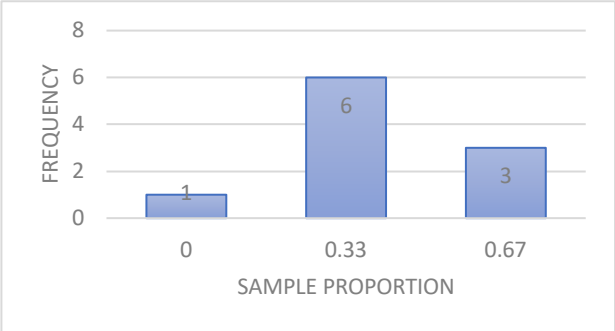

Figure 10: The Sampling Proportion Distribution

Please notice the nature of the sampling proportions distribution, it looks closer to Normal distribution curve. In fact, you can find that out by creating an example with 100 data points and sample size 30.

Given a sample size *n* and parameter proportion *p* of a particular category, then the sampling distribution for the given sample size would fulfil the following:

$$mean\ proportion = p \quad (14a)$$

$$Standard\ Deviation = \sqrt{\frac{p \times (1-p)}{n}} \quad (14b)$$

Let us extend the sampling distribution to interval variables. Following example explains different aspects sampling distribution:

Example 6: Consider a small population of age of just 5 person. The following table shows the population:

P1	P2	P3	P4	P5	Population mean (μ)
20	25	30	35	40	30

Figure 8: A hypothetical population

Suppose, you take a sample size (n) = 3, and collects random sample. The following are the possible set of random samples:

Sample	Sample Mean ($\bar{x}$)
P1, P2, P3	25
P1, P2, P4	26.67
P1, P2, P5	28.33
P1, P3, P4	28.33
P1, P3, P5	30
P1, P4, P5	31.67
P2, P3, P4	30
P2, P3, P5	31.67
P2, P4, P5	33.33
P3, P4, P5	35

Figure 11: Mean of Samples

The mean of all these sample means = 30, which is same as population mean μ. The histogram of the data is shown in Figure 12.

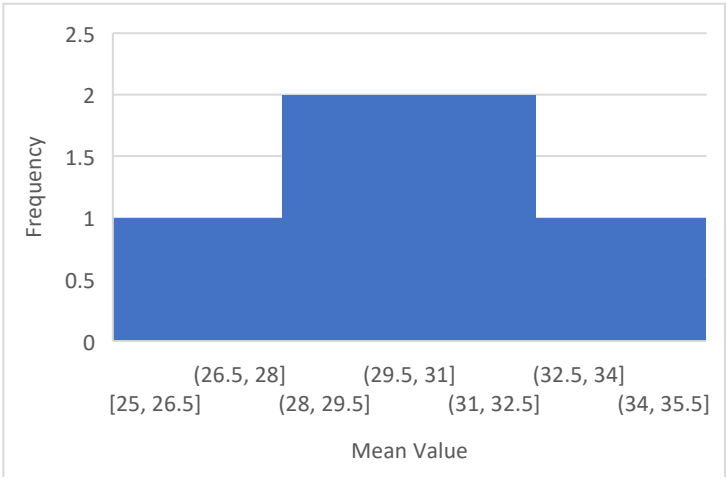

Figure 12: Frequency distribution of sample means

Given a sample size n and population mean μ, then the sampling distribution for the given sample size would fulfil the following:

$$Mean\ of\ sample\ means = \mu \quad (15a)$$

$$Standard\ Deviation\ of\ Sample\ Means = \frac{\sigma}{\sqrt{n}} \quad (15b)$$

Therefore, the *z*-score computation for sampling distribution will be as per the following equation:

Note: You can obtain this equation from equation (13), as this is a distribution of means, therefore, x of equation (13) is $\bar{x}$, and standard deviation of sampling distribution is given by equation (15b).

$$z = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \qquad (15c)$$

Please note that the histogram of the mean of samples is close to normal distribution.

Such exponentiations led to the Central limit Theorem, which proposes the following:
Central Limit Theorem: Assume that a sample of size is drawn from a population that has the mean μ and standard deviation σ. The central limit theorem states that with the increase in *n*, the sampling distribution, i.e. the distribution of mean of the samples, approaches closer to normal distribution.

However, it may be noted that the central limit theorem is applicable only if you have collected independent random samples, where the size of sample is sufficiently large, yet it is less than 10% of the population. Therefore, the Example 5 and Example 6 are not true representations for the theorem, rather are given to illustrate the concept. Further, it may be noted that the central limit theorem does not put any constraint on the distribution of population. Equation 15 is a result of central limit theorem.

Does the Size of sample have an impact on the accuracy of results?
Consider that a population size is100,000 and you have collected a sample of size *n*=100, which is sufficiently large to fulfil the requirements of central limit theorem. Will there be any advantage of taking a higher sample size say *n*=400? The next section addresses this issue in detail.

Check Your Progress 2

1. A fair dice is thrown 3 times, compute the probability distribution of the outcome number of times an even number appears on the dice.

2. What would be the probability of getting different number of heads, if a fail coin is tossed 4 times.

3. What would be the mean and standard deviation for the random variable of Question 2.

4. What is the mean and standard deviation for standard normal distribution?

5. A country has the population of 1 billion, out of which 1% are the students of class 10^{th}. A representation sample of 10000 students of class 10 were asked a question “Is Mathematics difficult or easy?”. Assuming that the population proportion of this question was reported to be 0.36, what would be possible standard deviation of the sampling distribution?

6. Given a quantitative variable, what is the mean and standard deviation of sampling distribution?

2.5 STATISTICAL HYPOTHESIS TESTING

In general, statistical analysis is mainly used in the two situations:

S1. To determine if students of class 12 plays some sport, a sample random survey collected the data from 1000 students. Of these 405 students, stated that they play some sport. Using this information, can you infer that students of class 12 give less importance to sports? Such a decision would require you to estimate the population parameters.

S2. In order to study the effect of sports on the performance of class 12th marks, a study was performed. It performed random sampling and collected the data of 1000 students, which included information of Percentage of marks obtained by the student and hours spent by the student in sports per week during class 12th. This kind of decision can be made through hypothesis testing.

In this section, let us analyse both these situations.

2.5.1 Estimation of Parameters of the Population

One of the simplest ways to estimate the parameter value as a point estimation. Key characteristics of this estimate should be that it should be unbiased, such as mean or median that lies towards the centre of the data; and should have small standard deviation, as far as possible. For example, a point estimate for situation S1 above would be that 40.5% students play some sports. This point estimate, however, may not be precise and may have some margin of error. Therefore, a better estimation would be to define an interval that contains the value of the parameter of the population. This interval, called confidence interval, includes the point estimate along with possible margin of error. The probability that the chosen confidence interval contains the population parameter is normally chosen as 0.95. This probability is called the confidence level. Thus, you can state with 95% confidence that the confidence interval contains a parameter. Is the value of confidence level as 0.95 arbitrary? As you know that sampling distribution for computing proportion is normal if the sample size (n) is large. Therefore, to answer the question asked above, you may study Figure 13 showing the probability distribution of sampling distribution.

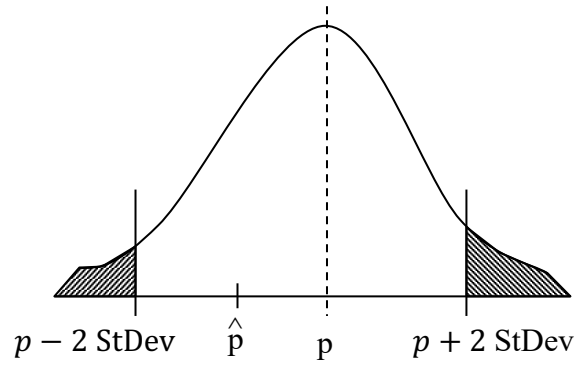

Figure 13: Confidence Level 95% for a confidence interval (non-shaded area).

Since you have selected a confidence level of 95%, you are expecting that proportion of the sample ($\hat{p}$)can be in the interval–(population proportion (p) -

2×(Standard Deviation)) to (population proportion (p) + 2×(Standard Deviation)), as shown in Figure 13. The probability of occurrence of $\hat{p}$ in this interval is 95% (Please refer to Figure 6). Therefore, the confidence level is 95%. In addition, note that you do not know the value of p that is what you are estimating, therefore, you would be computing $\hat{p}$. You may observe in Figure 13, that the value of p will be in the interval ($\hat{p}$ - 2×(Standard Deviation)) to ($\hat{p}$ + 2×(Standard Deviation)). The standard deviation of the sampling distribution can be computed using equation (14b). However, as you are estimating the value of p, therefore, you cannot compute the exact value of standard deviation. Rather, you can compute standard error, which is
computed by estimating the standard deviation using the sample proportion ($\hat{p}$), by using the following formula:

$$Standard\ Error(StErr) = \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}$$

Therefore, the confidence interval is estimated as ($\hat{p}$– 2×*StErr*) to ($\hat{p}$ + 2×*StErr*). In general, for a specific confidence level, you can specify a specific z-score instead of 2. Therefore, the confidence interval, for large n, is: ($\hat{p}$ – z×*StErr*) to ($\hat{p}$ + z×*StErr*)
In practice, you may use confidence level of 90% or 95% and 99%. The z-score used for these confidence levels are 1.65, 1.96 (not 2) and 2.58respectively.
Example 7: Consider the statement S1 of this section and estimate the confidence interval for the given data.
For the sample the probability that class 12th students play some sport is:

$\hat{p}$ = 405/1000=0.405

The sample size (n) = 1000

$$StErr = \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}} = \sqrt{\frac{0.405\times(1-0.405)}{1000}} = 0.016$$

Therefore, the Confidence Interval for the confidence level 95% would be:

(0.405 – 1.96 ×0.016) to (0.405 + 1.96 ×0.016)

0.374 to 0.436

Therefore, with a confidence of 95%, you can state that the students of class 12th, who plays some sport is in the range 37.4% to 43.6%
How can you reduce the size of this interval? You may please observe that *StErr*is inversely dependent on the square root of the sample size. Therefore, you may have to increase the sample size to approximately 4 times to reduce the standard error to approximately half.
Confidence Interval to estimate mean
You can find the confidence interval for estimating mean in a similar manner, as you have done for the case of proportions. However, in this case you need estimate the standard error in the estimated mean usingthe variation of equation 15b, as follows:

$$Standard\ Error\ in\ Sample\ Mean\ = \frac{s}{\sqrt{n}}$$

; where s is the standard deviation of the sample

Example 8: The following table lists the height of a sample of 100 students of class 12 in centimetres. Estimate the average height of students of class 12.

170	164	168	149	157	148	156	164	168	160
149	171	172	159	152	143	171	163	180	158
167	168	156	170	167	148	169	179	149	171
164	159	169	175	172	173	158	160	176	173

159	160	162	169	168	164	165	146	156	170
163	166	150	165	152	166	151	157	163	189
176	185	153	181	163	167	155	151	182	165
189	168	169	180	158	149	164	171	189	192
171	156	163	170	186	187	165	177	175	165
167	185	164	156	143	172	162	161	185	174

Figure 14: Random sample of height of students of class 12 in centimetres

The sample mean and sample standard deviation is computed and is shown below:

Sample Mean ($\bar{x}$) = 166; Standard Deviation of sample (s) = 11

Therefore, the estimated height confidence interval of the mean height of the students of class 12thcan be computed as:

Mean height ($\bar{x}$) = 166

The sample size (n) = 100

$Standard\ Error\ in\ Sample\ Mean\ = \frac{11}{\sqrt{100}}$ =1.1

The Confidence Interval for the confidence level 95% would be:

(166 – 1.96 × 1.1) to (166 + 1.96 × 1.1)

163.8 to 168.2

Thus, with a confidence of 95%, you can state that average height of class 12th students is in between 163.8 to 168.2 centimetres.

You may please note that in example 8, we have used t-distribution for means, as we have used sample's standard deviation rather than population standard deviation. The t-distribution of means is slightly more restrictive than z-distribution. The t-value is computed in the context of sampling distribution by the following equation:

$$t = \frac{(\bar{x}-\mu)}{s/\sqrt{n}} \qquad (16)$$

2.5.2 Significance Testing of Statistical Hypothesis

In this section, we will discuss about how to test the statement S2, given in section 2.5. A number of experimental studies are conducted in statistics, with the objective to infer, if the data support a hypothesis or not. The significance testing may involve the following phases:

1.Testing Pre-condition on Data:

Prior to preforming the test of significance, you should check the pre-conditions on the test. Most of the statistical test require random sampling, large size of data for each possible category being tested and normal distribution of the population.

2. Making the statistical Hypothesis: You make statistical hypothesis after the parameters . of the population. There are two basic hypothesis in statistical testing – the Null Hypothesis and the Alternative Hypothesis.

Null Hypothesis: Null hypothesis either defines a particular value for the parameter or specifies there is no difference or no change in the specified parameters. It is represented as H_0.

Alternative Hypothesis: Alternative hypothesis specifies the values or difference in parameter values. It is represented as either H_1 or H_a. We use the convention H_a.

For example, for the statement S2 of Section 2.5, the two hypothesis would be:

H_0: There is no effect of hours of study on the marks percentage of 12^{th} class.
H_a: The marks of class 12^{th} improves with the hours of study of the student.

Please note that the hypothesis above is one sided, as your assumption is that the marks would increase with hours of study. The second one sided hypothesis may relate to decrease in marks with hours of study. However, most of the cases the hypothesis will be two sided, which just claims that a variable will cause difference in the second. For example, two sided hypothesis for statement S2 would be hours of study of students makes a difference (it may either increase or decrease) the marks of students of class 12^{th}. In general, one sided tests are called one tailed tests and two sided tests are called two tailed tests.

In general, alternative hypothesis relates to the research hypothesis. Please also note that the alternative hypothesis given above is one way hypothesis as it only states the effect in terms of increase of marks. In general, you may have alternative hypothesis which may be two way (increase or decrease; less or more etc.).

3. Perform the desired statistical analysis:
Next, you perform the exploratory analysis and produce a number of charts to explore the nature of the data. This is followed by performing a significance statistical test like chi-square, independent sample t-test, ANOVA, non-parametric tests etc., which is decided on the basis of size of the sample, type and characteristics of data. These tests generate assumes the null hypothesis to be True. A test may generate parameter values based on sample and the probability called p-value, which is an evidence against the null hypothesis. This is shown in Figure 15.

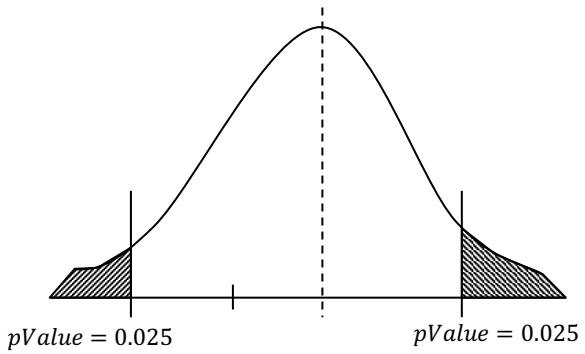

Figure 15: p-value of test statistics

4. Analysing the results:
In this step, you should analyse your results. As stated in Unit 1, you must not just draw your conclusion based on statistics, but support it with analytical reasoning.

Example 9: We demonstrate the problem of finding a relationship between study hours and Marks percentage (S2 of section 2.5), however, by using only sample data of 10 students (it is hypothetical data and just used for the illustration purpose), which is given as follows:

Weekly Study Hours (*wsh*)	96	92	63	76	89	80	56	70	61	81
Marks Percentage (*mp*)	21	19	7	11	16	17	4	9	7	18

In order to find such a relationship, you may like to perform basic exploratory analysis. In this case, let us make a scatter plot between the two variables, taking *wsh* as an independent variable and *mp* as a dependent variable. This scatter plot is shown in Figure 16

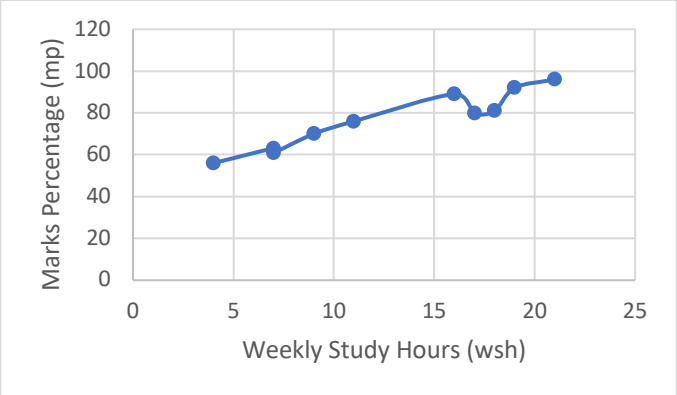

Figure 16: Scatter plot of Weekly Study Hours vs. Marks Percentage.

The scatter plot of Figure 16 suggests that the two variables may be associated. But how to determine the strength of this association? In statistics, you use Correlation, which may be used to determine the strength of linear association between two quantitative variables. This is explained next.

2.5.3 Example using Correlation and Regression

As stated correlation is used to determine the strength of linear association. But how the correlation is measured?
Consider two quantitative variables *x* and *y*, and a set of *n* pairs of values of these variables (for example, the *wsh* and *mp* values as shown in example 9), you can compute a correlation coefficient, denoted by *r* using the following equation:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(\frac{(x-\bar{x})}{sx}\right)\times\left(\frac{(y-\bar{y})}{sy}\right)}{(n-1)} \qquad (16)$$

The following are the characteristics of the correlation coefficient (*r*):

- The value of *r* lies between +1 and -1.
- A positive value of *r* means that value of *y* increases with increase in value of *x* and the value of *y* decreases with decrease in value of *x*.
- A negative value of *r* means that value of *y* increases with decrease in value of *x* and the value of *y* decreases with increase in value of *x*.
- If the value of *r* is closer to +1 or -1, then it indicates that association is a strong linear association.
- Simple scaling one of the variable does not change the correlation.
- Correlation does not specify the dependent and independent variables.
- Please remember a correlation does not mean cause. You have to establish it with reasoning.

The data of Example 9 shows a positive correlation. It can be computed as follows:

Mean of *wsh* = 12.9; Standard Deviation of *wsh* (Sample) = 5.98980616
Mean of *mp* = 76.4; Standard Deviation of *mp* (Sample) = 13.7210301

$$r_{wsh,mp} = \frac{8.61944034}{(10-1)} = 0.95771559$$

Therefore, the data shows strong positive correlation.

You may also use any statistical tool to find the correlation, we used MS-Excel, which gave the following output of correlation:

	Weekly Study Hours (wsh)	*Marks Percentage (mp)*
Weekly Study Hours (wsh)	1	
Marks Percentage (mp)	0.957715593	1

Figure 17: The Correlation coefficient

As the linear correlation between *wsh* and *mp* variables is strong, therefore, you may like to find a line, called linear regression line, that may describe this association. The accuracy of regression line, in general, is better for higher correlation between the variables.

Single Linear Regression:

A single linear regression predicts a response variable or dependent variable (say y) using one explanatory variable or independent variable (say x). The equation of single linear regression can be defined by using the following equation:

$$y_{predicted} = a + bx \quad (17)$$

Here, $y_{predicted}$ is the predicted value of response variable (y), x is the explanatory variable, a is the intercept with respect to y and b is called the slope of the regression line. In general, when you fit a linear regression line to a set of data, there will be certain difference between the $y_{predicted}$ and the observed value of data (say $y_{observed}$). This difference between the observed value and the predicted value, that is ($y_{observed}$ - $y_{predicted}$), is called the residual. One of the most used method of finding the regression line is the method of least square, which minimises the sum of squares of these residuals. The following equations can be used for computing residual:

$$Residual = y_{observed} - y_{predicted} \quad (18)$$

The objective of least square method in regression is to minimise the sum of squares of the residual of all the n observed values. This sum is given in the following equation:

$$SumOfResidualSquares = \sum_{i=1}^{n}\left(y_{observed} - y_{predicted}\right)^2 \quad (19)$$

Another important issue with regression model is to determine the predictive power of the model, which is computed using the square of the correlation (r^2). The value of r^2 can be computed as follows:

- In case, you are not using regression, then you can predict the value of y using the mean. In such a case, the difference in predicted value and observed value would be given by the following equation:

$$ErrorUsingMean = y_{observed} - \bar{y} \quad (20)$$

- The total of sum of square of this error can be computed using the following equation:

$$TotalSumOfSquare = \sum_{i=1}^{n}(y_{observed} - \bar{y})^2 \quad (21)$$

The use of regression line reduces the error in prediction of the value of y. Equation (19) represents this square error. Thus, use of regression results helps in reducing the error. The proportion r^2 is actually the predictive power of the regression and is represented using the following equation:

$$r^2 = \frac{\sum_{i=1}^{n}(y_{observed} - \bar{y})^2 - \sum_{i=1}^{n}\left(y_{observed} - y_{predicted}\right)^2}{\sum_{i=1}^{n}(y_{observed} - \bar{y})^2} \quad (22)$$

As stated earlier, r^2 can also be computed by squaring the value of r.

On performing regressing analysis on the observed data of Example 9, the statistics as shown in Figure 18 is generated.

Regression Statistics	
Multiple R	0.9577
R Square	0.9172
Adjusted R Square	0.9069
Standard Error	4.1872
Observations	10.0000

ANOVA

	df	*SS*	*F*	*Significance F*
Regression	1.0000	1554.1361	88.6407	0.0000
Residual	8.0000	140.2639		
Total	9.0000	1694.4000		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	48.0991	3.2847	14.6435	0.0000
Weekly Study Hours (*wsh)*	2.1939	0.2330	9.4149	0.0000

Figure 18: A Selected Regression output

The regression analysis results, as shown above are discussed below:

- Assumptions for the regression model:
  - Data sample is collected using random sampling.
  - For every value of x, the value of y in the population
    - is normally distributed
    - has same standard deviation
  - The mean value if y in the population follows regression equation (17)
- Various Null hypothesis related to regression are:
  - For the analysis of variance (ANOVA) output in the regression:
    - H_{0A}: All the coefficients of model are zero, therefore, the model cannot predict the value of y.
  - For the *Intercept*:
    - H_{0I}: *Intercept* =0.
  - For the *wsh*:
    - H_{0wsh}: *wsh* =0.
- The *Significance F* in ANOVA is 0, therefore, you can reject the Null hypothesis H_{0A} and determine that the this model can predict the value of y. Please note high F value supports this observation.
- The p-value related to *intercept* and *wsh* are almost 0, therefore, you can reject the Null hypothesis H_{01} and H_{0wsh}.
- The regression line has the equation:
  $mp_{predicted} = 48.0991 + 2.1939 \times wsh$
- You can compute the sum of squares (*SS*) using Equation (19) and Equation (21).
- The degree of freedom in the context of statistics is the number of data items required to compute the desired statistics.

- The term "Multiple R" in *Regression Statistics* defines the correlation between the dependent variable (say y) with the set of independent or explanatory variables in the regression model. Thus, multiple R is similar to correlation coefficient (r), expect that it is used when multiple regression is used. Most of the software express the results in terms of Multiple R, instead of r, to represent the regression output. Similarly, R Square is used in multiple regression, instead of r^2. The proposed model has a large r^2, therefore, can be considered for deployment.

You can go through further readings for more details on all the terms discussed above.

Figure 19 shows the regression line for the data of Example 9. You may please observe that residuals is the vertical difference between the Marks Percentage and Predicted marks percentage. These residuals are shown in Figure 20.

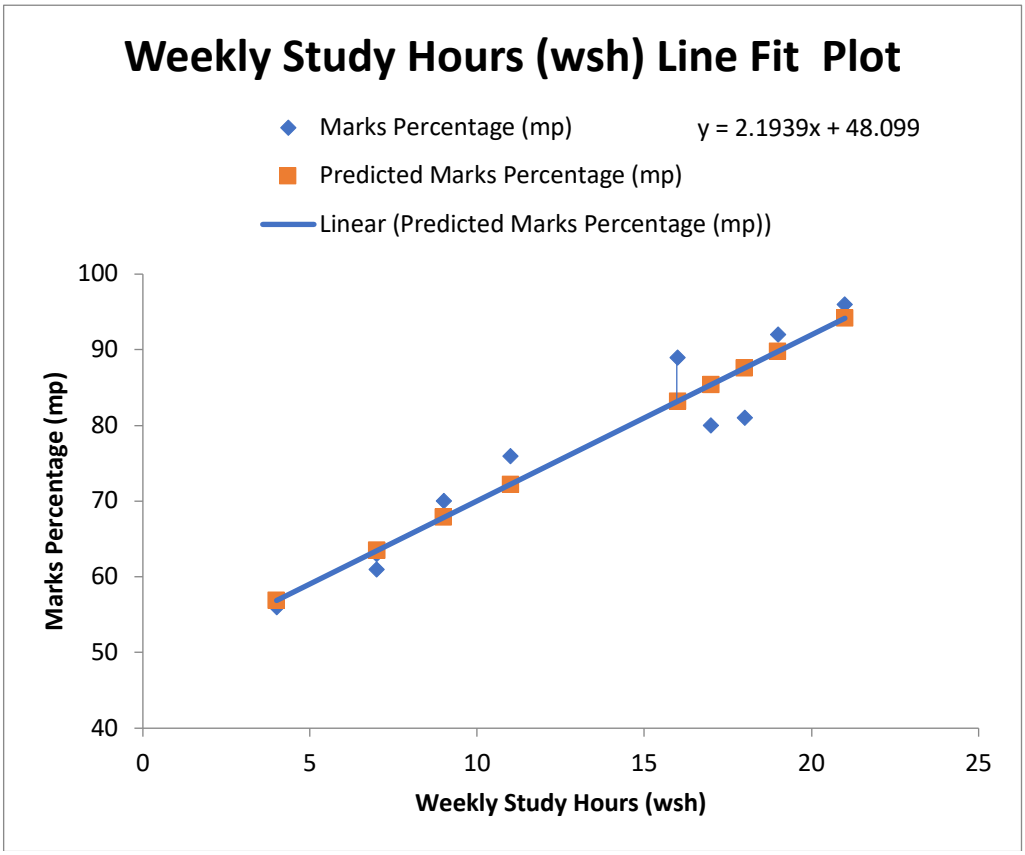

Figure 19: The Regression Line

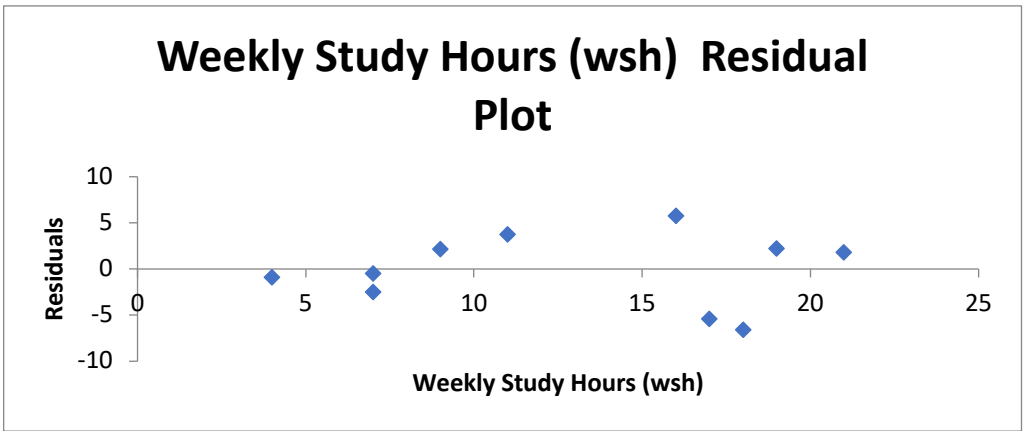

Figure 20: The Residual Plot

2.5.4 Types of Errors in Hypothesis Testing

In the section 2.5.1 and section 2.5.2, we have discussed about testing the Null hypothesis. You either Reject the Null hypothesis and accepts alternative hypothesis based on the computed probability or p-value; or you fail to Reject the Null hypothesis. The decisions in such hypothesis testing would be:

- You reject Null hypothesis for a confidence interval 95% based on the p-value, which lies in the shaded portion, that is p-value < 0.05 for two tailed hypothesis (that is both the shaded portions in Figure 15, each area of probability 0.025). Please note that in case of one tailed test, you would consider only one shaded area of Figure 15, therefore, you would be considering p-value < 0.05 in only one of the two shaded areas.
- You fail to reject the Null hypothesis for confidence interval 95%, when p-value > 0.05.

The two decisions as stated above could be incorrect, as you are considering a confidence interval of 95%. The following Figure shows this situation.

The Actual Scenario	*Final Decision*	
	H_0 is Rejected, that is, you have accepted the Alternative hypothesis	*You fail to reject H_0, as you do not have enough evidence to accept the Alternative hypothesis*
H_0 is True	This is called a TYPE-I error	You have arrived at a correct decision
H_0 is False	You have arrived at a correct decision	This is called a TYPE-II error

For example, assume that a medicine is tested for a disease and this medicine is NOT a cure of the disease. You would make the following hypotheses:

H_0: The medicine has no effect for the disease

H_a: The medicine improves the condition of patient.

However, if the data is such that for a confidence interval of 95% the p-value is computed to be less than 0.05, then you will reject the null hypothesis, which is Type-I error. The chances of Type-I errors for this confidence interval is 5%. This error would mean that the medicine will get approval, even though it has no effect on curing the disease.

However, now assume that a medicine is tested for a disease and this medicine is a cure of the disease. Hypotheses still remains the same, as above. However, if the data is such that for a confidence interval of 95% the p-value is computed to be more than 0.05, then you will not be able to reject the null hypothesis, which is Type-II error. This error would mean that a medicine which can cure the disease will not be accepted.

Check Your Progress 3

1. A random sample of 100 students were collected to find their opinion about whether practical sessions in teaching be increased? About 53 students voted for increasing the practical sessions. What would be the confidence interval of the population proportions of the students who would favour increasing the population percentage. Use confidence levels 90%, 95% and 99%.

2. The Weight of 20 students, in Kilograms, is given in the following table

65	75	55	60	50	59	62	70	61	57
62	71	63	69	55	51	56	67	68	60

Find the estimated weight of the student population.

3. A class of 10 students were given a validated test prior and after completing a training course. The marks of the students in those tests are given as under:

Marks before Training (*mbt*)	56	78	87	76	56	60	59	70	61	71
Marks after training (*mat*)	55	79	88	90	87	75	66	75	66	78

With a significance level of 95% can you say that the training course was useful?

2.6 SUMMARY

This Unit introduces you to the basic probability and statistics related to data science. The unit first introduces the concept of conditional probability, which defines the probability of an event given a specific event has occurred. This is followed by discussion on the Bayes theorem, which is very useful in finding conditional probabilities. Thereafter, the unit explains the concept of discrete and continuous random variables. In addition, the Binomial distribution and normal distribution were also explained. Further, the unit explained the concept of sampling distribution and central limit theorem, which forms the basis of the statistical analysis. The Unit also explain the use of confidence level and intervals for estimating the parameters of the population. Further, the unit explains the process of significance testing by taking an example related to correlation and regression. Finally, the Unit explains the concept of errors in hypothesis testing. You may refer to further readings for more details on these concepts.

2.7 SOLUTION/ANSWERS

☞ Check Your Progress – 1

1. Is $P(Y/X) = P(Y/X)$, No. Please check in Example 3, the probability *P(Red/BagB)* is 7/10, whereas, *P(BagB/Red)*is 7/12.
2. Consider two independent events A and B, first compute *P(A)* and *P(B).* The probability of any one of these events to occur would be computed by equation (2), which is:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
The probability of occurrence of both the events will be computed using the equation (4), which is:
$$P(X \cap Y) = P(X) \times P(Y)$$
3. Let us assume Event X, as "A student is selected from University A". Assuming, any of the University can be selected with equal probability, *P(UniA)* = 1/2.
Let the Event Y, as "A student who has obtained more that 75% marks is selected". This probability $P(StDis) = \frac{1}{2} \times \frac{10}{20} + \frac{1}{2} \times \frac{20}{30} = \frac{7}{12}$
In addition, $P(StDis/UniA) = \frac{10}{20} = \frac{1}{2}$

$$P(UniA/StDis) = \frac{P(StDis/UniA) \times P(UniA)}{P(StDis)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{7}{12}} = \frac{3}{7}$$

Check Your Progress 2

1. As the probability of getting the even number (E) or odd number (O) is equal in each two of dice, the following eight outcomes may be possible:

Outcomes	EEE	EEO	EOE	EOO	OEE	OEO	OOE	OOO
Number of times Even number appears (*X*)	3	2	2	1	2	1	1	0

Therefore, the probability distribution would be:

X	Frequency	Probability *P(X)*
0	1	1/8
1	3	3/8
2	3	3/8
3	1	1/8
Total	8	Sum of all *P(X)* = 1

2. This can be determined by using the Binomial distribution with *X*=0, 1, 2, 3 and 4, as follows (*s* and *f* both are 1/2):

$$P(X = 0)\ or\ p_0 = {}^4C_0 \times s^{\,0} \times f^{4-0} = \frac{4!}{0!\,(4-0)!} \times \left(\frac{1}{2}\right)^0 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$P(X = 1)\ or\ p_1 = {}^4C_1 \times s^{\,1} \times f^{4-1} = \frac{4!}{1!\,(4-1)!} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^3 = \frac{4}{16}$$

$$P(X = 2)\ or\ p_2 = {}^4C_2 \times s^{\,2} \times f^{4-2} = \frac{4!}{2!\,(4-2)!} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^2 = \frac{6}{16}$$

$$P(X = 3)\ or\ p_3 = {}^4C_3 \times s^{\,3} \times f^{4-3} = \frac{4!}{3!\,(4-3)!} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^1 = \frac{4}{16}$$

$$P(X = 4)\ or\ p_4 = {}^4C_4 \times s^{\,4} \times f^{4-4} = \frac{4!}{4!\,(4-4)!} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^0 = \frac{1}{16}$$

3. The number of tosses (n) = 4 and s = ½, therefore,

$$\mu = n \times s = 4 \times \frac{1}{2} = 2$$

$$\sigma = \sqrt{n \times s \times (1-s)} = \sqrt{4 \times \frac{1}{2} \times (1 - \frac{1}{2})} = 1$$

4. Mean = 0 and Standard deviation = 1.

5. Standard deviation of sampling distribution =

$$\sqrt{\frac{p \times (1-p)}{n}} = \sqrt{\frac{0.36 \times (1-0.36)}{10000}} = \frac{0.6 \times 0.8}{100} = 0.0048$$

The large size of sample results in high accuracy of results.

6. $Mean\ of\ sample\ means = \mu$

$Standard\ Deviation\ of\ Sample\ Means = \frac{\sigma}{\sqrt{n}}$

Check Your Progress 3

1. The value of sample proportion $\hat{p}$ = 53/100 = 0.53

Therefore, $StErr = \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} = \sqrt{\frac{0.53 \times (1-0.53)}{100}} = 0.05$

The Confidence interval for 90%:

$(0.53 \pm 1.65 \times 0.05)$, which is 0.4475 to 0.6125

The Confidence interval for 95%:

$(0.53 \pm 1.96 \times 0.05)$, which is 0.432 to 0.628

The Confidence interval for 99%:

$(0.53 \pm 2.58 \times 0.05)$, which is 0.401 to 0.659

2. Sample Mean ($\bar{x}$) = 61.8; Standard Deviation of sample (s) = 6.787

Sample size (n) = 20

$Standard\ Error\ in\ Sample\ Mean = \frac{6.787}{\sqrt{20}} = 1.52$

The Confidence Interval for the confidence level 95% would be:

$(61.8 \pm 1.96 \times 1.52) = 58.8$ to 64.8

3. Analysis: This kind of problem would require to find, if there is significant difference in the mean of the test results before and after the training course. In addition, the data size of the sample is 10 and the same group of person are tested twice, therefore, paired sample t-test may be used to find the difference of the mean. You can follow all the steps for this example of hypothesis testing.

1. Testing Pre-condition on Data:
   - The students who were tested through this training course were randomly selected.
   - The population test scores, in general, are normally distributed.
   - The sample size is small, therefore, a robust test may be used.
2. The Hypothesis

   $H_0: \overline{mbt} = \overline{mat}$

   $H_1: \overline{mbt} < \overline{mat}$
3. The results of the analysis are given below

(Please note H_1 is one sided hypothesis, as you are trying to find if training was useful for the students)

t-Test: Paired Two Sample for Means

	Marks before Training (mbt)	*Marks after training (mat)*
Mean	67.4	75.9
Variance	112.9333333	124.1
Observations	10	10
df	9	
t Stat	-2.832459252	
P(T<=t) one-tail	0.009821702	
t Critical one-tail	1.833112933	

4. Analysis of results: The one tail p-value suggests that you reject the null hypothesis. The difference in the means of the two results is significant enough to determine that the scores of the student have improved after the training.