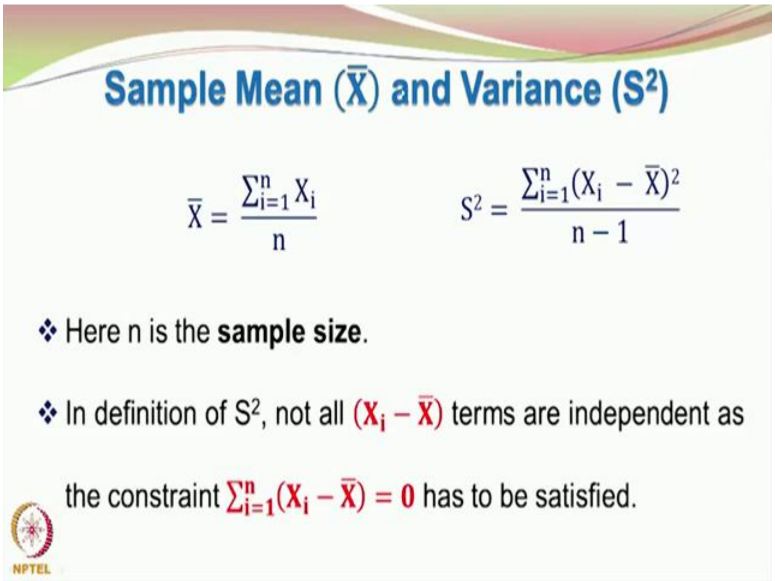
Introduction to Research
Prof. Kannan
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 03
Design of Experiments

Now that we have taken the sample, we can find the sample mean and sample variance.

(Refer Slide Time: 00:21)



Sample Mean ($\bar{X}$) and Variance (S^2)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

❖ Here n is the **sample size**.

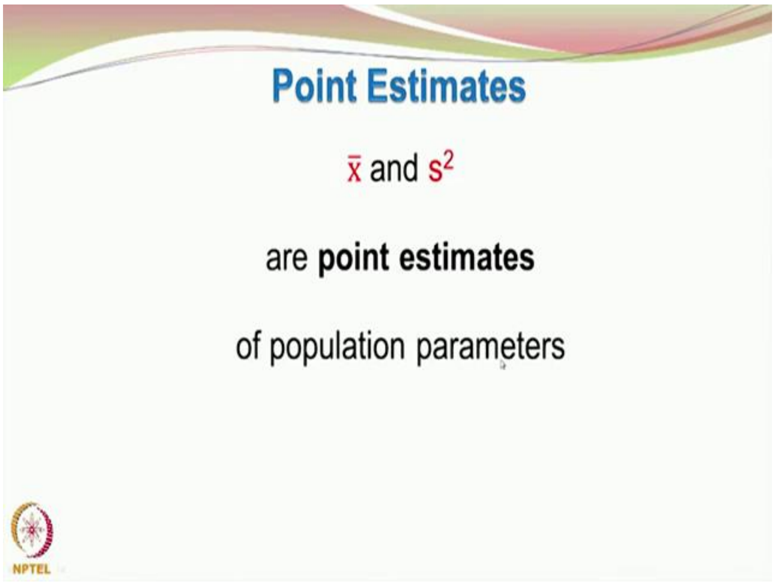
❖ In definition of S^2, not all $(X_i - \bar{X})$ terms are independent as the constraint $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ has to be satisfied.

Sample mean is denoted by x bar and sample variance is denoted by s square. So, x bar defined as sigma i equals 1 to n x i by n, x i is the ith random variable. Similarly, a sample variance s squared is defined as the square of the deviation of each random variable from the sample mean. So, each of the deviations is squared and then summed. We get sigma i is equal to 1 to n x i minus x bar whole squared by n minus one. Here, n is the sample size, and it also denotes the number of degrees of freedom. The degrees of freedom is a very interesting concept and refers to the number of independent entities in the collection, you are looking at the collection.

The collection we are looking at is x i minus x bar, we have n such terms, we have n random variables, but not all the x i minus x bar terms are independent because we know that the sum of the deviations from the mean is equal to zero. So, when you have n minus 1 deviation, the nth deviations should be such that the sum is having a value of zero. So, there are only n minus 1 independent entities. If you have n independent entities of the deviations, the sum may not be equal to zero and the constraint is violated.

Another thing to remember here is x bar and s squared are called as the estimators of the population mean and variance respectively. So, these are estimators. So, the formulae for these estimators are defined as shown in this slide. So, now, once you have actually taken a sample, and found out the values, and then calculated the mean and variance based on the sample values, and then we have what are called as sample estimates of the population mean mu and variance sigma squared. So, point estimates are denoted by small x bar and small s squared and these are point estimates of the population parameters.

(Refer Slide Time: 03:09)



Point Estimates

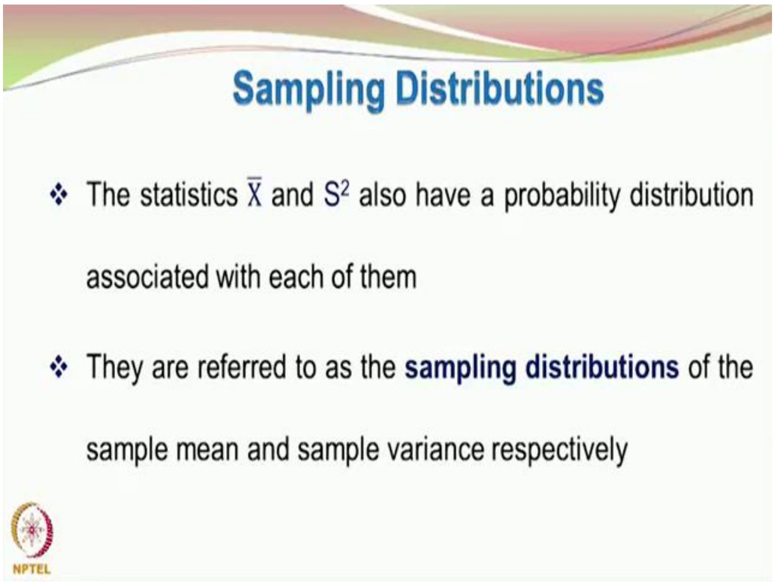
$\bar{x}$ and s^2

are **point estimates**

of population parameters

So, the definition is sample mean involving the random variables before their values are known is given here. And. similarly, for the sample variance it is given as shown here. Sum of the square of the deviations divided by n minus one, where n is the sample size, and these are measures of the population mean and the population variance. So, we can call them as the estimators of the population parameters. Once the sample has been taken and the values determined, we have the sample estimates and these are also point estimates. Why we call them as point estimates is because these values are single values; for example, you have one sample mean based on the sample you have collected and you have one sample variance. So, you are giving a specific value for x bar and s squared.

(Refer Slide Time: 04:15)

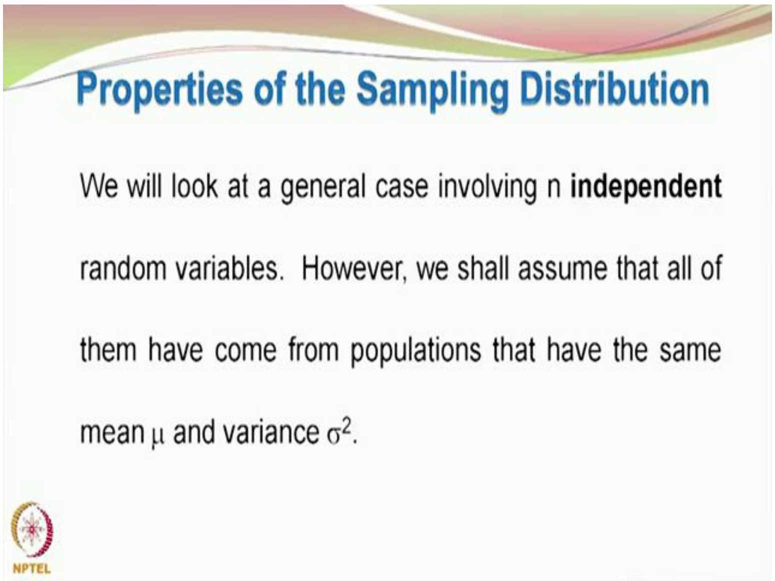


Sampling Distributions

- The statistics $\bar{X}$ and S^2 also have a probability distribution associated with each of them
- They are referred to as the **sampling distributions** of the sample mean and sample variance respectively

Now, we know that the random variable x is having its own probability distribution; that means, it can take a spectrum of values and there is a probability distribution associated with this spectrum of values it can take. When x can take multiple values a mathematical manipulation of x can also take a range of values. If x can have probability distribution x bar and s squared can also have probability distributions associated with them okay. And they are mathematical transformations of x into x bar and the s squared, and so, correspondingly you also have a probability distribution associated with the new random variables x bar and s squared. So, these are referred to as the sampling distributions of the sample mean and sample variance respectively.

(Refer Slide Time: 05:24)

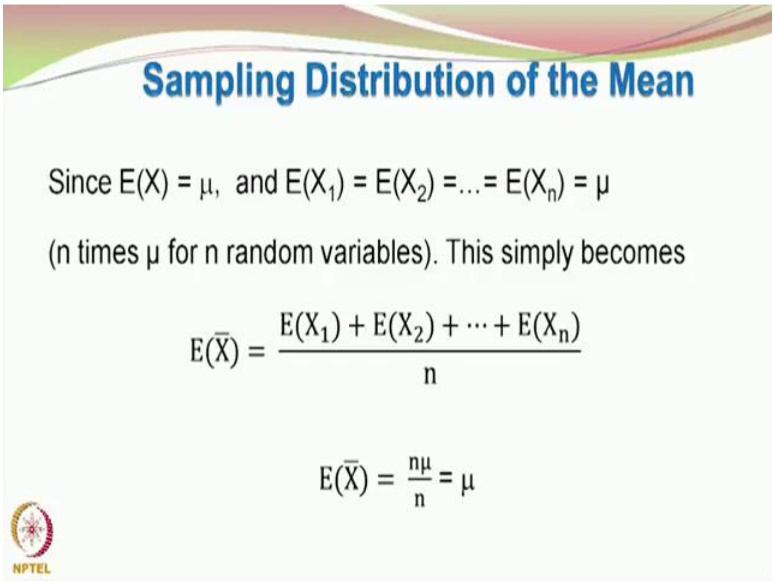


Properties of the Sampling Distribution

We will look at a general case involving n **independent** random variables. However, we shall assume that all of them have come from populations that have the same mean μ and variance σ^2.

So, now, let us look at the properties of the sampling distribution. We will look at the general case involving n independent random variables and we will assume that all of these have come from populations that have the same mean and variance sigma square. So, now, let us look at the sampling distribution of the mean.

(Refer Slide Time: 05:30)



Sampling Distribution of the Mean

Since $E(X) = \mu$, and $E(X_1) = E(X_2) = \ldots = E(X_n) = \mu$

(n times μ for n random variables). This simply becomes

$$E(\bar{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n}$$

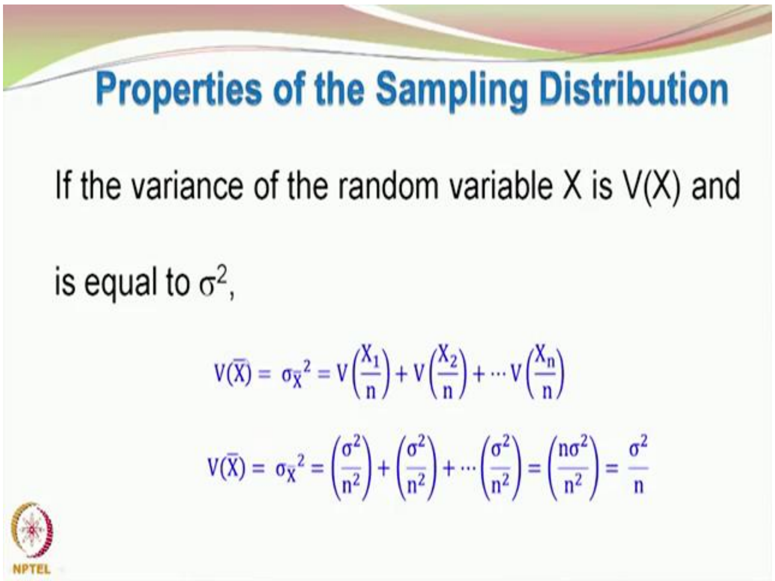
$$E(\bar{X}) = \frac{n\mu}{n} = \mu$$

First of all, what is meant by sampling distribution of the mean? From a population, you can take any number of samples okay. And There is no guarantee that the mean you get from the first sample should be identical to the mean you take from the second sample okay. So, in such a situation you have a distribution of the sample means just as you had a distribution of the random variable x. Now, we know by definition, the excepted value of the random variable x is equal to mu, and since all of these have come from population of the same parameter mu and sigma squared, expected value of x 1 will be equal to excepted value of x 2 so on to excepted value of x n, and they will all be equal to mu.

And if you look at the excepted value of x bar, this is also a random variable, it can be shown as given in the slide that e of x bar is also equal to mu. The random variable x is coming from a probability distribution which is having a mean mu. x bar is also coming from a probability distribution which is having the same mean mu as the parent population okay. You are taking samples from a population and that sampling distribution is also having a mean mu. What sampling distribution are we talking to here? The sampling distribution of the means, so the mean of the means is mu; slightly confusing, but if you think about it, it is pretty simple after all.

Now, let us look at the variance of x bar. The variance of the sampling distribution of the means, so the sampling distribution of the means is also having a spread; so when you have a spread, then you have a variance associated with the spread. So, as given in the slide, you have many random samples that may be drawn from a population, and each of them may have a different mean. So, there will be a distribution of the sample means and the variance of this distribution is v of x bar.

(Refer Slide Time: 07:50)



Properties of the Sampling Distribution

If the variance of the random variable X is V(X) and is equal to σ^2,

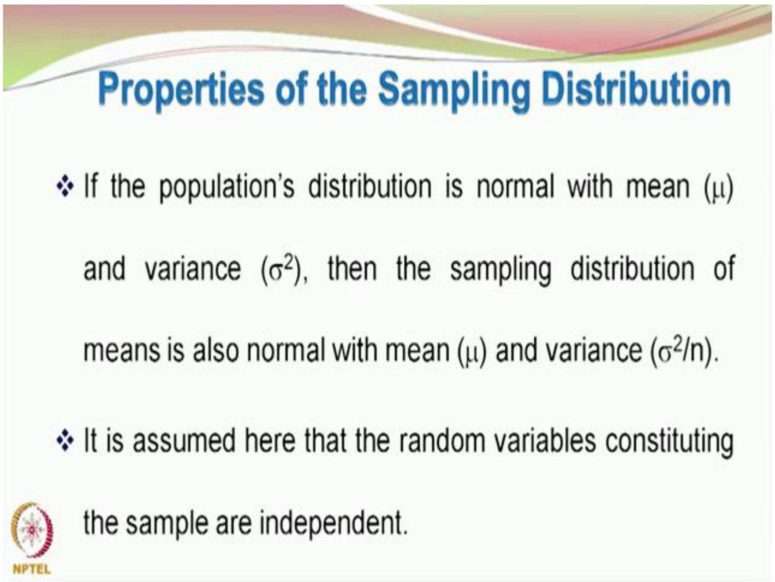
$$V(\bar{X}) = \sigma_{\bar{X}}{}^2 = V\left(\frac{X_1}{n}\right) + V\left(\frac{X_2}{n}\right) + \cdots V\left(\frac{X_n}{n}\right)$$

$$V(\bar{X}) = \sigma_{\bar{X}}{}^2 = \left(\frac{\sigma^2}{n^2}\right) + \left(\frac{\sigma^2}{n^2}\right) + \cdots \left(\frac{\sigma^2}{n^2}\right) = \left(\frac{n\sigma^2}{n^2}\right) = \frac{\sigma^2}{n}$$

So, as shown in the slide, the variance of x bar is equal to sigma squared by n; the mean of x bar was the same as the population mean mu. On similar lines, you might have expected the variance of x bar to be also equal to sigma squared, but it is not so. As given here, it can be seen that the variance of x bar is scaled down by the sample size from sigma squared the population variance to sigma squared by n the sampling distribution variance. How this was obtained? We can see in this particular slide. When you take the variance x 1 it is equal to sigma squared, but when you have variance of x 1 by n, then it becomes sigma squared by n squared. You remember that we defined x bar as x 1 plus x 2 plus so on to x n divided by n. So, when you apply variance of x bar, it will be variance of x 1 by n plus variance of x 2 by n plus so on to variance of x n by n. So, earlier it was the expected value, now we are applying the variance, and we get n sigma squared by n squared, which is sigma squared by n. So this is very interesting.

What it is telling us is the sampling distribution of the mean has a smaller spread than the population distribution of x. So, the larger the value of the sample size the smaller is the variability, which is also making lot of sense. If you a collect sample of large size, then several such samples of large sizes may not have much differences between them in terms of the mean - the sample mean - but if you take a very small sample size, and you take a ten such samples, there is a strong chance that all these ten samples have very different means. So, the mean values get spread, but when the sample size is large, then the means are pretty much close to one another and their variability is less. So, variance of x bar, the sampling distribution of the means, is reduced if you increase the sample size, and the variance of x bar, as i said earlier, is sigma squared by n where sigma squared is the population variance and n is the sample size.

(Refer Slide Time: 10:32)



Properties of the Sampling Distribution

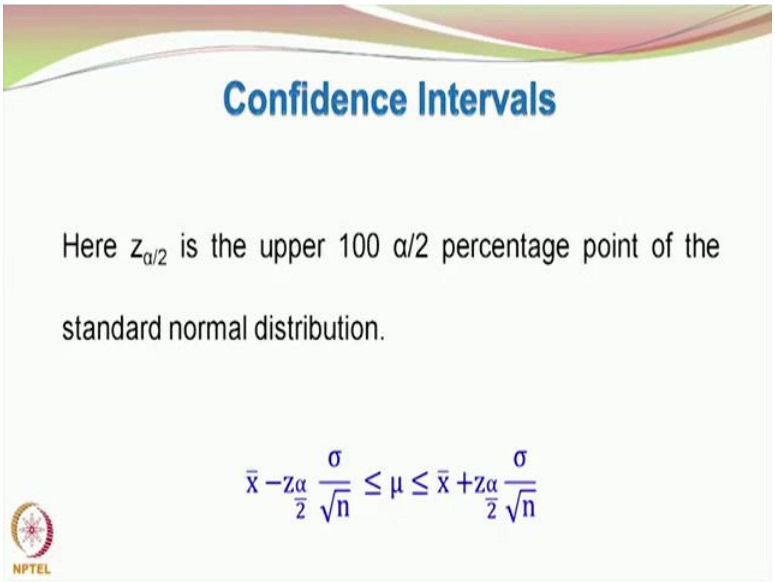
- If the population's distribution is normal with mean (μ) and variance (σ^2), then the sampling distribution of means is also normal with mean (μ) and variance (σ^2/n).
- It is assumed here that the random variables constituting the sample are independent.

So, to reiterate this, if the population distribution is normal with mean mu and variance sigma squared, then the sampling distribution of means is also normal with mean mu and variance sigma squared by n. Here we have introduced another rider to this, if the parent population is normal, then the sampling distribution can also be shown to be normally distributed and the mean and variance are mu and sigma squared by n respectively. So, an important assumption is made here that the random variables constituting the sample are independent of each other.

So far we have been talking about point estimators and point estimates, but we can also have interval estimates. I will give a very simple example for this. Let us say that you are going to a remote place where the train runs through the village only once in the day. Obviously, after finishing the work in the village you want to get back to your place as early as possible, and you don't want to miss that train, and be delayed for another day. So, you might ask the people at what time the train arrives to the station and you may get either a point estimate or you may get interval estimates. Point estimate may be the average time of arrival of the train to the station is let us say 2:10 okay; some people may say 2:10 pm, some people might 2:30 pm, some people may say 2:15 pm and so on.

On the other hand, there may be some people who may say that the train is going to come between 2 pm and 2:30 pm. So, a single valued estimate of a train's average time of arrival is a point estimate, whereas the interval specified on the average time of the train arrival is an interval estimate. So, from the same sample we can also construct an interval estimate.

(Refer Slide Time: 12:37)



Confidence Intervals

Here $z_{\alpha/2}$ is the upper 100 α/2 percentage point of the standard normal distribution.

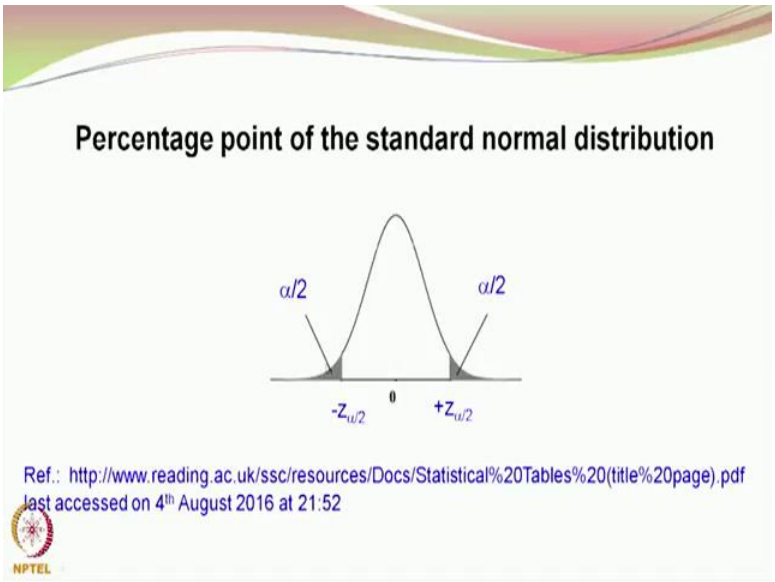
$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

So, if x bar is a sample mean of a random sample of size n obtained from a normal population with known variance sigma squared, then the hundred into 1 minus alpha percent confidence interval on mu is given by x bar minus z alpha by 2 sigma by root n less than or equal to mu less than or equal to x bar plus z alpha by 2 sigma by root n okay. Sounds a bit abstract, but it is quite straight forward after all. We will go through it one by one.

So, here we are assuming that sigma is known to us and then we can construct a confidence interval around the parameter mu, the population mean okay. Please note that we are always defining the confidence interval on the population parameter using the sample mean x bar. Different samples will give different values of x bar, and obviously, the intervals also will be different. And the unknown terms here are alpha and z alpha by 2 other terms are pretty straight forward sigma is the standard deviation of the population and as the sample size in x bar is the sample mean.

Okay now let us see what alpha and z alpha by 2 are. We can define z alpha by 2 as the upper 100 alpha by 2 percentage point of the standard normal distribution. We saw that capital Z is a random variable describing the standard normal variable. We defined capital Z as, if you recollect, Z is equal to x minus mu by sigma and that led any normal distribution to be reduced to a normal distribution with mean zero and variance or standard deviation unity. So, this is the standard normal distribution we are referring to, and then, what exactly is meant by z alpha by 2? Please refer to the following diagram.

(Refer Slide Time: 14:44)



Percentage point of the standard normal distribution

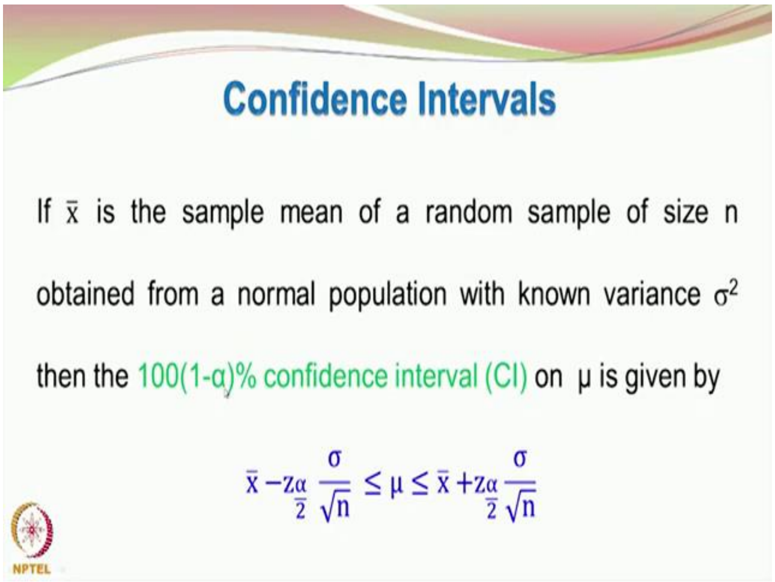
Ref.: http://www.reading.ac.uk/ssc/resources/Docs/Statistical%20Tables%20(title%20page).pdf last accessed on 4th August 2016 at 21:52

You have the standard normal distribution with mean zero and variance 1. Let us choose points z alpha by 2 and minus z alpha by 2 in this normal distribution. Please note that below the values of zero, the x-axis value are negative and above the value of zero it is positive. And we also note that the normal distribution is symmetric in nature. So, let us define points z alpha by 2 and minus z alpha 2. Due to the symmetric nature of the normal distribution z alpha by 2 and minus z alpha by 2 are equidistant from the origin, and similarly, by the same arguments the area under the curve will also be the same in both the cases. We are talking about the tail end of the normal distribution curve. So, you have this area to be alpha by 2 and this area is also alpha by 2. So, we identify points z and minus z such that the areas beyond those points are identically equal to alpha by 2.

So, this is alpha and this alpha is called as the level of significance. Usually we take alpha to be 0.05; that means, the area under the curve is 0.025 here and this is also 0.025 here. So, we have to look at the standard normal probability distribution table and see what is the value of z such that the area in the tail region beyond the value of z is alpha by 2. And so, if this for example, comes to a certain number minus of that number would be this number here. So, that takes care of this as I said alpha by 2, and this is also plus z alpha by 2, this is minus z alpha by 2. So, then knowing the value of x bar and assuming that the standard deviation is known of the population, we can construct the confidence interval.

What exactly is meant by the confidence interval? If you take hundred random samples, and the samples have different sample averages okay. So each will have a different value of x bar or most of them will have different values of x bar and z alpha by 2 sigma by root n is a constant number. So, when x bar changes from sample to sample, the intervals also will change from sample to sample. So, each interval may have different lower limit and upper limit. If you take 100 such samples, then if you are talking about a 95 percent confidence interval, then we mean that 95 percent of these 100 samples or 95 samples is expected to bracket the population mean mu. So, this is the meaning of the term confidence interval.

(Refer Slide Time: 18:23)



Confidence Intervals

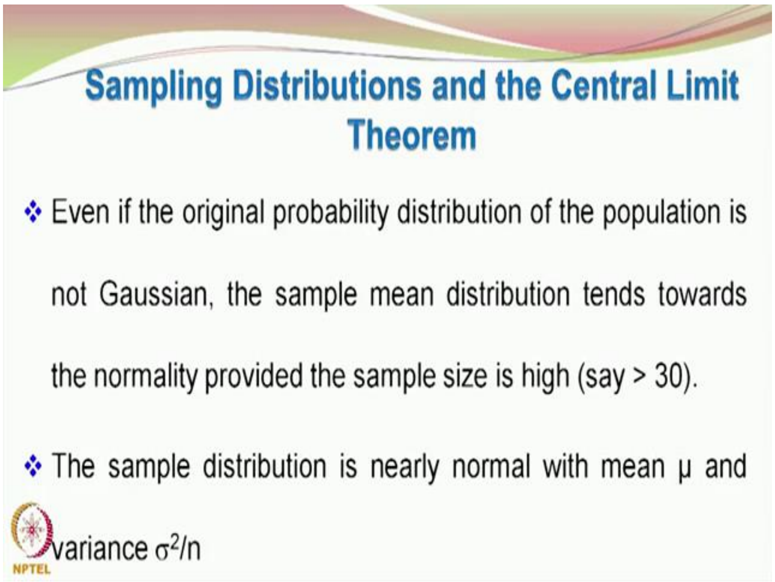
If $\bar{x}$ is the sample mean of a random sample of size n obtained from a normal population with known variance σ^2 then the 100(1-α)% confidence interval (CI) on μ is given by

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

So, what we are doing here is when you put alpha as 0.05, which is a common value as I said earlier, if you put the level of significance as 0.05, you get 1 minus 0.05, which is 0.95, you get then 95 percent confidence interval. And again, you can have a broad confident interval, and then, you can also have a narrow confidence interval. It is like a person saying that the train will be coming between 1 and 5. Obviously, this confidence interval is more or less definite to bracket the actual arrival of the train, but this is very vague. You may be having a tight schedule and we cannot be waiting in the station right from 1 ‘o’ clock.

On other hand, we can also have a very precise confidence interval which says that the train is going to come between 2 and 2:15. That sounds like a very good and precise confidence interval because it is not very big, but on other hand we are not very confident about this confidence interval because the train for all reasons might have come at 1:55 and left okay. So, again we are not very confident about highly precise confidence interval, but very broad confidence interval is less precise, but there is more confidence attached to it. So, these are some of the implications of the confidence interval definition and you will come across confidence intervals quite frequently in design of experiments. So, it’s very important to know what really the confidence interval is all about.

(Refer Slide Time: 20:11)



Sampling Distributions and the Central Limit Theorem

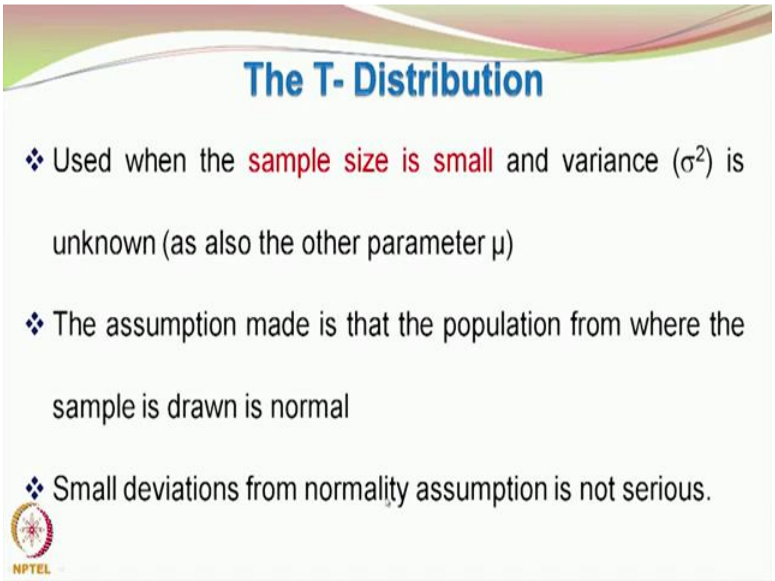
- Even if the original probability distribution of the population is not Gaussian, the sample mean distribution tends towards the normality provided the sample size is high (say > 30).
- The sample distribution is nearly normal with mean μ and variance σ^2/n

Now, we were discussing earlier that the normal distribution is preferable, it is unimodel and symmetric, and its properties are well known. So, it’s like having good thing to use. If the parent population is normal, and you take samples of any size from the normal distribution, the population distribution is normal, and you take samples of any size. It can be even a small size of 3 or 4, and then the resulting sampling distribution of the means is also going to be normal. On the other hand, if you do not know about the population probability distribution it may be normal or it may not be normal.

Let us assume that it is not normal, then you take samples from this population. We take different samples from this population, and each sample is let us saying of size 35. You are going to get a sampling distribution of the means. This sampling distribution is tending towards normality because you have taken a large sample size. Just because you have taken a large sample size you get an added benefit that the resulting sampling distribution of the means is tending towards normality. This is the a very useful thing to have, and even though we do not know much about the population distribution, we are getting a nice nearly normal distribution for the sampling distribution of the means. This is termed as the Central Limit Theorem.

So, even if the original probability distribution of the population is not normal or Gaussian - Gaussian is another term for the normal distribution - the sample mean distribution tends towards the normality provided the sample size is high, say greater than 30. A sample distribution is nearly normal with mean mu and variance sigma squared by n.

(Refer Slide Time: 22:21)

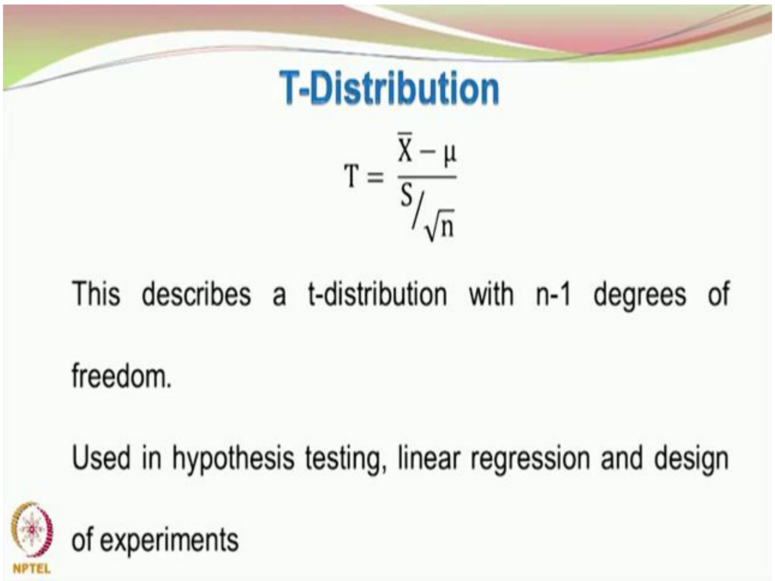


The T- Distribution

- Used when the sample size is small and variance (σ^2) is unknown (as also the other parameter μ)
- The assumption made is that the population from where the sample is drawn is normal
- Small deviations from normality assumption is not serious.

You might have also come across in several papers or books, the t-distribution; we will not be getting into all the details about the t-distribution. This t-distribution is used when the sample size is small and the variance is unknown. So far when discussing about the confidence intervals, for example, we have assumed that the sample, not sample, the population variance sigma squared is known. Now, when you have a small sample size, and the variance sigma square is not known, which is usually the case, for several reasons you might be forced to take small samples, and also, you may not know the real population variance value. So, now, the assumption made is that the population from where the sample is drawn is normal. So, this is not very serious assumption. So, many populations tend towards normality, and so, making the assumption that the population from which the sample is drawn is normal it is not a very serious one.

(Refer Slide Time: 23:26)



T-Distribution

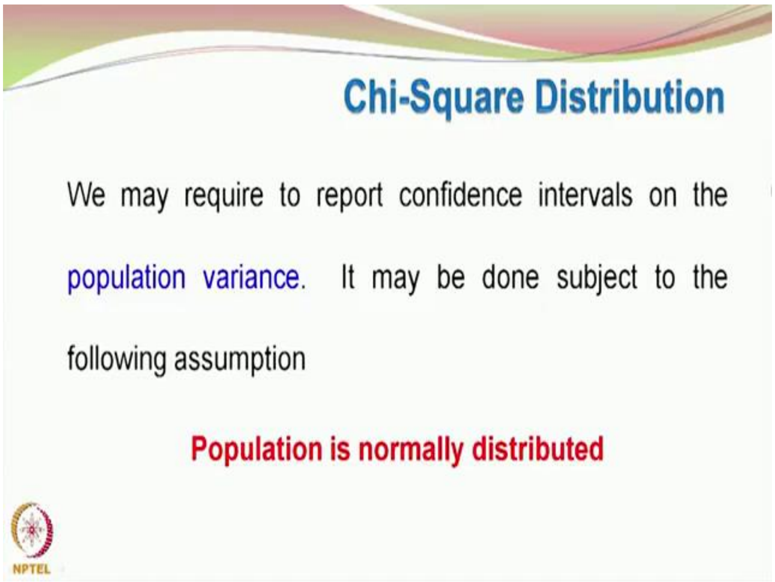
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

This describes a t-distribution with n-1 degrees of freedom.

Used in hypothesis testing, linear regression and design of experiments

Now, we define the t random variable as x bar minus mu by s by root 10. This describes a t-distribution with n minus 1 degrees of freedom and this t-distribution is often used in hypothesis testing, linear regression, and design of experiments. So, you can see the degrees of freedom again coming into the picture; s is the sample standard deviation, it is not the population standard deviation. Population standard deviation is given by sigma. The sample standard deviation is s. So, instead of using sigma, we are actually using s - the sample standard deviation. And again, we saw from the definition of the sample standard deviation, that only n minus 1 entities are independent, and so, we have only n minus 1 degree of freedom.

(Refer Slide Time: 24:22)



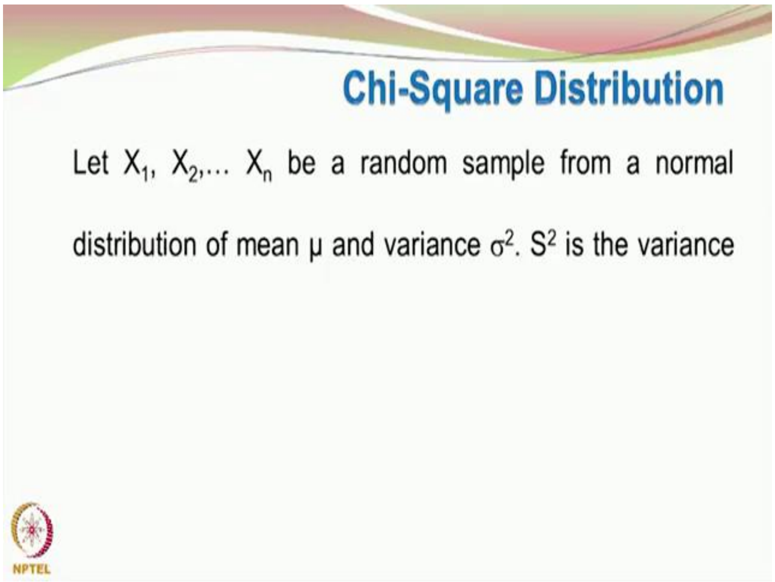
Chi-Square Distribution

We may require to report confidence intervals on the population variance. It may be done subject to the following assumption

Population is normally distributed

Now, we also come frequently across the chi-square distribution. Just as we report confidence interval on the mean, we saw it just a few minutes back, we may have to report confidence intervals on the population variance and in this connection the chi-square distribution is very useful, and again, it is based on the assumption that the population is normally distributed.

(Refer Slide Time: 24:47)

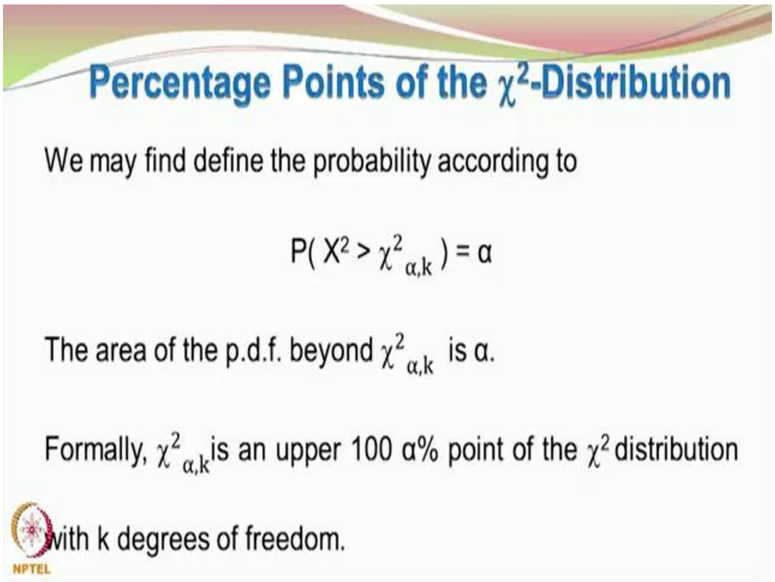


Chi-Square Distribution

Let X_1, X_2,… X_n be a random sample from a normal distribution of mean μ and variance σ^2. S^2 is the variance

So let x 1, x 2, so on to x n be a random sample from a normal distribution of mean mu and variance sigma squared; s squared is the variance of this sample. So, let us see now how to define the chi-square distribution. Let x 1, x 2, so on to x n be a random sample from a normal distribution of mean mu and variance sigma squared; s squared is the variance. So, we define the random variable capital chi-square as n minus 1 s squared by sigma squared and we call it as a chi-square distribution with n minus 1 degrees of freedom.

(Refer Slide Time: 25:33)



Percentage Points of the χ^2-Distribution

We may find define the probability according to

$$P(X^2 > \chi^2_{\alpha,k}) = \alpha$$

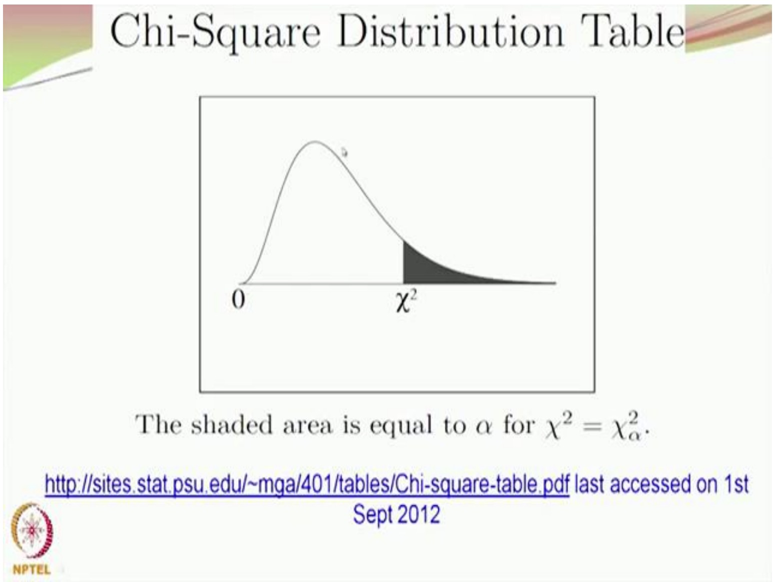
The area of the p.d.f. beyond $\chi^2_{\alpha,k}$ is α.

Formally, $\chi^2_{\alpha,k}$is an upper 100 α% point of the χ^2 distribution with k degrees of freedom.

We may find the probability according to the following equation, probability of capital chi-square is greater than chi-squared alpha k that is equal to alpha. In other words, we are finding out the value of the chi-squared random variable such that the area of the curve beyond that particular value is equal to alpha in the chi-squared probability distribution function.

The area of the probability distribution function beyond chi-squared alpha k is given by alpha. We may define formally, the chi-squared alpha k as an upper 100 alpha percent point of the chi square distribution with k degrees of freedom. So, this k degrees of freedom is represented by the subscript k by here. So, this is similar to what we saw earlier in the confidence interval where we defined z alpha by 2 with respect to the normal probability curve, and then we said we are identifying two points z alpha by 2 minus z alpha by 2, such that the area of the normal distribution curve beyond the z alpha by 2 and below minus z alpha by 2 is equal to alpha. Now, in this chi-square distribution called distribution function we are looking at only one end of the curve - the tail end of the curve. I will demonstrate this in a minute.

(Refer Slide Time: 27:03)



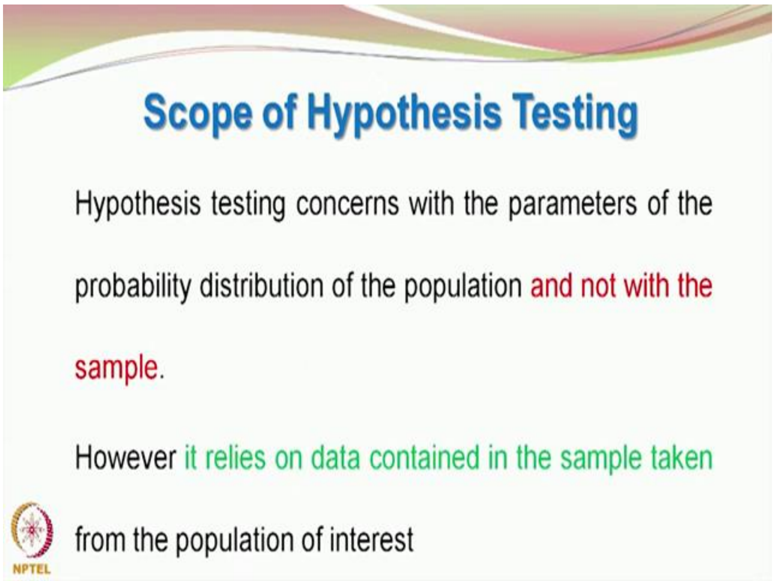
Chi-Square Distribution Table

The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf last accessed on 1st Sept 2012

So, let us look at the chi-square distribution. Here it is starting from zero, it is going to have only positive values, you can’t have negative values, and then, you are identifying the chi squared value for the appropriate degrees of freedom, such that the area in the curve beyond this chi-squared value is alpha okay - so that is the definition for the probability. Probability of capital chi-square greater than chi-squared alpha k is equal to alpha for the specified k degrees of freedom.

(Refer Slide Time: 27:39)



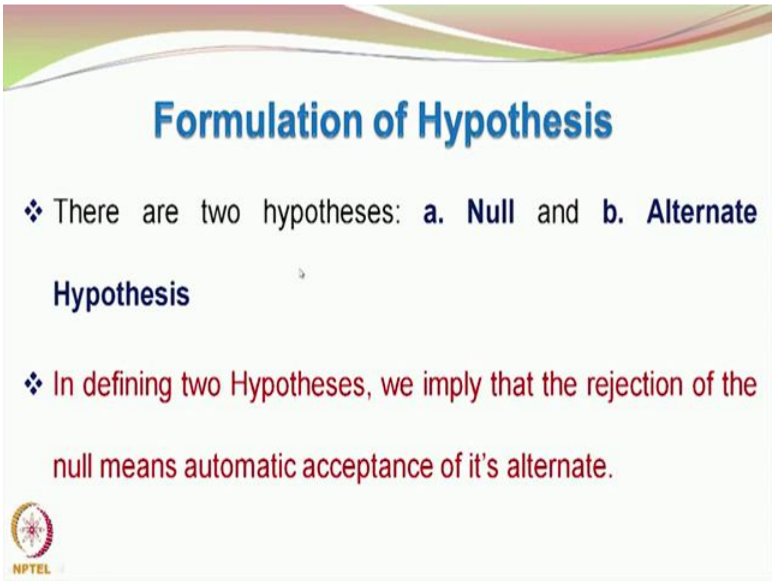
Scope of Hypothesis Testing

Hypothesis testing concerns with the parameters of the probability distribution of the population and not with the sample.

However it relies on data contained in the sample taken from the population of interest

Now, let us come to another important application namely the hypothesis testing okay. This is again going to play an important in design of experiments. What do you mean by hypothesis testing? You come up with the supposition or you come up with the claim or a statement and we are going to investigate whether this particular claim or supposition is valid or it has a strong refutation. So, this hypothesis testing concerns with parameters of the probability distribution of the population and not with the sample. You are making hypothesis pertaining to the parameters of the population; you are not making hypothesis with respect to the sample, but with the population; but in order to make suppositions or hypothesis regarding the population we use the information contained in the samples.

(Refer Slide Time: 28:37)

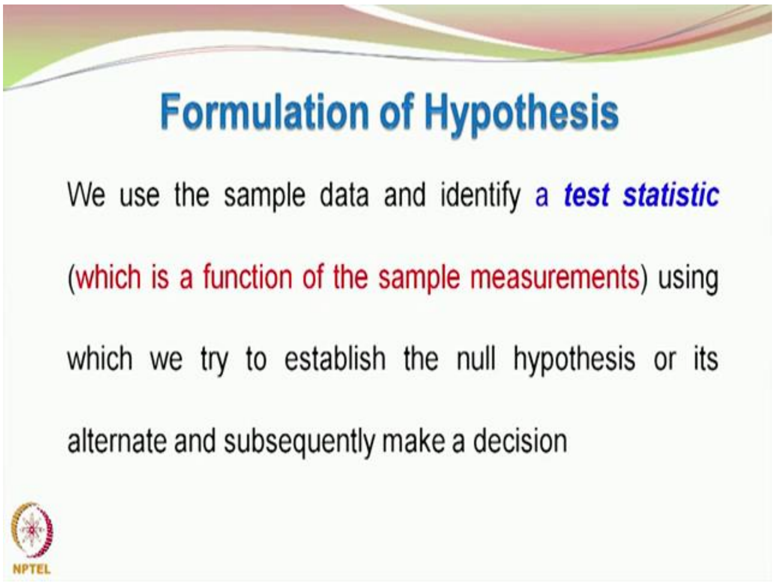


Formulation of Hypothesis

- There are two hypotheses: **a. Null** and **b. Alternate Hypothesis**
- In defining two Hypotheses, we imply that the rejection of the null means automatic acceptance of it's alternate.

There are two hypotheses: one is the null hypothesis and the other is the alternate hypothesis. The nullification of the original hypothesis is the alternate hypothesis. In defining the two hypotheses, we imply that the rejection of the null means automatic acceptance of its alternative. In other words, if you are not accepting the null hypothesis you accept the alternate hypothesis.

(Refer Slide Time: 29:09)

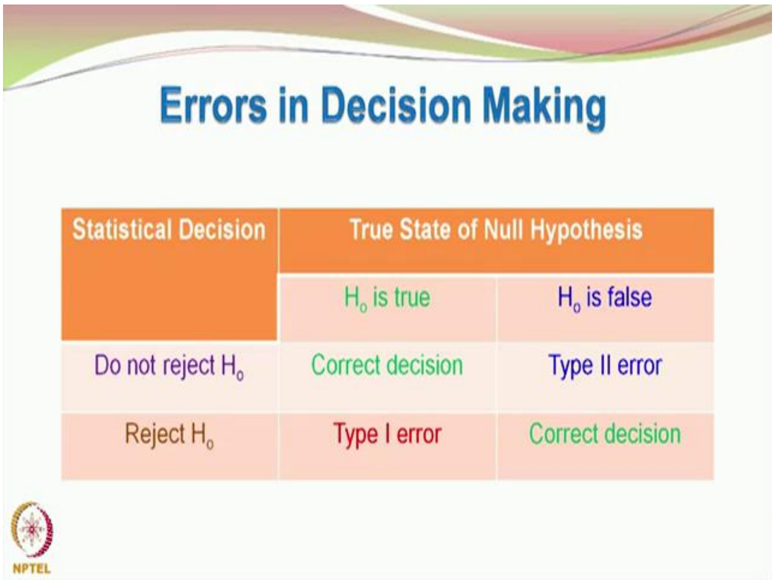


Formulation of Hypothesis

We use the sample data and identify a ***test statistic*** (which is a function of the sample measurements) using which we try to establish the null hypothesis or its alternate and subsequently make a decision

So, when you formulate the hypothesis we identify a test statistic. The two test statistics we have seen so far are the sample mean definition and the sample variance. So, we identify a test statistic using which we try to establish the null hypothesis or it's alternate and then subsequently make a decision.

(Refer Slide Time: 29:35)



Errors in Decision Making

Statistical Decision	True State of Null Hypothesis	
	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Now, let's start of, with an example. You are having particular running plant and a person is newly recruited from a reputed institution. He comes over there and says that this process is not good enough; I have an idea which will improve the process. The management is a bit skeptical not because it wants to discourage the youngster, but already you are having a well running process and it is making profits. So, why do you want to tinker with the process already existing or running and successfully, and commit money, man power, time, resources, etcetera to try out the new process.

The management is also skeptical that the new process may not be any significant or considerable improvement over an already existing one. So, the null hypothesis in this case would be the processes proposed is actually not producing any improvement. The alternate hypothesis, obviously, would then be the refutation of the null hypothesis - the suggested process is in fact bringing a significant improvement. So, the null hypothesis is the suggested process is not good or not producing any considerable improvement and the alternate hypothesis would be the new process is in fact better than the old process or the existing process. Another, more easier example is, suppose the court is investigating a particular crime and the prosecution is saying a particular person is guilty, the null hypothesis is the person is not guilty, the alternate hypothesis is the person is in fact guilty of committing a crime.

We will continue shortly.