

# **Stochastic Structural Dynamics**

**Prof. Dr. C. S. Manohar**

**Department of Civil Engineering**

**Indian Institute of Science, Bangalore**

**Module No. # 07**



**Lecture No. # 26**

**Monte Carlo Simulation Approach-2**

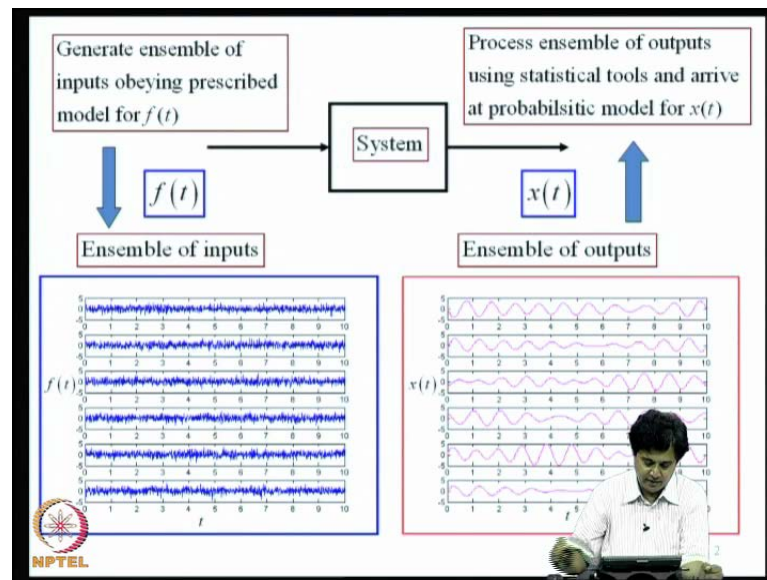
(Refer Slide Time: 00:21)

**Stochastic Structural Dynamics**  
Lecture-26  
Monte Carlo simulation approach-2

**Dr. C. S. Manohar**  
Department of Civil Engineering  
Professor of Structural Engineering  
Indian Institute of Science  
Bangalore 560 012 India  
[manohar@civil.iisc.ernet.in](mailto:manohar@civil.iisc.ernet.in)



(Refer Slide Time: 00:25)

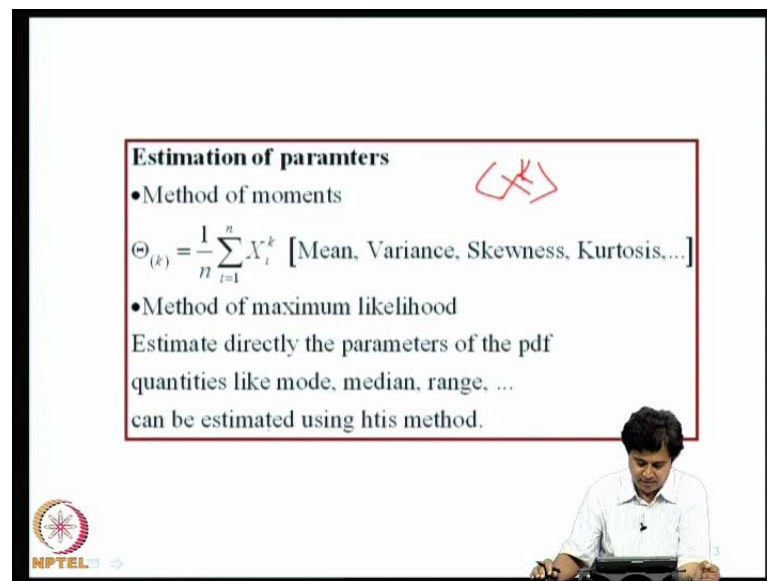


In the previous lecture, we started discussing an application of Monte Carlo simulation methods for analyzing response of randomly driven systems. So, the frame work of our discussion is as shown here: this is a system which is typically governed by set of stochastic differential equations and this is driven by Non symbol of inputs  $f t$ , which is the random process, and we would like to simulate samples  $f t$ , which are compatibles with a prescribed probabilistic model. Suppose if  $f t$  is a Gaussian random process with a given probabilistic density functions, suppose it is 0 mean and stationary, I should able to generate a Non symbol of time histories, which are, which are compatibles with the given P S D and probability density function.

Now, for each of this sample, we will integrate the governing equations of motion, and then get a ensemble of response quantities of interest; this ensemble of response quantities of interest, we will process statistically and arrive at probabilistic model for the response. So, given that we are basically approaching the problem through numerical simulation, the scope of this method is very worst, you can apply this method to any problem for which a sample calculation can be performed. One of the things that we have to appreciate at the outset is, when you simulate samples of  $f t$  compatibles with the target probabilistic model for  $f t$ , we need to ascertain that we have succeeded in doing so, and the tool that is needed is, to address that problem is methods of mathematical statistics.

Similarly, after we produce the ensemble of response, to process this ensemble of response time histories, to arrive at a model for the response process, again, we need to use statistical methods. So, we have initiated a discussion on statistical methods, so, we will continue that in this lecture. And we are discussing the problem of estimation of parameters; so, one of the methods is method of moments, where we find basically quantities like expected value of X to the power of k.

(Refer Slide Time: 02:29)



**Estimation of parameters**

- Method of moments (X)

$$\Theta_{(k)} = \frac{1}{n} \sum_{i=1}^n X_i^k \quad [\text{Mean, Variance, Skewness, Kurtosis, ...}]$$

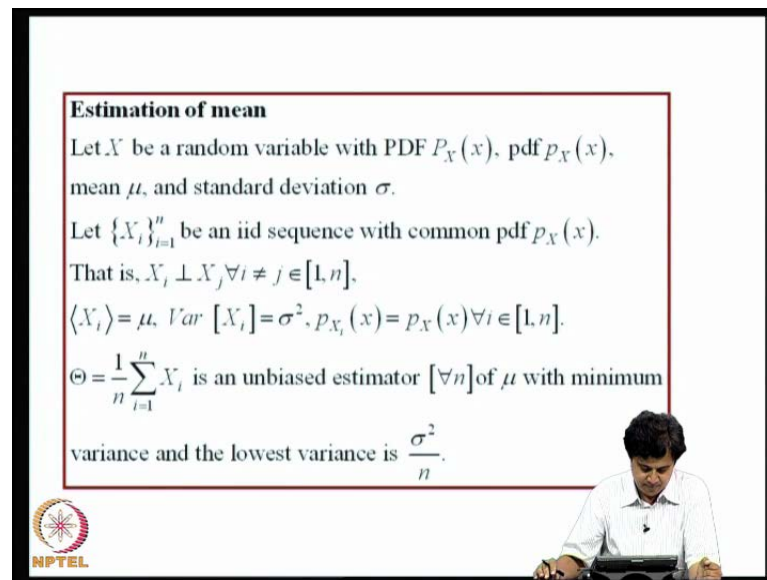
- Method of maximum likelihood

Estimate directly the parameters of the pdf  
quantities like mode, median, range, ...  
can be estimated using this method.

NPTEL

So, that would mean we can find mean, variance, skewness, kurtosis, etcetera. The other alternative method is, it directly estimates the parameter of an assumed probability density function- that is method of maximum likelihood. Here the parameter themselves need not be one of these moments, they can, they will be related to these moments, but they may not be directly those quantities; similarly, quantities like mode, median, range etcetera., cannot be estimated using method of moments.

(Refer Slide Time: 03:26)



**Estimation of mean**



Let  $X$  be a random variable with PDF  $P_X(x)$ , pdf  $p_X(x)$ , mean  $\mu$ , and standard deviation  $\sigma$ .

Let  $\{X_i\}_{i=1}^n$  be an iid sequence with common pdf  $p_X(x)$ .

That is,  $X_i \perp X_j \forall i \neq j \in [1, n]$ ,

$\langle X_i \rangle = \mu$ ,  $Var [X_i] = \sigma^2$ ,  $p_{X_i}(x) = p_X(x) \forall i \in [1, n]$ .

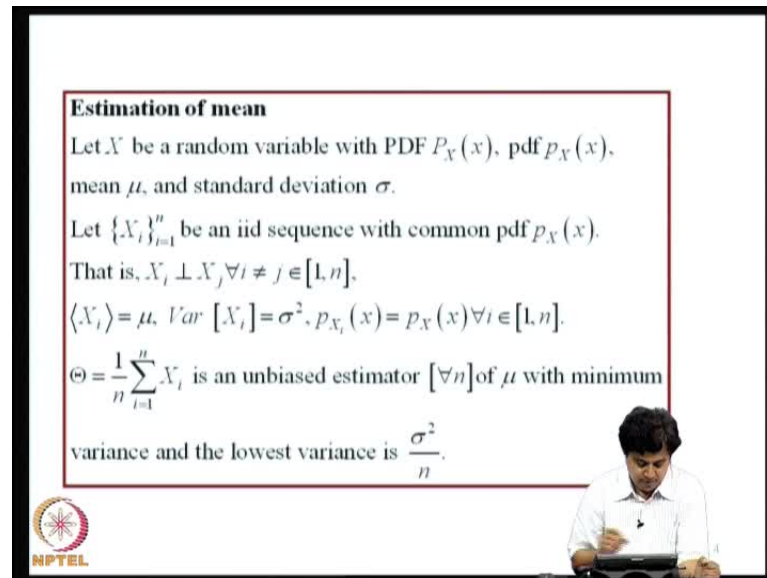
$\Theta = \frac{1}{n} \sum_{i=1}^n X_i$  is an unbiased estimator  $[\forall n]$  of  $\mu$  with minimum variance and the lowest variance is  $\frac{\sigma^2}{n}$ .

So, the maximum likelihood estimation method helps us to address some of these issues. So, we are discussing estimation of the mean, so, we started by assuming that  $x$  is the random variable with the, probability density function,  $P_x$  of  $x$ , probability density function, **lower case  $P_x$  of  $x$** , mean  $\mu$  and standard deviation  $\sigma$ ; we formed a sequence of iid that is, identical and independently distributed random variables with the common probability distribution function which is  $P_x$  of  $x$ , which agrees with the probability density function of the random variable that I am talking about that is,  $X_i$  is independent of  $X_j$  for all  $i$  not equal to  $j$  from 1 to  $n$ , and each one of this exercise has mean  $\mu$ , variance  $\sigma^2$  and probability density function  $P_x$  of  $x$ .

In the previous lecture, I showed that  $\theta$  given by  $\frac{1}{n} \sum_{i=1}^n X_i$  is an unbiased estimator for all  $n$  of  $\mu$  with minimum variance, and the lowest variance is  $\sigma^2/n$  - that means, this an unbiased estimator irrespective of the size of the sample. And the lowest variance is  $\sigma^2/n$ , so, as  $n$  become large this variance reduces; on an average the statistic provides an exact solution to the problem of parameter estimation of the mean of the population.

(Refer Slide Time: 04:55)



**Estimation of mean**



Let  $X$  be a random variable with PDF  $P_X(x)$ , pdf  $p_X(x)$ , mean  $\mu$ , and standard deviation  $\sigma$ .

Let  $\{X_i\}_{i=1}^n$  be an iid sequence with common pdf  $p_X(x)$ .

That is,  $X_i \perp X_j \forall i \neq j \in [1, n]$ .

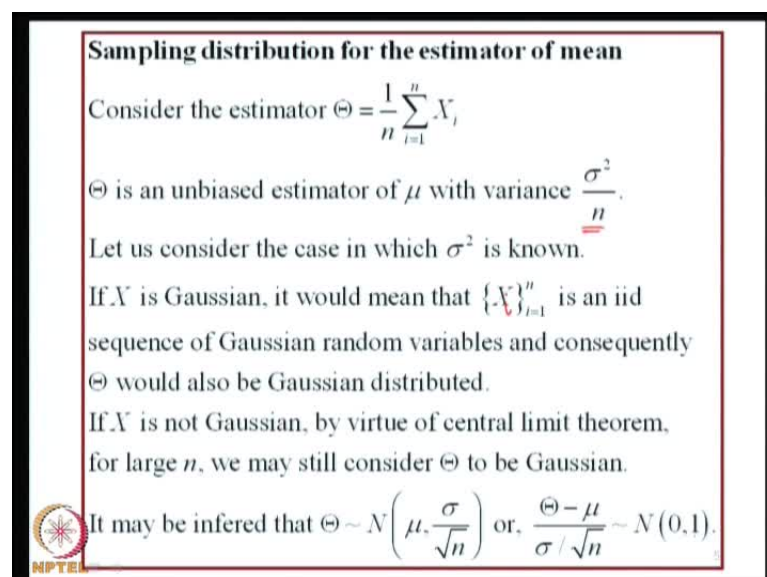
$\langle X_i \rangle = \mu$ ,  $Var [X_i] = \sigma^2$ ,  $p_{X_i}(x) = p_X(x) \forall i \in [1, n]$ .

$\Theta = \frac{1}{n} \sum_{i=1}^n X_i$  is an unbiased estimator  $[\forall n]$  of  $\mu$  with minimum variance and the lowest variance is  $\frac{\sigma^2}{n}$ .



Now, we now talk, we, it is clear that since  $X_i$ 's are random variables here, and  $\theta$  is a transformation on random variables,  $\theta$  is also a random variable; the probability distribution function of  $\theta$  is known as sampling distribution for mean, and its standard deviation is known as standard error. So, we are now interested in postulating a model for this sampling distribution of  $\theta$ ; so,  $\theta$  is an unbiased estimator of  $\mu$  with variance  $\sigma^2/n$ .

(Refer Slide Time: 05:15)



**Sampling distribution for the estimator of mean**

Consider the estimator  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i$


$\Theta$  is an unbiased estimator of  $\mu$  with variance  $\frac{\sigma^2}{n}$ .

Let us consider the case in which  $\sigma^2$  is known.

If  $X$  is Gaussian, it would mean that  $\{X_i\}_{i=1}^n$  is an iid sequence of Gaussian random variables and consequently  $\Theta$  would also be Gaussian distributed.

If  $X$  is not Gaussian, by virtue of central limit theorem, for large  $n$ , we may still consider  $\Theta$  to be Gaussian.

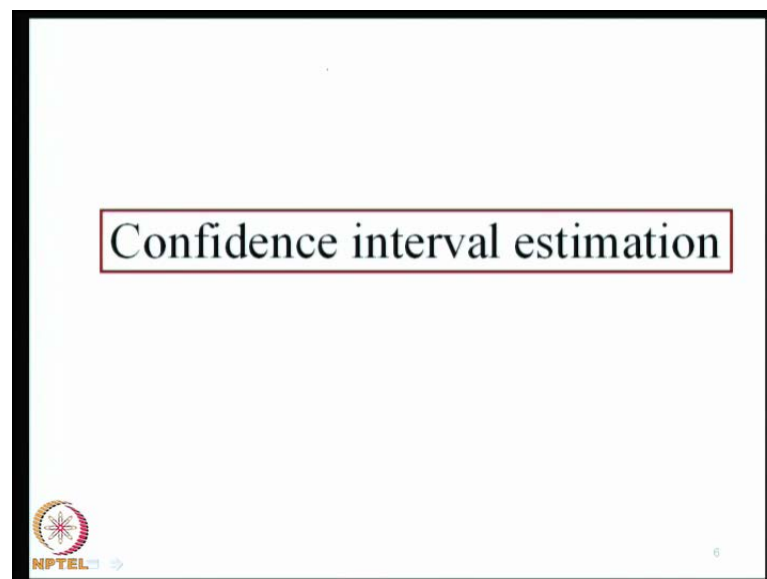
It may be inferred that  $\Theta \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$  or,  $\frac{\Theta - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ .



Now, let us begin by considering the case in which variance is known. Now, if  $x$  is Gaussian, it would mean that all these  $X_i$ 's would also be Gaussian because they are i.i.d. sequence with the common pdf, which is Gaussian, and since we are adding this Gaussian random variables,  $\theta$  would also be Gaussian; so, in this case the sampling distribution for  $\theta$  would be Gaussian with mean  $\mu$  and variance  $\sigma^2/n$ .

However, if  $x$  is not Gaussian, by virtue of central limit theorem and for large  $n$ , we may still consider  $\theta$  to be Gaussian; so, this is an approximation which we generally make and therefore, we assume that  $\theta$  is normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ , or if we form a standard normal variable, we remove the mean and divide by standard deviation, and this  $(\theta - \mu)/(\sigma/\sqrt{n})$  is normally distributed with 0 mean and unit standard deviation. So, this is the sampling distribution for the mean

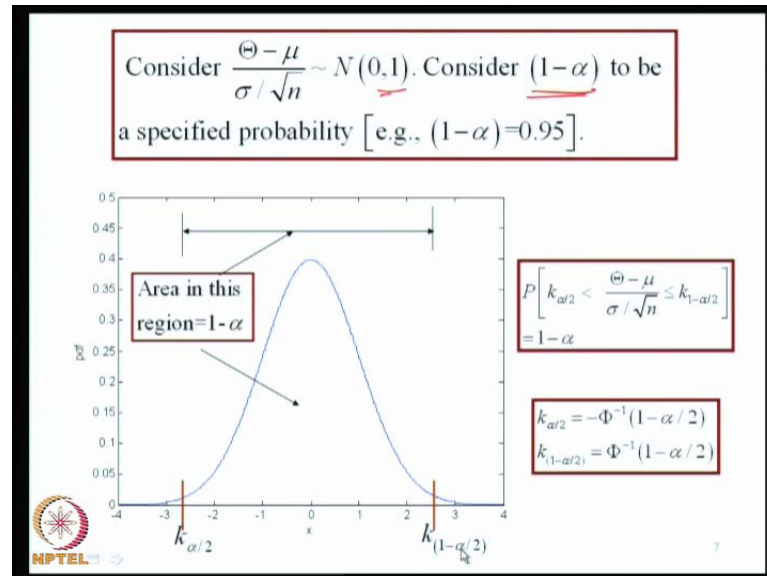
(Refer Slide Time: 06:36)



Now, based on the idea of sampling distribution, we can construct what are known as confidence interval estimation. So, in the discussion that we had till now, if you use this for a given realization of  $X_i$ , we will get a realization of  $\theta$ , and this is the point estimator, you get one realization of  $\theta$ . So, this is one way of answering the question, but there is another way known as confidence interval estimation, so, let us discuss what it is.

Now, we have just now showed that the random variable theta minus mu divided by sigma divide by square root n is normal with zero mean and unit standard deviation.

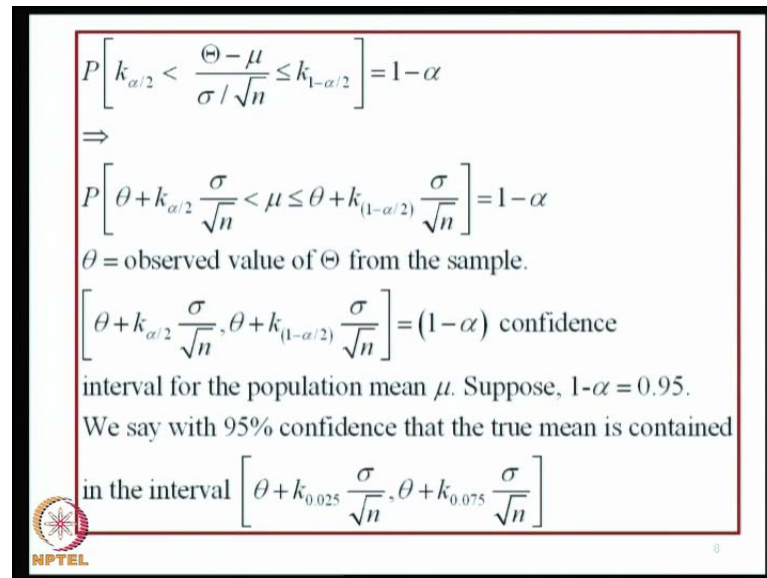
(Refer Slide Time: 07:08)



Now, let us consider a probability level 1 minus Alpha associated with this random variable for example, 1 minus Alpha could be .95 that means, Alpha is .05. Now, if you draw the probability density function of the random variable as shown here, we define two points such that the area between these two points is equal to 1 minus Alpha, so, that means, the definition is, I consider probability, I am defining two points K of Alpha by 2 and K of 1 minus Alpha by 2, that are defined as follows. The probability that the random variable theta minus mu divided by this sigma divided by square root n, lies in the interval K Alpha by 2 K 1 minus Alpha by 2, is 1 minus Alpha.

Now, this K Alpha by 2 is actually the inverse probability distribution function evaluated at 1 minus Alpha by 2 **minus of that**, and K of 1 minus Alpha by 2 is phi inverse of 1 minus Alpha by 2, Alpha is given here, so, we are determining K Alpha by 2 and K 1 minus Alpha by 2.

(Refer Slide Time: 08:37)


$$P\left[k_{\alpha/2} < \frac{\Theta - \mu}{\sigma / \sqrt{n}} \leq k_{1-\alpha/2}\right] = 1 - \alpha$$
$$\Rightarrow$$
$$P\left[\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$\theta$  = observed value of  $\Theta$  from the sample.

$$\left[\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right] = (1 - \alpha) \text{ confidence}$$

interval for the population mean  $\mu$ . Suppose,  $1 - \alpha = 0.95$ .

We say with 95% confidence that the true mean is contained

in the interval  $\left[\theta + k_{0.025} \frac{\sigma}{\sqrt{n}}, \theta + k_{0.975} \frac{\sigma}{\sqrt{n}}\right]$

So, we have now the statement, probability that theta minus mu divided by sigma divided by square root n, lying between  $K_{\alpha/2}$  and  $K_{1-\alpha/2}$  is  $1 - \alpha$ .

Now, I can rearrange these terms, and I can write this as probability that the interval  $\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  and  $\theta + k_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  contains this mean is  $1 - \alpha$ .  $\theta$  is observed value of  $\Theta$  from the sample, so what we say is that, this interval  $\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ,  $\theta + k_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$  is the confidence interval on population parameter  $\mu$  with confidence  $1 - \alpha$ . Suppose  $1 - \alpha$  is  $0.95$  so, what we are telling is, with 95 percent confidence I can say that the true mean is contained in the interval  $\theta + k_{0.025} \frac{\sigma}{\sqrt{n}}$  to  $\theta + k_{0.975} \frac{\sigma}{\sqrt{n}}$ .



(Refer Slide Time: 10:13)

$$P\left[\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

**Remark**

- This should be interpreted as the probability that the random interval  $\left(\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right)$  contains the population mean  $\mu$  is  $1 - \alpha$ .
- Remember  $\mu$  is a deterministic quantity.
- $\theta = \frac{1}{n} \sum_{i=1}^n x_i$  is a point estimate &  $\left(\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right)$  is a confidence interval estimate.

So, instead of telling the estimate of population parameter is a single number, now, I am providing an interval, that means this interval encloses the true population mean with a given level of confidence; so, these are much more useful way of providing the answer.

So, we can make some remarks now: this statement that  $\mu$  lies between these two numbers should be interpreted as the probability that the random interval, the random interval  $\theta + K \text{ Alpha by } 2 \text{ sigma by square root } n$  and  $\theta + K 1 \text{ minus Alpha by } 2 \text{ by sigma by square root } n$  contains the population  $\mu$ , that probability is  $1 \text{ minus Alpha}$ ;  $\mu$  is not a random variable,  $\mu$  is a deterministic constant, so, this should not be constituted as a probability statement made on  $\mu$ ;  $\mu$  is a not at all random variable, the random variables are here, so, this is the random interval, this another random interval, so, this is a one random variable, this is another random variable, so the difference between the two can be constituted as a random interval, and that probability, that random interval encloses the population mean is  $1 \text{ minus Alpha}$ .

So,  $\theta = \frac{1}{n} \sum_{i=1}^n x_i$  is a point estimate, and this interval is the confidence interval estimate for population mean.

(Refer Slide Time: 11:31)

Example


$x =$	
-0.4326	
-1.6656	
0.1253	
0.2877	
-1.1465	
1.1909	
1.1892	
-0.0376	
0.3273	
0.1746	

$$\theta = \frac{1}{10} \sum_{i=1}^{10} x_i = 0.0013$$
$$\alpha = 0.05$$
$$k_{\alpha/2} = -1.96 \text{ \& } k_{1-\alpha/2} = 1.96$$

95% confidence interval

$$= \left( 0.0013 - \frac{1.96}{\sqrt{10}}, 0.0013 + \frac{1.96}{\sqrt{10}} \right)$$
$$= (-0.6068, 0.6328)$$

The point estimate of mean is 0.0013.  
With 95% confidence we say that the population mean is contained in the interval  $(-0.6068, 0.6328)$ .

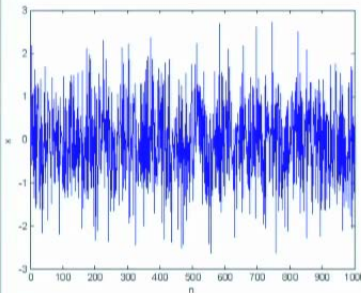


10



Some quick examples: suppose I take ten numbers as displayed here, and I find a point estimator, you can verify that it is 0.013. Now, if Alpha is 0.05, I can find out K Alpha by 2 is 1 point minus 1.96 and K of 1 minus Alpha by 2 1.96; so, with 95 percent confidence I can say that the population mean lies between minus 0. 6068 and 0.6328 whereas, according to this, the estimate for the population mean is 0.0013 with ten samples. That is all that I can say with this statement whereas, here what I am able to say is with 95 percent confidence the interval minus 0.6068, 0.6328 contains the mean.

(Refer Slide Time: 12:39)

$n=1000$

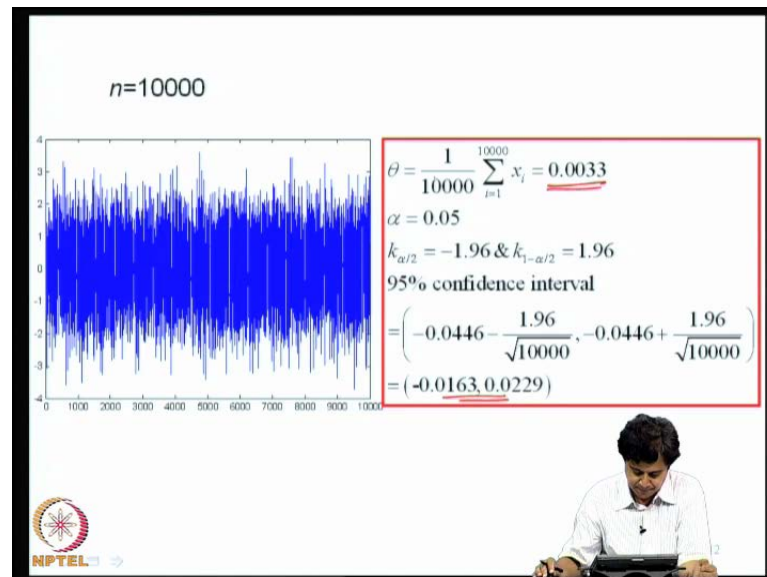

$$\theta = \frac{1}{1000} \sum_{i=1}^{1000} x_i = -0.0446$$
$$\alpha = 0.05$$
$$k_{\alpha/2} = -1.96 \text{ \& } k_{1-\alpha/2} = 1.96$$

95% confidence interval

$$= \left( -0.0446 - \frac{1.96}{\sqrt{1000}}, -0.0446 + \frac{1.96}{\sqrt{1000}} \right)$$
$$= (-0.1065, 0.0174)$$


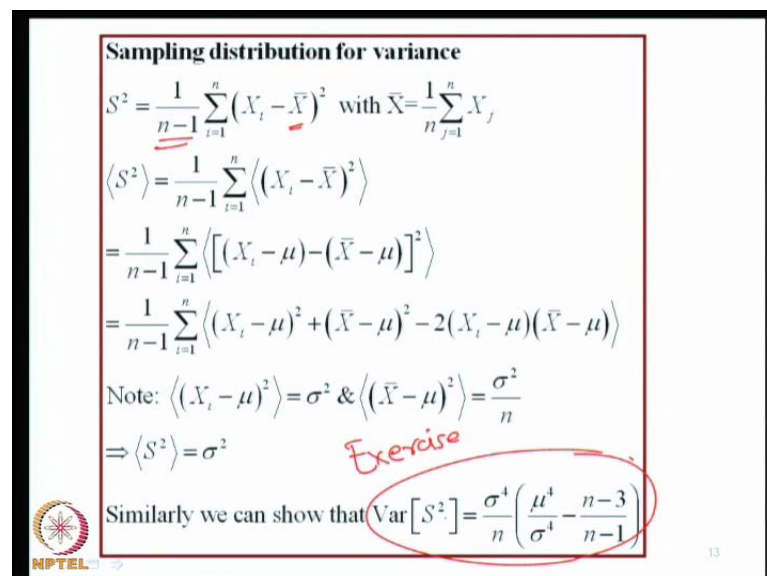
This is, you can see here, there were only ten samples and this interval is fairly big. Now, if I use thousand samples- I just plotted these 1000 numbers- I get point estimate as point minus 0.0446, and again with 95 percent confidence I can say that the interval 0.1065, 0.0174 contains the population mean, I can say that with 95 percent confidence.

(Refer Slide Time: 13:07)



Now, with 10,000, this answer seems to be close to 0, but this confidence interval is shrinking; we are getting narrower and narrower confidence intervals because I am using larger number of samples. That is what the conclusion we can draw from this exercise.

(Refer Slide Time: 13:30)



I talked about sampling distribution for mean, so, in principle we should be able to construct sampling distribution for any statistic that you are interested in. For example, if you are interested in sampling distribution for variance, so, we use this estimator  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ , where  $\bar{X}$  is the point estimate  $\frac{1}{n} \sum_{j=1}^n X_j$ , this is the estimate for the mean. Now, you find the mean square value, I mean expected value of  $s^2$ , which is this, and for  $\bar{X}$  I am going to add and subtract  $\mu$  and rewrite this in a slightly different form, expand this, I get three terms, and if I manipulate these terms and use the fact that  $X_i$ 's are all identical having mean  $\mu$  and variance  $\sigma^2$ , and  $\bar{X}$  is an unbiased estimator with a known variance, which is  $\sigma^2/n$ , if I use this information I can show that this estimator is unbiased; please, see that I am dividing by  $n-1$  not by  $n$ , that is to be expected because  $\bar{X}$  is a number that has been computed from the sample. So, if I use  $n$  here instead of  $n-1$ , if I use  $1/n$ , it will be a biased estimator for variance, so this  $n-1$  ensures that it is unbiased.

We can show that the variance of this estimator is given by this; I leave this as an exercise. You have to understand carefully what is being said here, we are talking about variance of a random variable  $X$ , that itself, the estimator for that itself is the random variable, which has a mean, which is agrees with the population variance, but it being a random variable it has its own variance and that is given by this; and here you can see that as  $n$  tends to infinity this variance comes down therefore, this estimator is consistent.

(Refer Slide Time: 15:44)

$$\Rightarrow S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$
 is an unbiased and consistent estimator of  $\sigma^2$ .

$$\Rightarrow \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 - \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

If population is Gaussian,  $X_i$  and  $\bar{X}$  are Gaussian.  
 RHS: sum of squares of Gaussian random variables such sums have  $\chi^2$ -distributions.

$$\Rightarrow \frac{(n-1)S^2}{\sigma^2}$$
 is  $\chi^2$ -distributed with  $(n-1)$  dof.

The pdf of such a random variable is given by

$$p(u) = \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} u^{\frac{(n-1)}{2}-1} \exp\left(-\frac{u}{2}\right); 0 < u < \infty$$

NPTEL

So,  $s^2$  is equal to  $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  is an unbiased and consistent estimator for  $\sigma^2$ . Now, what is my objective of our discussion, the objective of discussion is to arrive at sampling distribution for variance. So, I will rearrange this term, I rewrite this  $n-1$  into  $s^2$  by  $\sigma^2$  and I rewrite in this form, if the population is Gaussian  $X_i$  and  $\bar{X}$  are Gaussian, then this  $\sum (X_i - \bar{X})^2$  will be sum of squares of Gaussian random variables and such sums have distribution known as chi square distribution; if you add  $n$  squares of Gaussian random variables, the resulting random variable has a distribution known as chi square distribution that can be shown, and the form of the distribution is displayed here.

So, this is a chi square distribution, probability density function of chi square distribution, chi square random variable with  $n-1$  degrees of freedom. So, this is the well studied probability density function, its properties are tabulated, and we can use that information to analyze the properties of estimator for the variance.

(Refer Slide Time: 16:59)

**Student's t-distribution**

$$T = \frac{X}{\sqrt{Y/n}}$$

$X \sim N(0,1)$

$Y \sim \chi^2$  - distribution with  $n$ -dofs

$X \perp Y$

$$\Rightarrow p_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}}; -\infty < t < \infty$$

Now, in estimating mean we assume the standard deviation of the population is known, but that may not be always the case; so, if you are going to estimate the, while estimating mean, you are going to estimate standard deviation of the population again from the same sample, so, then what happens is, the sampling distribution for the mean will not be Gaussian, it will have certain other property and that is known as student's t distribution;

So, what that is, if you take 2 random variables X and Y, where X is normally distributed zero mean and unit standard deviation, and Y is chi squares distributed with n degrees of freedom, and we form the ratio X divided by square root Y divided by n, this random variable, T, has this probability density function and this is known as student's t probability density function. Student was a pen name of a scientist was writing papers in the, with that name, and this distribution goes with that name.

So, here, you can see here this is, this is gamma function and n is on the right hand side, this t is the state variable and t takes values from minus infinity to plus infinity.

(Refer Slide Time: 18:24)

**Sampling distribution for the estimator of mean with variance not known**

Consider the estimator  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i$


$\Theta$  is an unbiased estimator of  $\mu$  with variance  $\frac{\sigma^2}{n}$ .

$\frac{\Theta - \mu}{s / \sqrt{n}} \sim$  Student t-distribution with  $n-1$  dofs.

$s$  = estimate of standard deviation from the sample.

$p_{\Theta}(\theta) = \frac{\Gamma[(f+1)/2]}{\sqrt{\pi f} \Gamma(f/2)} \left(1 + \frac{\theta^2}{f}\right)^{-\frac{1}{2}(f+1)} ; -\infty < \theta < \infty$

$f = \text{dof}$



Equipped with this description we can now talk about sampling distribution for the estimator of a mean with variance not known, so, that means the variance that is needed to construct the sampling distribution for the mean will be estimated from the same sample; to get point estimator for the mean you do not need the variance, but to write the sampling distribution and hence the confidence interval you need the variance.

Now, theta given by  $\frac{1}{n} \sum_{i=1}^n x_i$  is an unbiased estimator of  $\mu$  with variance  $\sigma^2/n$ . Now,  $s^2$ , which is the sampling, the sample variance as chi square distribution, so, I form now the sum, the ratio  $\theta - \mu$  divided by  $s / \sqrt{n}$ , and this is chi square, this is gaussian therefore, this will be a student's t distribution with  $n - 1$  degrees of freedom. Therefore, the

sampling distribution for estimator of a mean when variance of the population is not known is given by this.

(Refer Slide Time: 19:48)

$$P_{\Theta}(\theta) = \frac{\Gamma\left[\frac{(f+1)}{2}\right]}{\sqrt{\pi f} \Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{\theta^2}{f}\right)^{-\frac{1}{2}(f+1)} ; -\infty < \theta < \infty$$

Consider the statement

$$P\left(\theta_{\frac{\alpha}{2}, n-1} < \frac{\Theta - \mu}{s / \sqrt{n}} \leq \theta_{\frac{1-\alpha}{2}, n-1}\right) = 1 - \alpha$$

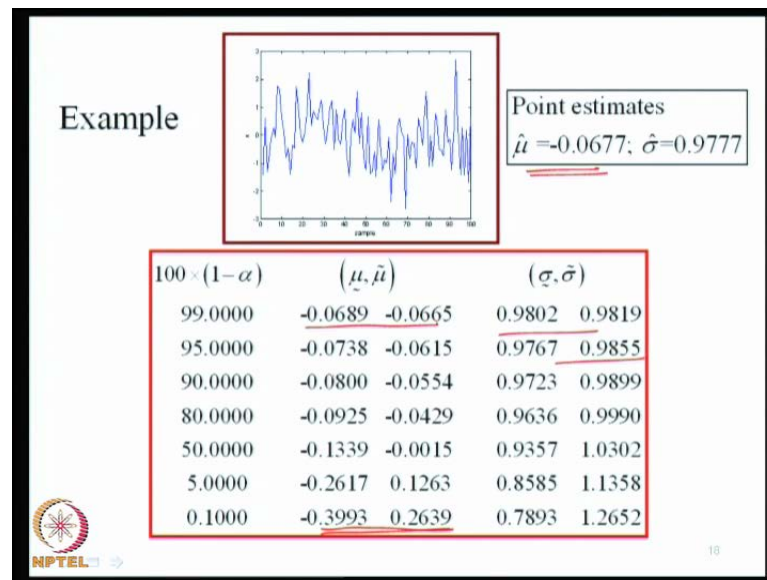
From this one can obtain the confidence interval for the population mean.

NPTEL

17

Now, if you want to construct confidence interval, from this you have to use this density function; so, earlier, if you recall in the derivation of confidence interval, we used this, this curve was Gaussian, now, this Gaussian density function has to be replaced by the student's t distribution while computing the confidence interval that would mean, if this is the sampling probability density function, if we make the statement theta Alpha by 2 n minus 1 theta minus mu divided by s divided by square root of n, that is, this random variable lying in this interval is one minus Alpha, where 1 minus Alpha is the confidence level, from this we can construct the confidence interval with given level of confidence.

(Refer Slide Time: 20:21)



So, the thing is, as I said instead of using Gaussian density function you need to use student's t distribution function. Now, an example: I have selected 100 samples of random numbers, and I am constructing, first, the point estimator for the mean is minus 0.9677 and sigma **hat** is 0.9777 for this sample. Now, here, in this table I have shown the confidence intervals for mean and standard deviation at different levels of confidence; so, with 99 percent confidence the interval is this and with 10 percent confidence the interval is this, and this is for standard deviation, this is for standard deviation and this is for mean.

(Refer Slide Time: 21:14)

Data used in the example  $n=100$

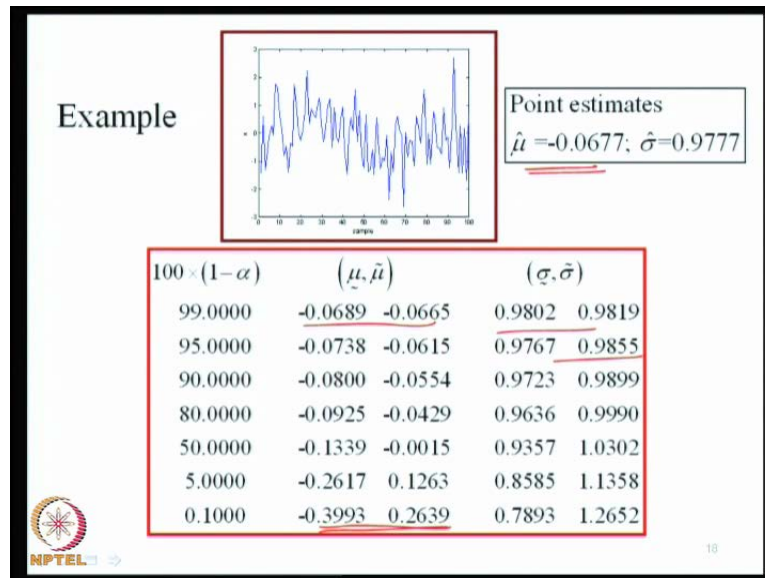
-1.4440	0.6123	-1.3235	-0.6616	-0.1461	0.2481	-0.0766	1.7382	1.6220
0.6264	0.0918	-0.8076	-0.4613	-1.4060	-0.3745	-0.4709	1.7513	0.7532
0.0650	-0.2928	0.0828	0.7662	2.2368	0.3269	0.8633	0.6794	0.5548
1.0016	1.2594	0.0442	-0.3141	0.2267	0.9967	1.2159	-0.5427	0.9122
-0.1721	-0.3360	0.5415	0.9321	-0.5703	-1.4986	-0.0503	0.5530	0.0835
1.5775	-0.3308	0.7952	-0.7848	-1.2631	0.6667	-1.3926	-1.3006	-0.6050
-1.4886	0.5585	-0.2774	-1.2937	-0.8884	-0.9865	-0.0716	-2.4146	-0.6943
-1.3914	0.3296	0.5985	0.1472	-0.1014	-2.6350	0.0281	-0.8763	-0.2655
-0.3276	-1.1582	0.5801	0.2398	-0.3509	0.8921	1.5783	-1.1082	-0.0259
-1.1106	0.7508	0.5002	-0.5173	-0.5592	-0.7534	0.9258	-0.2485	-0.1498
-1.2584	0.3126	2.6903	0.2897	-1.4228	0.2468	-1.4358	0.1486	-1.6931
				0.7192				

NPTEL 19

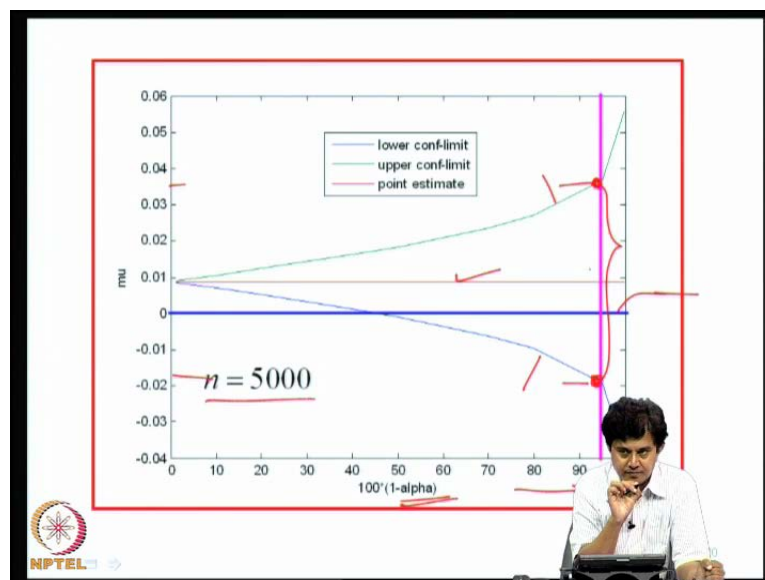


So, you can try to simulate this and see what information this conveys. Now, the data used in this example is provided here, so, you could actually replicate this table and I leave that as an exercise for you to do that.

(Refer Slide Time: 21:19)



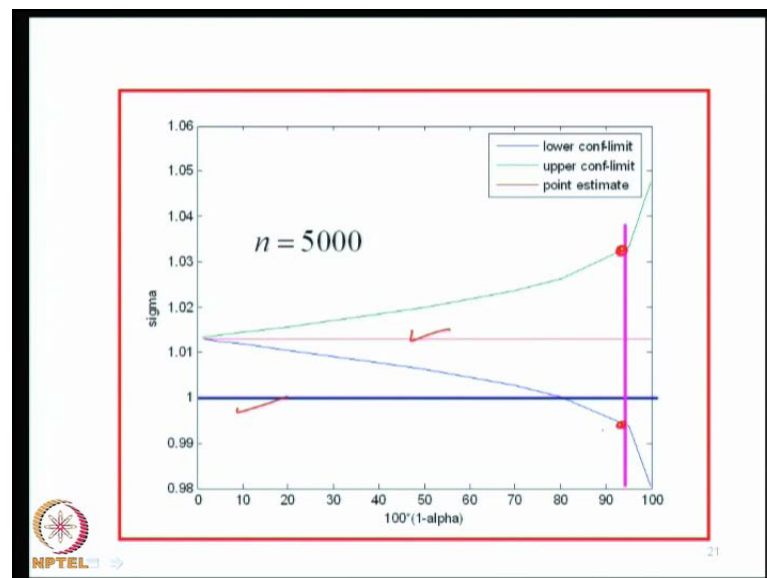
(Refer Slide Time: 21:24)



In this graph what I have shown is, I use five 5000 numbers and try to find the confidence intervals for different levels of Alpha, so, see, with 100 percent confidence I can say that the confidence interval is from minus infinity to plus infinity; so, as we go as our confidence level increases the width of the confidence band widens.

So, at 95 percent the confidence level interval is this, so, this is the upper limit of the confidence interval, this is a lower limit of the confidence interval; so, this is actually the population mean, and this 5000 numbers are generated synthetically with 0 mean and unit standard deviation, I will explain how that can be done in due course, but right now you can believe that there of 5000 numbers drawn from a population whose mean is zero and standard deviation is 1. So, we are getting in answer of a something less than 0.1 as point estimator, this red line is the point estimator and this blue line is actually the population mean, and this red line is an approximation to this blue line, that is one answer, the other answer is, you take any you take any value of confidence, so, 95 percent confidence I can say that this number here on the Y axis, this number here on the Y axis, this range contains the population mean, and I am able to make that statement with 95 percent confidence.

(Refer Slide Time: 23:06)



Now, this is the result on mean and a similar result on standard deviation; the standard deviation as I said is unity for the population, and I am getting an answer, which is between 1.01 to 1.02 as the point estimator; and the confidence bands, the lower confidence band value at 95 percent confidence level and the upper 1 is here, so, with 95 percent confidence I can say that the standard deviation population standard deviation is contained in this interval.

(Refer Slide Time: 23:40.

**Factors influencing confidence interval**

- The statistic used as the estimate
- The observations made
- Confidence level
- Sampling distribution
- Sample size

NPTEL

So, we can summarize that some of the factors that influence confidence interval are: the statistic that we are using as the estimate, the actual observations made, the confidence level that you prescribed and the sampling distribution for the statistic, and actually a sample size. They are some of the factors that influence the confidence interval.

(Refer Slide Time: 24:03)

**Number of samples needed for a given width of confidence interval**

Consider the estimator for population mean with known variance.

•  $\bar{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$  or,  $\frac{\bar{\Theta} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ .

•  $P\left[\theta + k_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \theta + k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$

Let  $w = k_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$  = half width of confidence interval be specified. Minimum number of samples required

$$\tilde{n} = \frac{1}{w^2} \left[ \sigma k_{(1-\alpha/2)} \right]^2$$

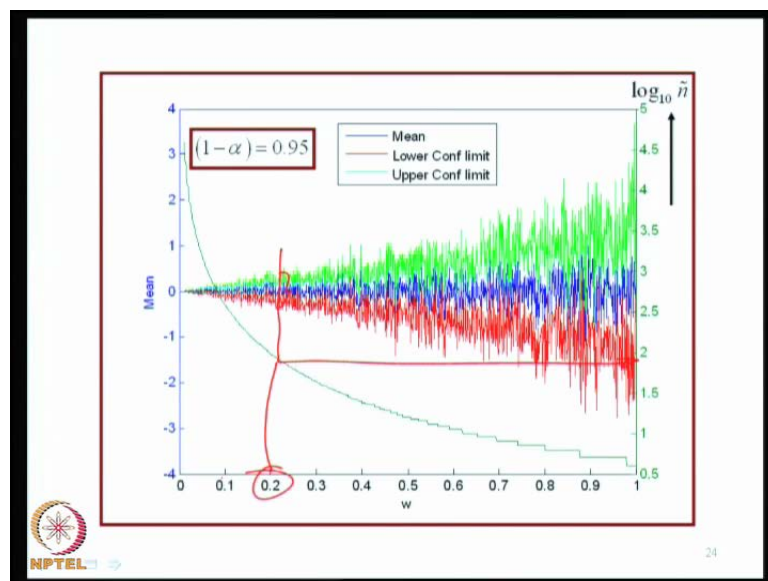
NPTEL 23

Now, the confidence interval is the function of the sample size. Now, we can ask the question, if I fix the width of the confidence interval, can I determine what is the number of samples needed. So, that problem can be addressed as shown here: the number of

samples needed for a given width of confidence interval, what is that? Now, consider the estimator for population mean with known variance, so that the sampling distribution is Gaussian, and this is our sampling distribution, mean  $\mu$  and standard deviation  $\sigma$  by square root of  $n$ , where  $n$  is the sample size, or this standard normal random variable,  $\theta - \mu$  divided by  $\sigma$  by square root  $n$ , which is 0 mean unit standard deviation normal random variable; and this is the confidence interval that, this is a statement that helps us to define the confidence interval.

Now I define  $w$  as  $k \cdot 1 - \alpha$  by  $2 \sigma$  by square root of  $n$ , as half width of the confidence interval to be specified that means, as a user I will say this must this is a width, that I want an estimate with this width, how many samples I should use. Now, minimum number of samples required, we can compute from this, you solve for  $n$  here, and I get this, where  $w$  is the width that you specified.

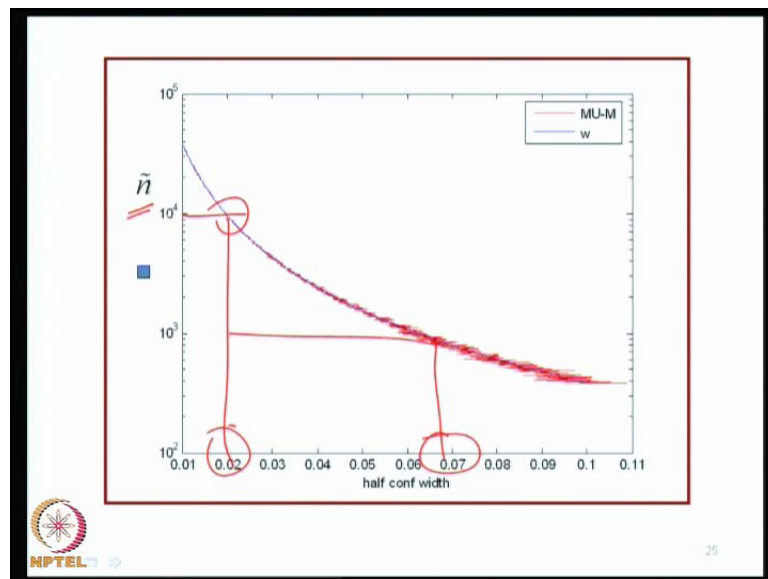
(Refer Slide Time: 25:31)



Now, so, what we have done here is, in this graph the Y axis's has to be read on left and the right, on the left it is the estimate of the mean and on the right this is the minimum number of samples that we have to use, and on the X axis I have  $w$ , the width of a confidence interval; so, that would mean, if the width of the confidence interval is point 2, so, I will go along this curve and determine that this much of samples I need to use, and that gives a- this is plotted on logarithmic scale- so, from this you will find out the number of samples needed to achieve this level of width.

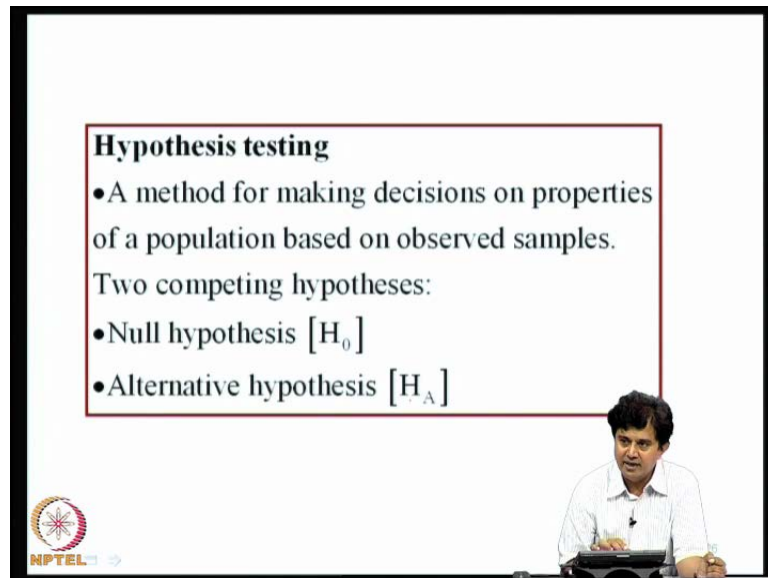
Now, if you actually performance these simulations, the blue line here is the mean point estimate, red is the lower limit of confidence interval and green is the upper limit of confidence interval; and if you exactly find the width here, that would be meeting the requirement that we are specified that means, with this number of samples this width will be this. So, that helps you to select sample size, if you want you, the narrower the confidence limit better is the answer, so you need more samples if you want narrower confidence intervals.

(Refer Slide Time: 26:48)



So, this is the plot of simply the minimum samples needed against the half confidence width. So, you want narrower confidence width, you need larger number of sample for example, with 10 to the power of 4 samples you are width will be 0.02, but if we are willing to only 100 samples your width will be somewhere 0.07. So, this is the much better solution than this, but you have to pay in terms of larger number of samples.

(Refer Slide Time: 27:24)



**Hypothesis testing**

- A method for making decisions on properties of a population based on observed samples.

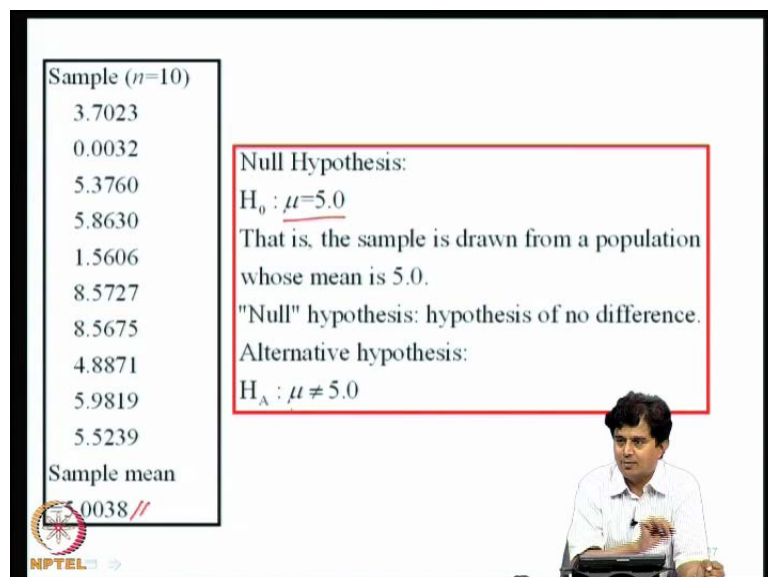
Two competing hypotheses:

- Null hypothesis [ $H_0$ ]
- Alternative hypothesis [ $H_A$ ]

NPTEL

Now, I move on to a next topic. What is known as hypothesis testing? Hypothesis testing is a method for making decisions on properties of a population based on observed samples. Typically, we postulate 2 competing hypothesis: one is what is known as null hypothesis denoted as  $H_0$ ; and other one known as alternative hypothesis denoted by  $H_a$ .

(Refer Slide Time: 27:56)



Sample ( $n=10$ )

3.7023
0.0032
5.3760
5.8630
1.5606
8.5727
8.5675
4.8871
5.9819
5.5239
Sample mean
5.0038 //

Null Hypothesis:  
 $H_0 : \mu = 5.0$   
That is, the sample is drawn from a population whose mean is 5.0.  
"Null" hypothesis: hypothesis of no difference.

Alternative hypothesis:  
 $H_A : \mu \neq 5.0$

NPTEL

Now, we can imagine a situation where, suppose there is a mass production of some product say, some steel rod or a yarn is being produced, and you are looking at say,

weight of the steel rod across say, one meter, and you have, you want that, that should be some prescribed number, suppose you want that the population mean should be 5 in some unit for some physical quantity.

Now, as the production is proceeding you draw 10 samples and find out it is sample mean; it is 5.0038, this is different from 5.0, why the difference occurs? There could be two reasons: one is there are inevitable random fluctuations, which are beyond our control, so, nothing can be done about that, and also I am finding this sample mean with only 10 samples so there is the fluctuations due to sampling fluctuations, I mean sample size, limited sample size and inherent randomness; but the other one could be, the production could be defective, that is could be something going wrong, that is what is produce producing this difference.

So, we are interested in knowing whether these 10 samples are actually being drawn from a population whose mean is 5.0 or not. So, we make the null hypothesis that  $\mu$  is 5.0 that means, the sample is drawn from a population whose mean is 5.0- the word null in null hypothesis means hypothesis of no difference; the alternative hypothesis here is the negation of this, no this 10 samples are do not drawn from a population whose mean is 5, the population from which this is being drawn, the mean of that is not 5, that is the alternative hypothesis.

(Refer Slide Time: 30:08)

Is the observed difference between estimate and the population mean due to sampling fluctuations (that is, random causes) or due to systematic (non-random) causes?

OR


Is the observed variation  $|\theta - \mu|$  arising due to some assignable causes or due to non-assignable causes?

OR

Is the observed variation  $|\theta - \mu|$  significant?

Significant: variation due to assignable causes.

If the difference is due to random causes, no action is needed. Otherwise, action is necessary.



28

Now, we test this hypothesis. The question that we are trying to answer is, the observed difference between estimate and the population mean is due to sampling fluctuations, that is, random causes or due to systematic non random causes, or is the observed variation,  $\theta - \mu$ , arising due to some assignable cause or due to non assignable causes. If it is due to an assignable cause, you need to take an action, may be you have to stop the production and examine what is go on, so, you have to take an action if it is assignable; if it is random fluctuations, you can continue with the production process. So, is observed variations significant? The word significant here means variation is due to assignable causes- something is indeed going wrong, there is something significantly wrong here I have to correct; if the difference is due to random causes, no action is needed, otherwise action is necessary. The decision that we need to make is accept or reject the null hypothesis.

(Refer Slide Time: 31:12)

**Decisions:** accept or reject the hypothesis

**Errors**

- Reject the hypothesis when it should have been accepted. [Type I error: error of commission]
- Accept the hypothesis when it should have been rejected. [Type II error: error of omission]
- Type I error: action when no action was needed.
- Type II error: inaction when action was needed.

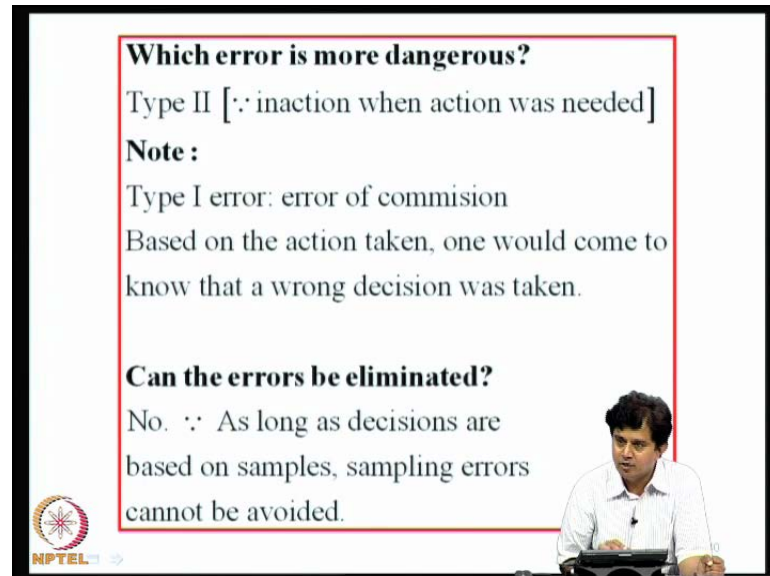
NPTEL 29

Now, the errors in making the decisions. Now, we may reject the hypothesis when it should have been accepted, this is known as Type 1 error, it is error of commission; the actual change, differences that you have you observe is due to random fluctuations, but you think it is due to systematic causes and you stop production, so you are making an error, that is known as Type 1 error. Now, accept the hypothesis when it should have been rejected, this is Type 2 error, this is error of omission; you should actually stop the production, but you think the difference that you are seeing is due to random fluctuations, so you permit productions to go ahead.



Now, Type 1 error is action when no action was needed and Type 2 error is inaction when action was needed.

(Refer Slide Time: 32:10)



**Which error is more dangerous?**  
Type II [∵ inaction when action was needed]

**Note :**  
Type I error: error of commission  
Based on the action taken, one would come to know that a wrong decision was taken.

**Can the errors be eliminated?**  
No. ∵ As long as decisions are based on samples, sampling errors cannot be avoided.

NPTEL


Now, which error is more dangerous? Actually, Type 2 errors are more dangerous because inaction when action was needed is more dangerous, if you take an action, you will come to know that the action was not needed, so, it is lesser evil.

So, Type 1 error is the error of commission, based on the action taken one would come to know what a wrong, that a wrong decision was taken, but of course, you have to pay the price because you stopped the production. The price that you would pay for pay for Type 2 error is that a faulty product would go into a final product, the faulty component will get into a final product.

Now, can these errors be eliminated? There is no way you can eliminate this as long as decisions are based on samples, sampling errors cannot be avoided; the only way you can avoid this is you have to measure everything that is produced, every meter of yarn or steel rod that you produce, you have to weigh and make sure it meets the criterion, then of course, there would not be any error, but you cannot be doing that. So therefore, there is no way that we can eliminate the error.

(Refer Slide Time: 33:17)

Null Hypothesis: $H_0 : \mu=5.0$		
Alternative hypothesis: $H_A : \mu \neq 5.0$		
Decision ↓	$H_0$ is true	$H_0$ is not true
Accept $H_0$	<u>Correct decision</u>	<u>Type II error</u>
Reject $H_0$	<u>Type I error</u>	<u>Correct decision</u>
P[Committing Type I error] = $\alpha$		
P[Committing Type II error] = $\beta$		
P[Accepting $H_0$ when $H_0$ is true] = $1 - \alpha$		
P[Accepting $H_A$ when $H_A$ is true] = $1 - \beta$		

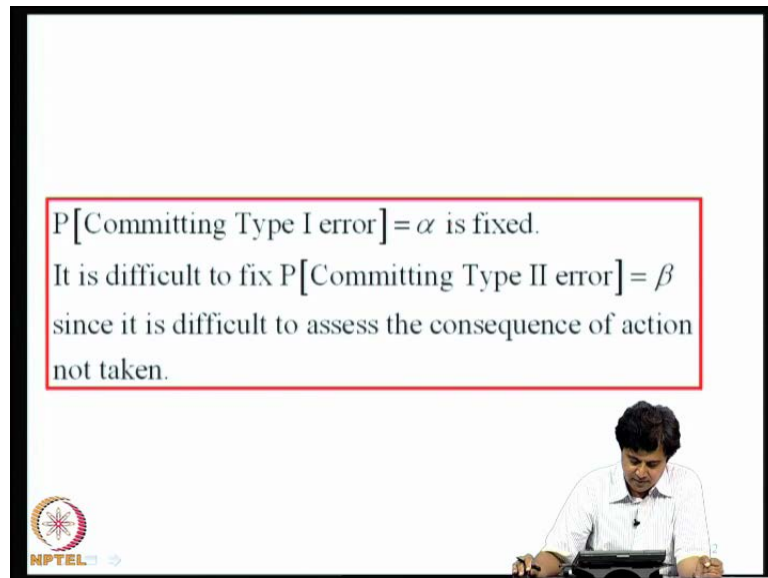


31



So, what we can do? Now, the null hypothesis is  $H_0$ ,  $\mu$  is 5 that means, sample is drawn for a population whose mean is five point zero; alternative hypothesis is  $\mu$  not equal to 5.0 that means, the sample is not drawn from a population, the sample is drawn for a population whose mean is not 5.0. Now, the decisions are:  $H_0$  is true,  $H_0$  is not true. You accept  $H_0$  when  $H_0$  is true, it is a correct decision; you accept  $H_0$  when  $H_0$  is not true, you are making Type 2 errors. When you reject  $H_0$  when  $H_0$  is true, you are making Type 1 error. You reject  $H_0$  when  $H_0$  is not true, it is a correct decision. So, there are two wrong decisions and two correct decisions.

So, what you do is, probability of committing Type 1 error, we call it as Alpha, and probability of committing Type 2 error, we call it as beta. Therefore, probability of accepting  $H_0$  when  $H_0$  is true is 1 minus Alpha and probability of accepting alternative hypothesis when, alternative hypothesis true is 1 minus beta.

(Refer Slide Time: 34:28)

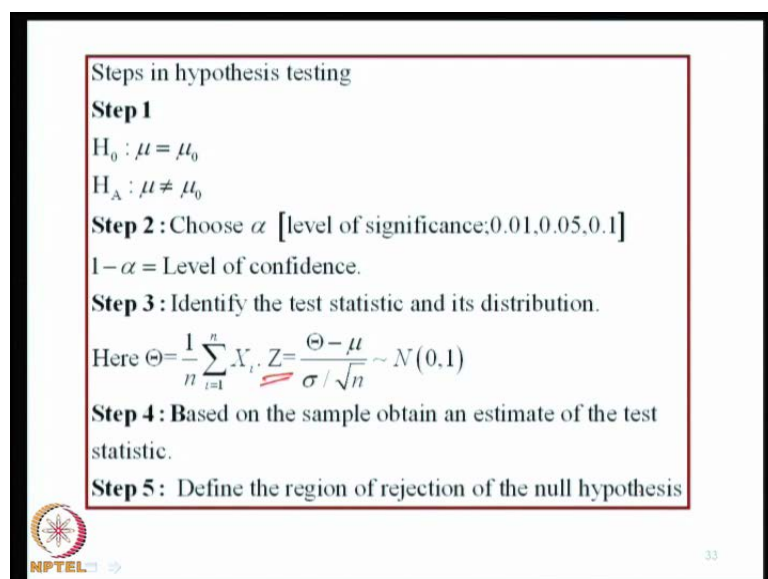


P[Committing Type I error] =  $\alpha$  is fixed.  
It is difficult to fix P[Committing Type II error] =  $\beta$   
since it is difficult to assess the consequence of action  
not taken.



Now, what we do is, probability of committing Type 1 error, Alpha, we fix that. It is difficult to fix probability of committing Type 2 error- that you have to assign a value for beta- since it is difficult to assess the consequence of action which has not taken. You can have experience on fixing Alpha, but not on beta. Ideally, you will try to minimize, but if you minimize Alpha, beta will increase, and if we minimize beta, Alpha will increase, I am not going to show that, but that is result.

(Refer Slide Time: 35:01)



Steps in hypothesis testing



**Step 1**  
 $H_0 : \mu = \mu_0$   
 $H_A : \mu \neq \mu_0$

**Step 2 :** Choose  $\alpha$  [level of significance: 0.01, 0.05, 0.1]  
 $1 - \alpha =$  Level of confidence.

**Step 3 :** Identify the test statistic and its distribution.  
Here  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

**Step 4 :** Based on the sample obtain an estimate of the test statistic.

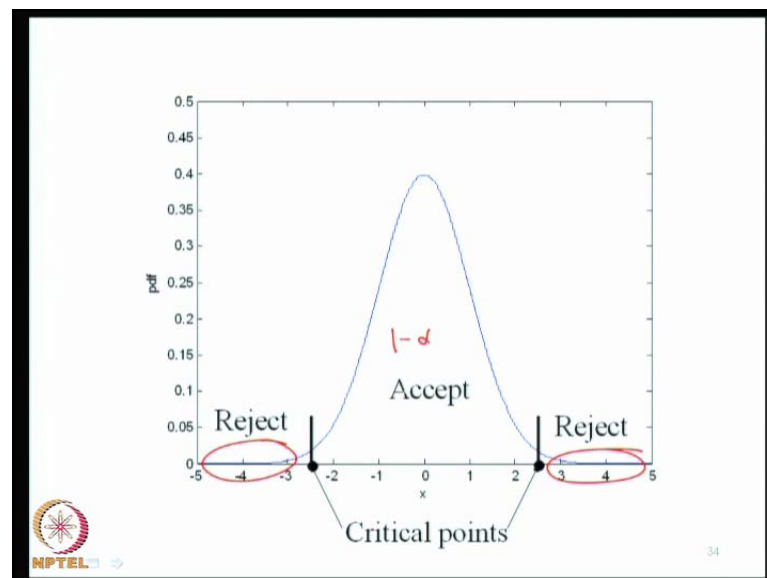
**Step 5 :** Define the region of rejection of the null hypothesis



So, equipped with this method we can now go through the steps in hypothesis test. Step one: we formulate the null and alternative hypothesis, here,  $\mu$  is equal to  $\mu_0$  and  $\mu$  is not equal to  $\mu_0$ . We choose Alpha, that is level, this is called level of significant 0.01, 0.05, 0.1, it is arbitrarily; actually, we have to make the choice, as a convention it is taken as 0.01 or 0.05 or 0.1, this is an a significant level. And one minus Alpha is known as confidence level. So, if you select 0.05 as significance level, you have 95 percent confidence in what you are saying.

Now, you have to identify the test statistic and its distribution. To test this hypothesis you need to define a statistic, so, obviously, we are now taking about mean, so the test statistic would be related to the estimate of the mean that is,  $\frac{1}{n} \sum_{i=1}^n X_i$  - this is our statistic. Now, the sampling distribution for this would be needed, so, we know that Z, which is  $\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$  is normal with zero mean and unit standard deviation; so, this Z will take it as the test statistic. Now, we have a sample therefore, we can find out realization of Z. Now, we define the region of rejection of the null hypothesis, how do we do that?

(Refer Slide Time: 36:29)



This is a probability density function of Z, and this area is 1 minus Alpha. If the observed statistic, that is the observed value of Z is in these region, we reject the null hypothesis otherwise you accept the null hypothesis. This is how we proceed.

(Refer Slide Time: 36:55)

$x =$

- 0.1867
- 0.7258
- 0.5883
- 2.1832
- 0.1364
- 0.1139
- 1.0668
- 0.0593
- 0.0956
- 0.8323
- 0.2944
- 1.3362
- 0.7143
- 1.6236
- 0.6918

**Step 1**  
 $H_0 : \mu = 0.0$   
 $H_A : \mu \neq 0.0$

**Step 2 :**  $\alpha = 0.05$

**Step 3 :**  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i ; Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

**Step 4 :**  $\theta = 0.1943 ; z = 0.7524$

**Step 5 :**  
 $\Phi^{-1}(0.025) = -1.96$   
 $\Phi^{-1}(1 - 0.025) = 1.96$   
 $-1.96 < z = 0.7524 < 1.96$   
Accept the null hypothesis at 5% significance level

NPTEL

So, again, let us take a set of 15 numbers drawn from a population whose standard deviation is 1, and let samples is 15. Now, I make a hypothesis that this sample is drawn from a population whose mean is zero, that is, my null hypothesis; the alternative hypothesis mu is not equal to zero.

I select Alpha to be 0.05. So, the estimator is  $\frac{1}{n} \sum_{i=1}^n X_i$  and the statistic is  $\frac{\theta - \mu}{\sigma / \sqrt{n}}$ , which is normal zero to 1. So, based on the these numbers, I get the point estimator for mean to the 0.1943, and substituting that into this and using  $n = 15$ , I get realization of Z to be 0.7524. Now, you look at the first critical point here, this one, that will be 5 inverse of 0.025, which is minus 1.96, and 5 inverse of the next critical point is 1 minus 0.025, which is 1.96.

(Refer Slide Time: 38:21)

$x =$

0.8617
0.1555
0.7128
0.2034
0.7749
0.7823
0.7970
0.8862
0.1036
5.1879
1.2127
3.0126
0.3665
0.4306
0.0172
15, $\sigma = 1$

**Step 1**  
 $H_0 : \mu = 0.0$   
 $H_A : \mu \neq 0.0$

**Step 2 :**  $\alpha = 0.05$

**Step 3 :**  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i ; Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

**Step 4 :**  $\theta = 1.0336 ; z = 4.0033$

**Step 5 :**  
 $\Phi^{-1}(0.025) = -1.96$   
 $\Phi^{-1}(1-0.025) = 1.96$   
 $z$  does not lie in the acceptance region  
 $-1.96 < z \leq 1.96$   
Reject the null hypothesis at 5% significance level

Now, this observed value of Z is indeed, contain in this interval 1.96 minus 1.96 to 1.96; so, Z lies in the region of acceptance, so we accept the null hypothesis at 5 percent significance level. Now, another example, I again take 15 numbers drawn from the population standard deviation is taken to be known, which is equal to one, and I go through this exercise again; the null hypothesis is mu is zero, alternative hypothesis mu not equal to zero, and I select Alpha to be 0.05, and again I get Z value to be 4.0033, this 4.0033 is not contain in my acceptance region. So, for this set of numbers I have to reject the null hypothesis at 5 percent significance level.

So, the sample mean is 1.03 whereas, the hypothesis that we are testing that it is zero, so what I have actually done here is, I have generated 15 random numbers from exponentially distributed random variables- I will come to that shortly- and that has mean 1; so, obviously, if I look at this data, it is clear that mean is not zero, but is this due to sampling fluctuations or is this due to systematic cause, that is what we are trying to discover. Now, for this same data, now, I make the null hypothesis that it is drawn from the population whose mean is 1, the earlier hypothesis it is drawn from a population whose mean is zero; now, I will test, since I know that I have generated the number for exponentially distributed random variable whose mean is one, I can now test whether this simulation of this random number is correct or not.

(Refer Slide Time: 39:29)

$\bar{x} =$

0.8617
0.1555
0.7128
0.2034
0.7749
0.7823
0.7970
0.8862
0.1036
5.1879
1.2127
3.0126
0.3665
0.4306
0.0172
15, $\sigma = 1$

**Step 1**  
 $H_0 : \mu = 1.0$   
 $H_A : \mu \neq 0.0$

**Step 2 :**  $\alpha = 0.05$

**Step 3 :**  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i; Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

**Step 4 :**  $\theta = 1.0336; z = 0.1301$

**Step 5 :**  
 $\Phi^{-1}(0.025) = -1.96$   
 $\Phi^{-1}(1 - 0.025) = 1.96$   
 $z$  does lie in the acceptance region  
 $-1.96 < z \leq 1.96$   
Accept the null hypothesis at 5% sig

NPTEL 37

So, null hypothesis is, mu is 1, alternative hypothesis mu not equal to zero, and again significance level is 0.05, and I get the Z statistic to be 0.130 and that lies in the acceptance region, and now I can accept the null hypothesis that, this sample is drawn from a population whose mean is 1 and at 5 percent significance.

((Refer Slide Time: 40:28))

**Step 1**  
 $H_0 : \mu = 0.0$   
 $H_A : \mu \neq 0.0$

**Step 2 :**  $\alpha = 0.05$

**Step 3 :**  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i; Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

**Step 4 :**  $\theta = 0.0123; z = 0.8732$

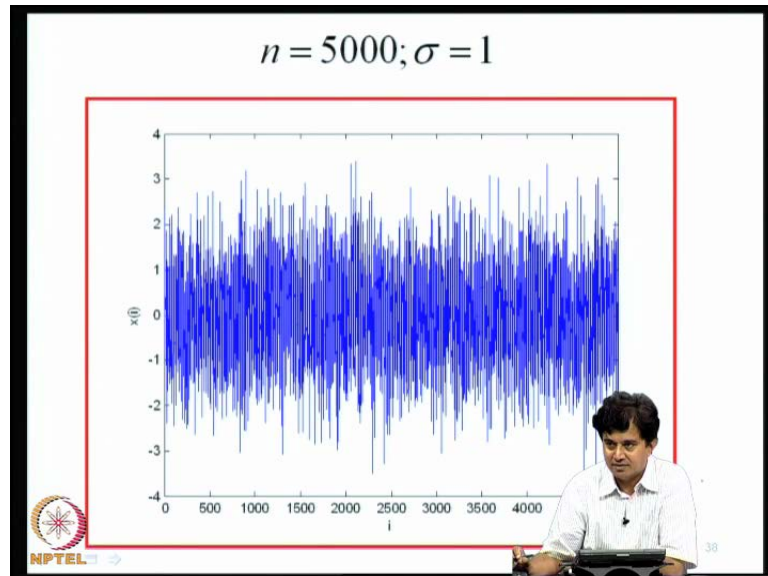
**Step 5 :**  
 $\Phi^{-1}(0.025) = -1.96$   
 $\Phi^{-1}(1 - 0.025) = 1.96$   
 $z$  does lie in the acceptance region  
 $-1.96 < z \leq 1.96$   
Accept the null hypothesis at 5% signific

NPTEL 39

Now, if samples, I take 5000 samples, so this again an exercise, this is hypothesis, null hypothesis this drawn from a population of mean is zero, and I get the sample statistics

of 0.8732, and it leads to the conclusion that we can accept the null hypothesis at 5 percent significance level.

(Refer Slide Time: 40:44)



Similarly, 5000 numbers drawn from exponential random variable.

(Refer Slide Time: 40:53)

**Step 1**  
 $H_0 : \mu = 1.0$   
 $H_A : \mu \neq 0.0$

**Step 2 :**  $\alpha = 0.05$

**Step 3 :**  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i; Z = \frac{\Theta - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

**Step 4 :**  $\theta = 0.990; z = -0.7040$

**Step 5 :**  
 $\Phi^{-1}(0.025) = -1.96$   
 $\Phi^{-1}(1 - 0.025) = 1.96$   
 $z$  does lie in the acceptance region  
 $-1.96 < z \leq 1.96$   
 Accept the null hypothesis at 5% significance level

41

This null hypothesis is drawn from a population with mean one, and here again variable to show that Z is 0.7040 and it passes the acceptance criteria. Now, what is to be noted here is that, the population here is not Gaussian, I know that I am drawing it for my exponential population whose basic distribution is exponential random variable, but still



by a virtual center limit theorem, because the sampling distribution even for that population we are assuming that it is Gaussian, so, this result kind of illustrates how center limit theorem leads to what seems to be a correct answer.

(Refer Slide Time: 41:40)

**Population standard deviation not known**

**Step 1**  
 $H_0 : \mu = \mu_0$   
 $H_A : \mu \neq \mu_0$


**Step 2 :** Choose  $\alpha$  [level of significance; 0.01, 0.05, 0.1]  
 $1 - \alpha =$  Level of confidence.

**Step 3 :** Identify the test statistic and its distribution.  
 Here  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i; S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .

$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim$  Student's t-distribution with  $n$  dof.

**Step 4 :** Based on the sample obtain an estimate of the test statistic.

**Step 5 :** Define the region of rejection of the null hypothesis


42

Now, if population standard deviation is not known, how do you do hypothesis test? Now, I make the null hypothesis that  $\mu$  is  $\mu_0$  and alternative hypothesis  $\mu$  is not equal to  $\mu_0$ . Now, again I choose significance level and level of confidence. Now, we have to identify the test statistic, now, and its distribution. Earlier, I knew, I assumed that variance is known therefore, sampling distribution was Gaussian, now variance is also estimated from the same samples therefore, the test statistic will be now related to student's t distribution; I define capital t as  $\bar{X} - \mu_0$  divide by square root  $n$  where this  $s$  is the sample variance.

(Refer Slide Time: 42:50)

**Step 1**  
 $H_0 : \sigma^2 = 100$   
 $H_A : \sigma^2 > 100$

**Step 2 :** Choose  $\alpha$  [level of significance; 0.01, 0.05, 0.1]  
 $1 - \alpha =$  Level of confidence. ✓

**Step 3 :** Identify the test statistic and its distribution.  
Here  $\Theta = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \Theta)^2$ .

$C = \frac{(n-1) S^2 \Theta - \mu}{\sigma^2} \sim \chi^2$  with  $(n-1)$  dof.

**Step 4 :** Based on the sample obtain an estimate of the test statistic.

**Step 5 :** Define the region of rejection of the null hypothesis.

NPTEL

Now, based on the sample obtain the estimate of the test statistic; now, I have to get value of t for that from observed value of theta and s, and we have to define the region of rejection of the null hypothesis, again based on the t distribution. This can be done, it is tedious, but it can be done. Now, we can similarly, now, we can argue of the logic or the steps in constructing hypothesis test procedures for other statistics for example, I make the null hypothesis that the sample is drawn from a population whose variance is 100, the alternative hypothesis is variance is greater than hundred, how do we test it ?

So, it is same story, we select the confidence level, significance level and confidence level, now you have to identify the test statistics; the test statistic depends on what you are testing as hypothesis, it is now one variance, so, I take the variance estimator of variance to be one by n minus one X i theta whole square where this theta is the unbiased estimator for the mean, this is also unbiased estimator for a variance. Now, the test statistic that we will select is related to the sampling distribution of variance and I have shown that this is the chi square random variable, and I define this to be n minus one whole square s square theta minus mu divided by sigma square, so, this is chi square with n degrees of freedom. Based on this, now, I calculate the sample realization of the test statistic and identify the region of acceptance and rejection, and I can then decide whether I should accept hypothesis or reject.

(Refer Slide Time: 44:17)

**Probability papers**

Let  $X$  be a random variable with PDF  $P_X(x)$ .

Let  $\{x_i\}_{i=1}^n$  be a sample of  $X$ .

Probability paper is a special plotting device in which y-axis is scaled in such a way that the PDF function appears as a straight line.

Example

$P_X(x) = 1 - \exp(-\lambda x)$   $x \geq 0$

$1 - P_X(x) = G_X(x) = \exp(-\lambda x)$

$\log G_X(x) = -\lambda x$

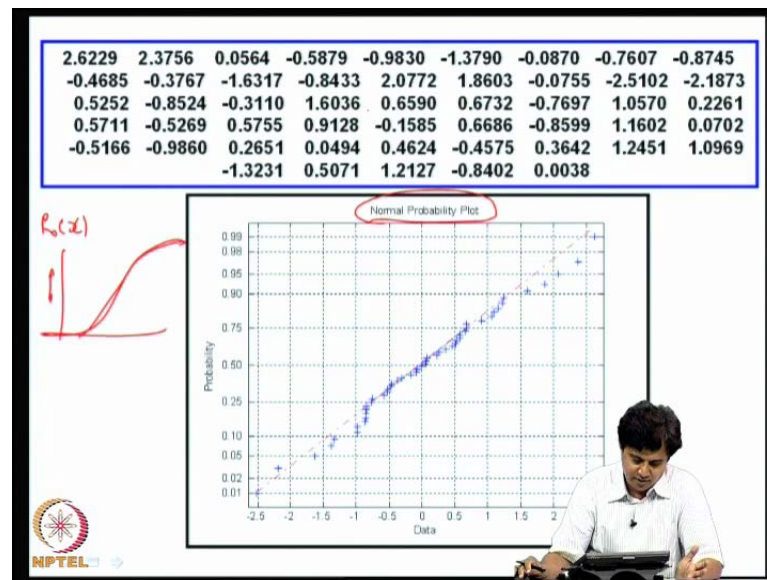
The complement of the cumulative PDF appears as a straight line.

NPTEL

Now, till now, I have been talking about moments, now how about probability distributions? So, if I want to see whether a sample is drawn from a population whose probability distribution is Gaussian or not with given mean and variance, how do we test that? So, we need to now think of modeling probability distributions and there are certain tools available for that, one of that is that is probability paper. So, let  $X$  be a random variable with PDF  $P_X$  of  $x$  and this **lower case X i** be a sample of  $X$ . This probability paper is a special plotting device in which y axis is scaled in such a way that the probability distribution function appears as a straight line, we distort the y axis in such a way that the probability distribution function becomes a straight line, probability distribution function or it is complement for example, if I take an exponential random variable,  $P_X$  of  $x$  is one minus exponential minus lambda  $X$  where  $X$  takes values from zero to infinity.

Now, one minus  $P_X$  of  $x$ , I call it as  $g_X$  of  $x$ , is a complementary probability distribution function is exponential minus lambda  $X$ . Now, you take logarithm of this I get  $\log$  of  $g_X$  of  $X$  is minus lambda  $X$ , so, if I now distort the  $Y$  axis, instead of plotting  $P_X$  of  $x$  if I plot  $\log g_X$  of  $x$ , I will get a straight line for the probability distribution function; So, on this paper if I now plot the observations that I am making, those points will lie on this line if the numbers are drawn from an exponential random variable.

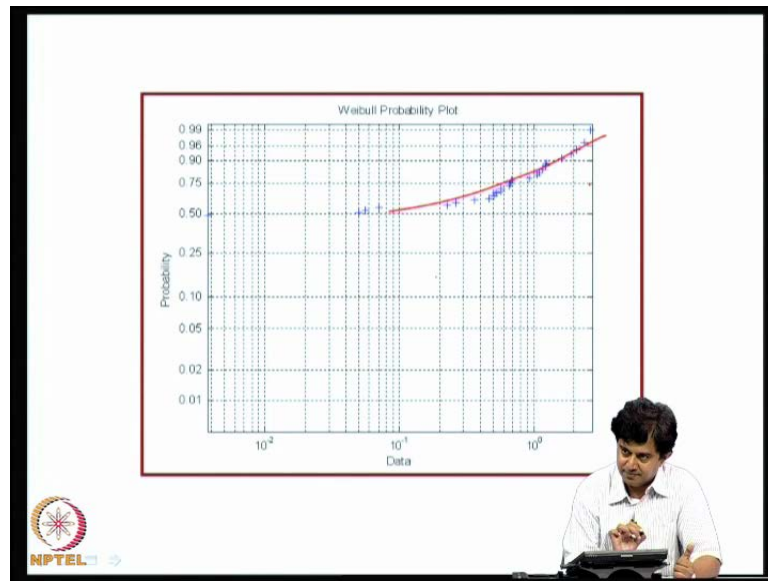
(Refer Slide Time: 45:56)



So, listed that, let us consider, I think this is fifty numbers from population whose mean is zero, standard deviation is one, and this is the probability paper for normal probability distribution function. The probability paper, for every probability distribution there will be one paper; the normal probability paper is meant for Gaussian random variables. Now, this red line is a straight line, straight line, this red straight line is actually the theoretical probability distribution function, and this distorted the y axis, the probability distribution function appears as a straight line. If you plot the Y axis in an arithmetic scale, this is what we have been doing, you get this familiar curve like this, but now we have adjusted the Y axis so that this curve appears is the straight line.

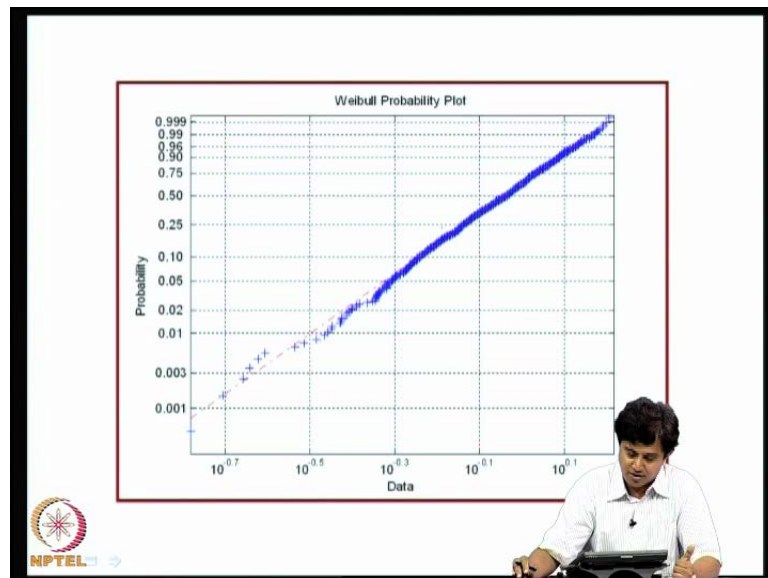
Now, if I plot these numbers on this, they are following this straight line. So, this is a simple device to see whether the numbers are Gaussian or not.

(Refer Slide Time: 47:13)



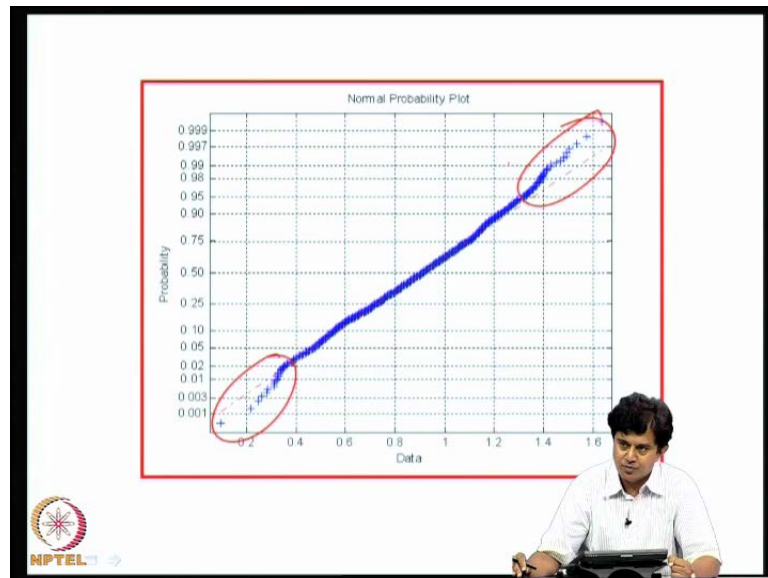
Now, the same data if I plot on another probability paper, that is probability paper corresponding to another random variable for example, if I plot the same numbers on Weibull probability paper, it is not a straight line, see I am getting something like this, which it is not a straight line, so, this clearly says the numbers are not Weibull.

(Refer Slide Time: 47:29)



Now, if I indeed simulate numbers which are according to Weibull distribution, this what has been done here, they appear along the straight line. These numbers if you plot on normal distribution paper, it will be appearing distorted on that.

(Refer Slide Time: 47:45)



So that is what is shown here, the Weibull numbers are shown on normal probability paper, you can see that there is a distortion with two ends. So, probability paper is a useful tool in modeling for a quick assessment on a nature of probability distribution function.

(Refer Slide Time: 48:00)

**Kolmogov - Smirnov test**

**Step 1**  
 $H_0$  :  $X$  has a specified distribution  $P_X(x)$ .  
 $H_A$  : The PDF of  $X$  is other than what is specified.

**Step 2** : Choose  $\alpha$  [level of significance: 0.01, 0.05, 0.1]  
 $1 - \alpha$  = Level of confidence.

**Step 3** : Define  $X^{(i)} = i^{\text{th}}$  largest value in observed sample.  
 $P^*(X^{(i)}) = \frac{i}{n}$  //  $\alpha = X^{(i)}$

$D_2 = \max_{i=1}^n \left[ P^*(X^{(i)}) - P_X(x) \right]$

**Step 4** : Based on the sample obtain an estimate of the test statistic.

**Step 5** : Define the region of rejection of the null hypothesis  
 Accept  $H_0$  if  $D_2 \leq c$ .

Now, can we formulate hypothesis testing procedures to verify the null hypothesis? For example,  $H_0$ , the null hypothesis  $X$  as specified distribution  $P_X$  of  $x$  that means, the population has this prescribed probability distribution function. You have a sample now,

based on properties of sample you have to test this hypothesis. The null hypothesis is PDF of  $x$  is other than what is specified. This  $P_x$  of  $x$  need to be completely specified for example, this has to be, population can be Gaussian, but with different parameters, so, your null hypothesis should be for example,  $X$  as a normal probability distribution function with mean equal to say, 1, standard deviation equal to 0.5; if it is drawn from a population whose mean is 20 and standard deviation is 30, then you cannot accept the null hypothesis. It is not on the Gaussian nature that we are testing the hypothesis, we are specifically testing for a given distribution with all the parameters as a given specific values.

We choose Alpha, again 0.01,0.05,0.1, and the test statistic here is, we, this, we plot the empirical probability distribution function, for that what we do we rank order all the observations and plot the  $X$ , we define the probability distribution  $i$  of  $n$  at that value  $X$  of  $i$ ; then we define a statistic known as  $D_2$ , which is the maximum of the difference between the observed empirical probability distribution function and the corresponding theoretical probability distribution function, I think this has to be evaluated at  $X$  of  $i$ . Based on the sample obtain an estimate of the test statistic. Now, again we have to define the region of rejection of the null hypothesis, accept  $H_0$  if  $D_2$  is less than or equal to  $c$ . To be able to do that we need the sampling distribution of the test statistic; in this test known as kolmogoy smirnov test, this sampling distribution is tabulated and we can refer to that and conduct this test.

(Refer Slide Time: 50:23)

**Kolmogov - Smirnov test : Example**

**Step 1**  
 $H_0$  :  $X$  has a specified distribution  $N(0,1)$ .  
 $H_A$  : The PDF of  $X$  is other than what is specified.

**Step 2** : Choose  $\alpha = 0.05$


**Step 3** : Define  $X^{(i)} = i^{\text{th}}$  largest value in observed sample.

$$P^*(X^{(i)}) = \frac{i}{n}$$

$$D_2 = \max_{i=1}^n \left[ P^*(X^{(i)}) - P_X(x) \right]$$

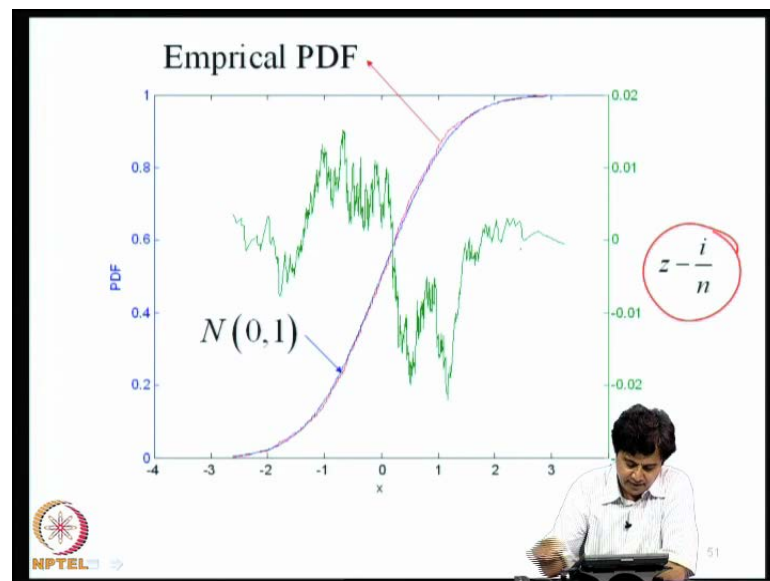
**Step 4** : Based on the sample obtain an estimate of the test statistic.

**Step 5** : Define the region of rejection of the null hypothesis  
Accept  $H_0$  if  $D_2 \leq c$ .


50

So, I will give some examples. The null hypothesis,  $H_0$ , is a specified distribution  $N(0, 1)$  that means, Gaussian is zero mean and  $(1)$  standard deviation; the alternate hypothesis  $H_1$  is other than what is specified. We select 5 percent significance level and go through this exercise, and I have done this plotting here.

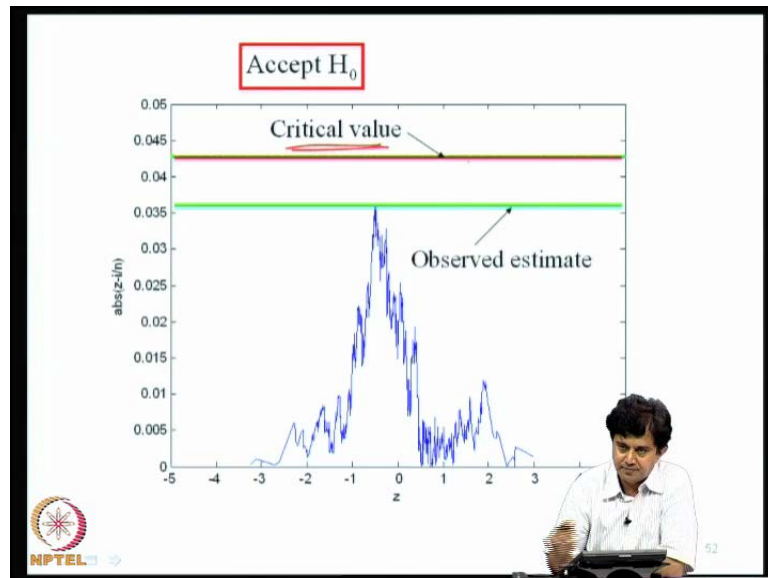
(Refer Slide Time: 50:46)



So, you see here, the blue line is the theory, theoretical population probability distribution function, which is normal zero one; this red one is the probability distribution function constructed from the data; this green line, which we are seeing here, which is to be read in conjunction with axis on Y axis on the right hand side, is the difference between the  $\bar{x}$ ,  $Z$  minus  $\mu$  by  $\sigma / \sqrt{n}$ .

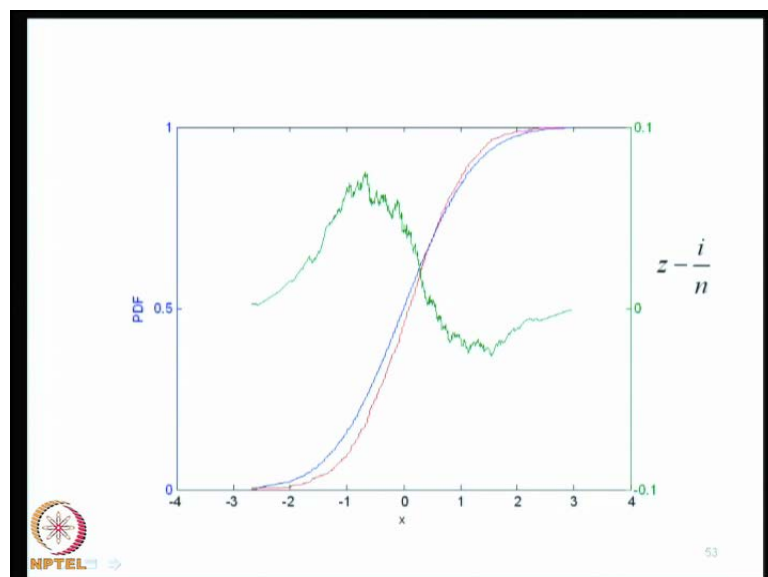


(Refer Slide Time: 51:16)



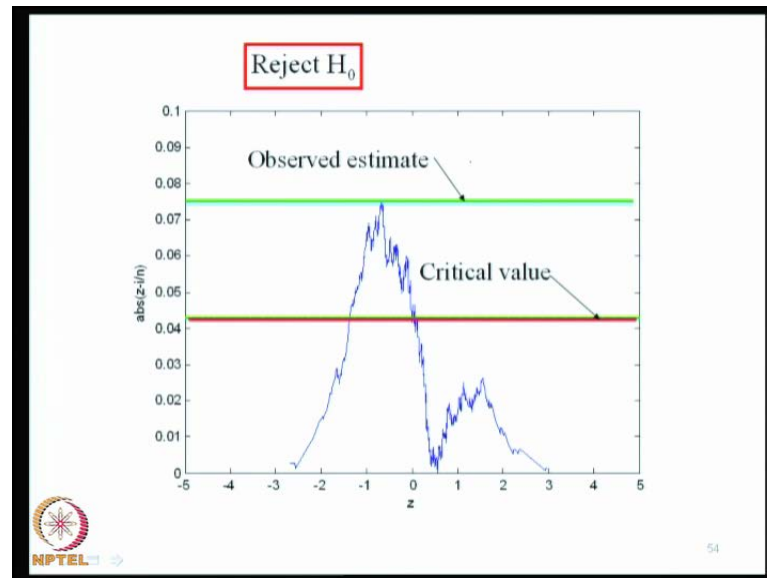
Now, if you take the absolute value of that, we can re plot that, and the maximum difference is the observed estimate of statistic that I am looking for, and for this gain sample size I can find out what is the critical value. So, the observed estimate of the observed statistic is less than the critical value therefore, we can accept the null hypothesis.

(Refer Slide Time: 51:40)



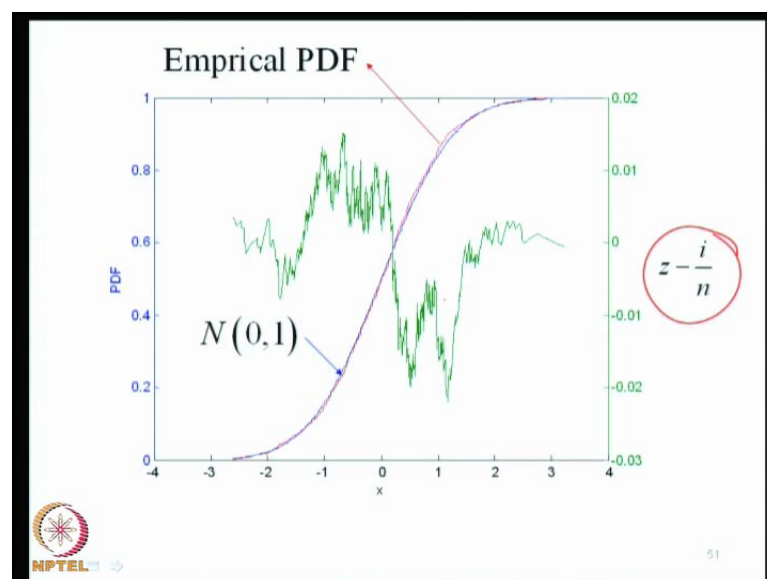
This is another case, same calculations.

(Refer Slide Time: 51:48)

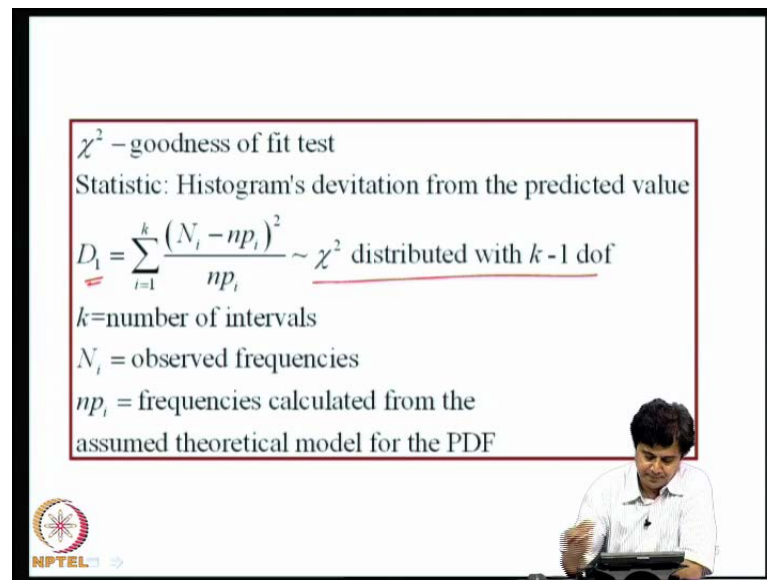


But here, the critical value is here, the observed estimate is here, and we need to reject the null hypothesis that  $(())$  the sample is drawn from a population whose mean is 0 and standard deviation is 1. So, what I basically did was, the same random numbers that I used in this study, I added artificially some mean and multiplied the sample, I mean standard deviation I enhanced artificially by some number so as to make the numbers, I mean they are still Gaussian, but not having the mean and standard deviation that is being proposed in the null hypothesis. So, the null hypothesis has to be rejected.

(Refer Slide Time: 52:05)



(Refer Slide Time: 52:26)



$\chi^2$  – goodness of fit test  
Statistic: Histogram's deviation from the predicted value

$$D_1 = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \sim \chi^2 \text{ distributed with } k-1 \text{ dof}$$

$k$  = number of intervals  
 $N_i$  = observed frequencies  
 $np_i$  = frequencies calculated from the assumed theoretical model for the PDF

There is another test known as chi square test, which is again helpful in verifying the probability nature of probability density function. Here, the statistic is defined in terms of deviation, histogram's deviation not on distribution, but on histogram's. So, we make  $k$  number of bins and find out how many points in your sample lying in each bin, and according to theory how many points you expect in in each of this bins; you can estimate knowing the sample size, and we can form this statistic, and we can show that this  $D_1$  has a chi square distribution with  $k$  minus 1 degrees of freedom where  $k$  is an number of intervals in making histogram's,  $N_i$  are observed frequencies,  $np_i$  are frequencies calculated from the assumed theoretical model for the PDF. So, there are this chi squared distributions, again well tabulated, and for given level of Alpha you can always find the critical value and therefore, you can develop the procedure to accept or reject the null hypothesis.

(Refer Slide Time: 53:31)

**Digital simulation of samples of random variables**

Let  $X$  be a random variable with PDF  $P_X(x)$ .

How to generate samples  $\{x_i\}_{i=1}^n$  of  $X$  on a computer so that the estimated model for PDF of  $X$  from the data  $\{x_i\}_{i=1}^n$  matches with the target PDF  $P_X(x)$ ?

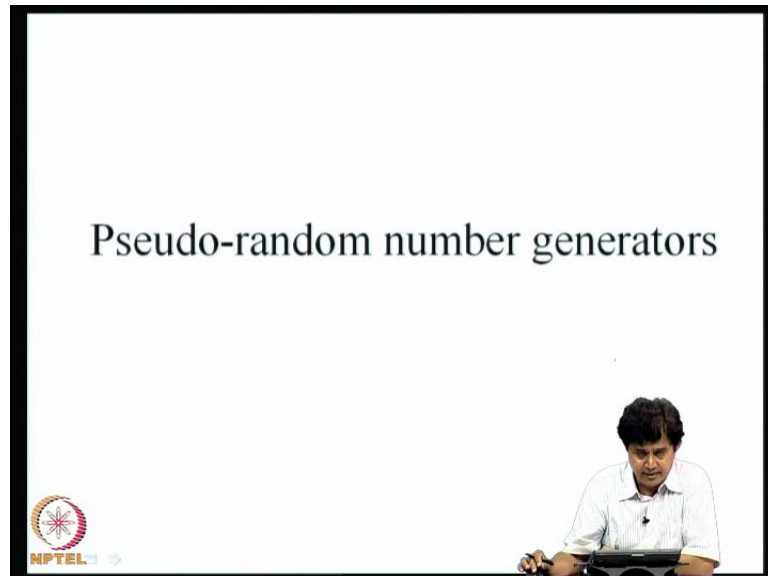
The slide features a red-bordered box containing the text above. Below the box, a man in a white shirt is seated at a desk, looking at a laptop. In the bottom left corner, there is a circular logo with a star and the text 'NPTEL'.

Now, what I have done till now is I have quickly reviewed the main results in theory, mathematical theory of statistics, and now we need to move on to the problem of digital simulation of samples of random variables. Our basic aim is to be able to stimulate samples of random quantities, which can be random variable, random processes, random processes evolving in time, random processes evolving in space; like a wind load on a chimney it evolves both in space and time, so, that has to be done; and ocean wave if you take, in space it is multidimensional, and there is a time; so, and at given point in ocean you may look at the displacement of the wave or some other field variable like pressure or something; so, you have a vector random field evolving in multiple parameters.

So, we need to now, we have now learnt how to completely characterize, specify such stochastic quantities, the question that we are now is asking how to simulate numerically samples of realizations of those random quantities. Suppose  $X$  is log normal, how do I generate 100 samples from the given probability distribution function with the parameter specified, how I can generate hundred number whose, if I were to empirically estimate the probability distribution function, you should be able to accept the hypothesis that this 100 numbers are drawn from a population of log normal random variables, which prescribed mean is standard deviation, so how do you achieve that? So, the question is let  $X$  be a random variable with a prescribed probability distribution function, and the question we are asking is how to generate samples of  $X$  on a computer so that estimated

models for probability distribution of  $X$  from the data matches with the target PDF, how do we do that?

(Refer Slide Time: 55:29)



This is a question that we will consider now, and the starting point for that is what are known as pseudo random number generators. That means, on a computer, how can we simulate random numbers?

So, I will begin discussion on this in the next lecture, and this will be the starting point to construct samples of random variables, random process evolving in time, random process evolving space and time, vector random processes, Gaussian random process and Non Gaussian random processes; so, the mathematics of simulation for each one of this differs from each other and we will see some of these details in the next lecture. So, we conclude this lecture at this stage.