

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

Lecture – 01
Design of Experiments

Hello, welcome to the introductory lectures on Design of Experiments.

(Refer Slide Time: 00:33)



First, we look at the relevant books that are useful for this topic. The first book is by Montgomery and Runger, *Applied Statistics and Probability for Engineers*; Fifth Edition, New Delhi, John Wiley India, 2011. This is a basic book which starts from Statistics, and then goes on to Probability, and also later on into Design of Experiments and Regression. The next book is by Montgomery, which deals with the Design and Analysis of Experiments, the various possible designs, and their method of implementation, and analysis.

The third book is a slightly more advanced one; it requires a background in linear algebra, and with sufficient background in this particular linear algebra you can work through most of the contents in the text book, and discover the fascinating aspects of design of experiments. Then, you also have the book written by Ogunnaike; this is a new book, and some of the complex topics are dealt in a nice manner; there are some interesting examples also in this book. For those of you who want to get started, the first

book is recommended, and once you develop interest in the subject, and you want to really implement the design of experiments concept in your experimental work, then you can start looking at the second book. And those of you who want to give a further theoretical background to your results and explore into more advanced design of experiments options can choose the book by Myers et al.

So, why do we do experiments? Experiments are necessary in order to prove the theory, and also, when sometimes theory is not developed or it's very difficult to develop the theory for a particular process or application, then experiments are necessary. Experiments are necessary also as proof of concept. So we all like to do experiments and we enjoy doing experiments, but sometimes we get frustrated with the results.

But design of experiments help us to overcome these frustrations and present the results in a scientifically acceptable manner. So once you have done the experiments, you have a certain amount of data with you, and you will be studying about representation of the experimental data. You will be looking at Histograms, Box Charts, Scatter Plots, Normal Probability Plots and so on. So, when you do experiments most of the time you do repeats just to convince yourself that the results are pretty much the same or more or less the same during each repeat, and then, you present the data in an average form. You take the average of all the responses from your experiments or output from your experiments. Next step is to quantify the scatter in the experimental results.

(Refer Slide Time: 03:55)



Key Features in Experimentation

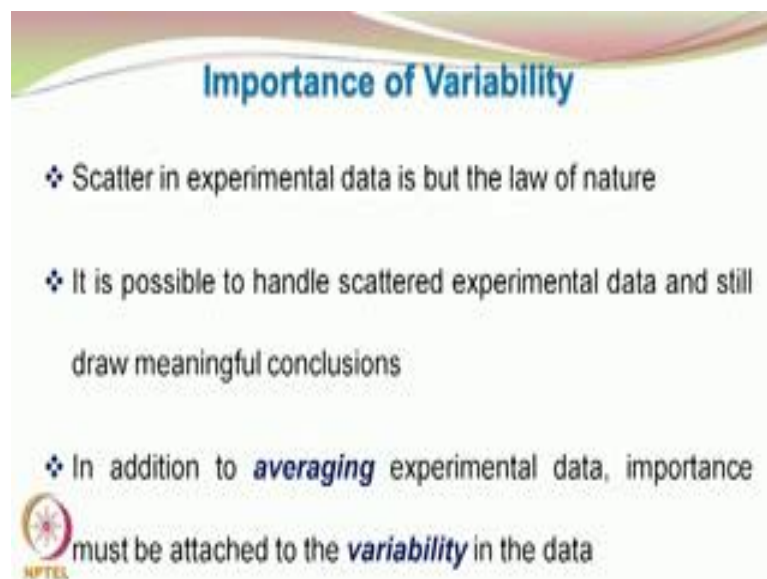
- ❖ Data representation
- ❖ Averaging of the experimental results
- ❖ Quantifying scatter in the experimental runs
- ❖ Identifying important variables that affect the response of the experiment
- ❖ Where to do the experiments next?

NPTEL

It is very unlikely that when you repeat the experiment you will get the identical response or the identical output, there is going to be some amount of deviation between the responses. So, it is very important for you to quantify the scatter in the experimental data in a suitable way. Once you have done these, it's very important to see which variables in the process are actually influencing your experiment. If a particular variable is not having that much of sensitivity, probably you can fix that variable at a certain level, and then, look at the other variables for their effects.

If a variable is not having an important or a significant effect on the outcome of the process, then it may be kept fixed, and you also reduce the number of experiments you are planning to do further. So once you have done a basic set of experiments called as Screening Experiments, you want to know where exactly you should do the next set of experiments so that you get the maximum yield or minimum power.

(Refer Slide Time: 05:33)



Importance of Variability

- ❖ Scatter in experimental data is but the law of nature
- ❖ It is possible to handle scattered experimental data and still draw meaningful conclusions
- ❖ In addition to *averaging* experimental data, importance must be attached to the *variability* in the data

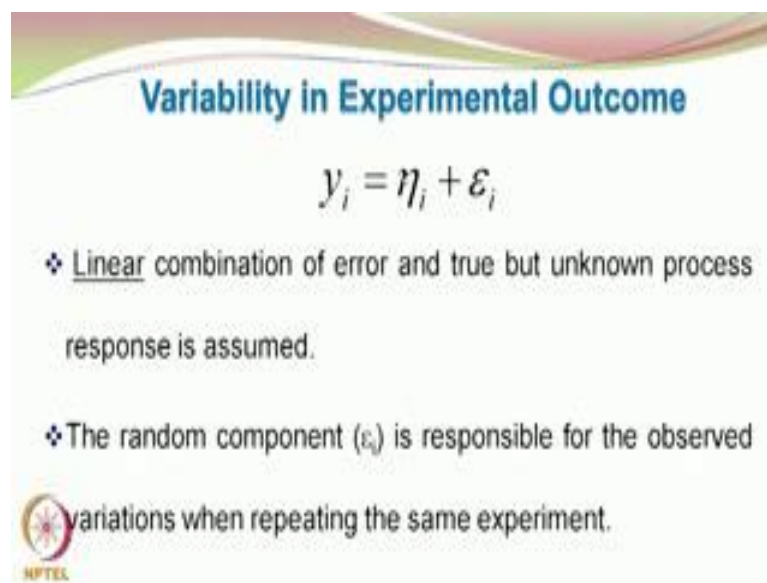
NPTEL

So, depending upon your objective, you would like to explore the experimental design space, and see where the next set of experiments are going to quickly lead you to the optimum set of conditions. As I said earlier, the variability in the experimental data has to be accounted for. The scatter in the experimental data is inevitable, because there are a lot of random effects which are influencing your experiment and you cannot control all of them. So you don't have to get frustrated when you are unable to identically get the response when you repeat the experiments, and if there is scatter in your data, sometimes

you may feel – ok, my data is not good and I am doing something wrong in the experimentation, so you may want to stop.

But **it's** very important to note that you have to live with this scatter experimental data and still draw meaningful conclusions. **So** in addition to averaging the experimental data importance must be also attached to the variability in the data. The design of experiments you will see how to account for the random variability, and also the systematic variability caused by changing the variables, and compare the differences between the two.


(Refer Slide Time: 06:36).



Variability in Experimental Outcome

$$y_i = \eta_i + \epsilon_i$$

- ❖ Linear combination of error and true but unknown process response is assumed.
- ❖ The random component (ϵ_i) is responsible for the observed variations when repeating the same experiment.

 NPTEL

Let us take a simple case where you have the response for the i th run y_i that is equal to $\eta_i + \epsilon_i$. For a moment imagine that ϵ_i is 0, what is ϵ_i ? ϵ_i is the random error component. If it is 0, then when you repeat the experiment, and assuming that all the components in your experimental setup are working perfectly, you will get the same response. But, because of this random error component you are having different responses when you repeat the experiments. Thus, ϵ_i may be either positive or negative. So, your response may be varying on the positive side and also on the negative side.


What I am trying to say is, suppose in the first case you are getting 100, and the next time you will repeat the experiment you may get a 101, in the third time you may get 97. So, the variability can be on either side **that's** because ϵ_i may be either positive or negative. η_i is the true value, it is the absolute value, but unfortunately, we do not

know that, and so we need to estimate what the true value may be by doing experiments, and also living with the associated variability.

(Refer Slide Time: 08:06)

Random Variable

- ❖ Denoted by X and once its value is known after the experiment, it is termed as x .
- ❖ Is the genesis for probability distribution functions
- ❖ Probability distributions are used to predict the behavior of a group of randomly behaving entities

 NPTEL

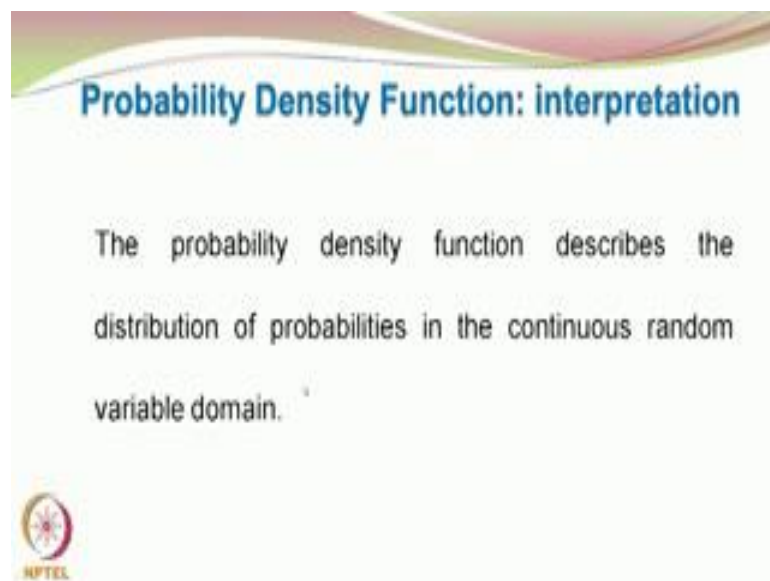
So before we jump into the design of experiments, it is really worthwhile to invest some time in understanding the basic concepts. Only then we will be able to appreciate the design of experiments concepts and understand what the results are actually telling us. Without doing this background material or without knowing this back-ground material you may not be able to understand what is meant by P value or confidence intervals or level of significance, mean square error, and things like that. So, let us spend a bit of time in looking at the basics.

First concept is the random variable. It is a kind of an abstract entity; the random variable is denoted by capital X and once its value is known after the experiment it is labeled as small x . **So** this is the nomenclature and this is the genesis for the probability distribution functions. You might have come across probability distribution functions earlier, so this is very, very important, and this gives the distribution of the random variable. And Probability distributions are used to predict the behavior of a group of randomly behaving entities. Let us take **a** class, and if we want to predict the performance in the class of a single individual, then it is very difficult, and if you assume that all the students in the class are independent of one another, each student is going to be difficult to predict beforehand on how his performance is going to be.

But an experienced teacher will know that when you have a collection of students, then their behavior is easier to predict, their collective behavior is easier to predict than the individual behavior; because, most of the collection of individuals form on or fall under one distribution or the other. So, the instructor may be able to say, okay in this class may be about 10 percent of the students will do very well in the exams, may be 5 percent of the class will do pretty badly.

The average level of the performance in the class is likely to be 60 percent, and more often than not you will find that his predictions are valid because he has handled large number of data sets. So, the random variable is the originator for these probability distributions. So we will be looking at continuously varying random variables, there may also be examples of discrete random variables, but in our design of experiments we deal with usually continuously varying quantities.

(Refer Slide Time: 11:17)



So the probability density function describes the distribution of probabilities in the continuous random variable domain.


(Refer Slide Time: 11:25)

Definition

The probability distribution function $f(x)$ for continuous random variables is defined as

$$P(X \leq z) = \int_{-\infty}^z f(x) dx = F(z)$$

$F(z)$ is the cumulative distribution function.



So what is the definition for the probability distribution function? It is defined as probability of the random variable taking the value less than or equal to small z . Small z is any particular value; it can be 0.5, 1, 3 or even infinity, is equal to minus infinity to z , integral of f of x dx , and that is denoted by f of z , capital F of z . So, what you are doing here is the upper limit you plug in the value of z and then you integrate the probability distribution function. **So** when you do that, you get a particular value if you know the form of the function f of x ; either you can do this integration analytically or if the form of the probability distribution functions is quite complex, then you may want to do it numerically.


Any way, you will get a value and that value is representing the cumulative probabilities from minus infinity to z and that is given as the cumulative distribution function. What will happen when z goes to plus infinity? By definition the total area under the curve is representing the sum of the probabilities which we know is equal to 1. So, the cumulative distribution function is referring to the sum of the probabilities. If it is plus infinity, it is the overall sum of the probabilities or if it is value less than plus infinity, it will be the sum of the probabilities up to that value z . We are not doing direct addition, but we are doing integration.

(Refer Slide Time: 13:12).

Features of Probability Density Functions

What is the probability of x lying between values 'a' and 'b'?

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$
$$= \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = F(b) - F(a)$$



So, people who are familiar with integral calculus can very easily figure out this particular slide. Probability of a less than or equal to X less than or equal to b ; that means, the probability that the random variable is lying between a and b . And that is given by integral of a to b f of x dx which may be shown to be the cumulative distribution function at b minus cumulative distribution function at a .

(Refer Slide Time: 13:46)


Mean and Variance of Probability Density Function

The mean for the continuous distribution is given by

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

The variance of the continuous probability distribution is

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$



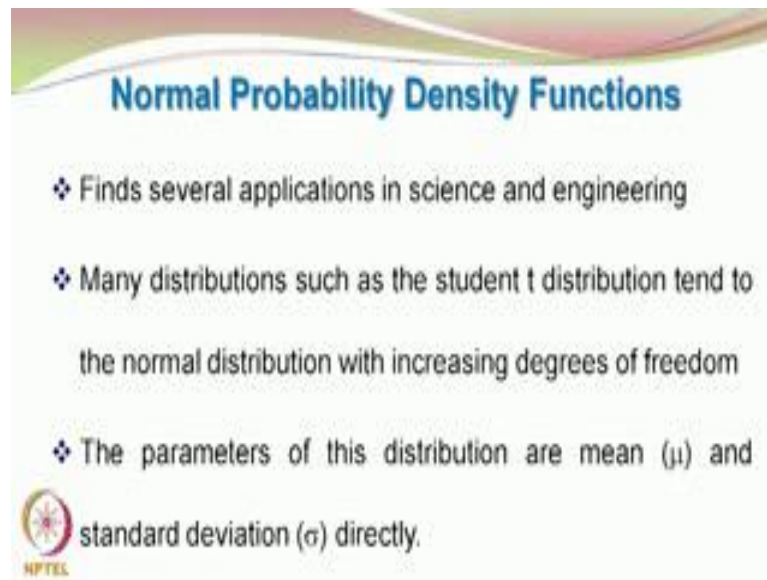
Now, whenever you have a probability distribution function you can imagine that there is a distribution of the probabilities. So, when you have a distribution you may want to

know where the center of the distribution is; you may also want to know the extent of spread of the distribution. So you would like to know the mean and you would like to know the variance. Mean is the representative of the average value of the distributions. Sometimes, depending upon the shape of the distribution, the mean may be at the geometric center of the distribution. Sometimes, in the case of skewed distributions, the mean may not be in the center. What I am trying to say by center is suppose the random variable x is going from minus infinity to plus infinity, in some symmetric distributions the mean may be at 0, but in some other cases it may be at minus 5 or plus 10 and so on depending upon the nature of the curve.

So, before we get too ahead of ourselves let us have some definitions out of the way. What is μ ? μ is the mean of the continuous probability distribution function. The mean for the continuous distribution is given by μ is equal to expected value of X . So, what is the expected value of the random variable X and that is nothing but the μ or the mean of the distribution. That is given by $\int_{-\infty}^{+\infty} x f(x) dx$; this need not be equal to 1 because we are multiplying the probability distribution function by x and so it will not be always equal to 1, only $\int_{-\infty}^{+\infty} f(x) dx$ will be equal to 1, but here it can be any value. It can even be a negative value, because if the distribution is predominantly on the negative side of the x -axis, then you will have a negative mean.


To quantify the spread of the probability distribution function, we have sigma squared representing the variance; so sigma squared is the variance and that is defined as expected value of the X minus μ whole squared okay. So we are finding the deviation of the random variable from the mean, and then we are squaring it, and then we get sigma squared. So, we have $\int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$.

(Refer Slide Time: 16:37).



Normal Probability Density Functions

- ❖ Finds several applications in science and engineering
- ❖ Many distributions such as the student t distribution tend to the normal distribution with increasing degrees of freedom
- ❖ The parameters of this distribution are mean (μ) and standard deviation (σ) directly.

 NPTEL

So, now, we will be looking at one of the most important and popular Probability Density Function. This is the normal distribution; you might have come across this normal distribution several times in the past. It finds applications in science and engineering and many of the other distributions also tend to the normal distribution under certain conditions. For example, the T distribution tends to the normal distribution when the degrees of freedom tends to infinity. What is meant by degrees of freedom we will see a bit later. And as I said earlier, whenever we look at a probability distribution function we are interested in knowing its average and also the spread.

In the case of the normal distribution, the parameters of the distribution themselves are mean and standard deviation. The standard deviation is nothing but the square root of the variance. There may be other distributions where the parameters may be a and b, for example, and you have to manipulate these parameters mathematically to get the mean and standard deviation. But the advantage in normal distribution is that the μ and the standard deviation σ are themselves the parameters of the distribution.


(Refer Slide Time: 18:05)

DEFINITION

A random variable X with the following p.d.f. is a normal random variable with parameters μ ($-\infty < \mu < \infty$) and σ ($0 < \sigma < \infty$)

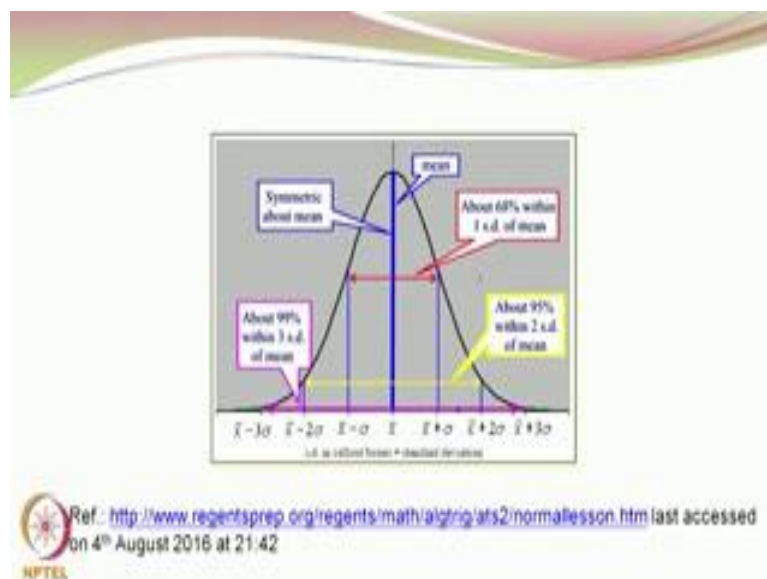
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ with } -\infty < x < +\infty$$

The notation $N(\mu, \sigma^2)$ is used to represent the distribution.



So what do you mean by parameters? I will show in the next slide. So, this is the form of the probability distribution function and that is given by $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. **So** this is a mathematical form for f of x and the parameters in this distribution are μ and σ . So, you can get different types of the distribution by changing the μ and σ . That is why these are called as parameters.

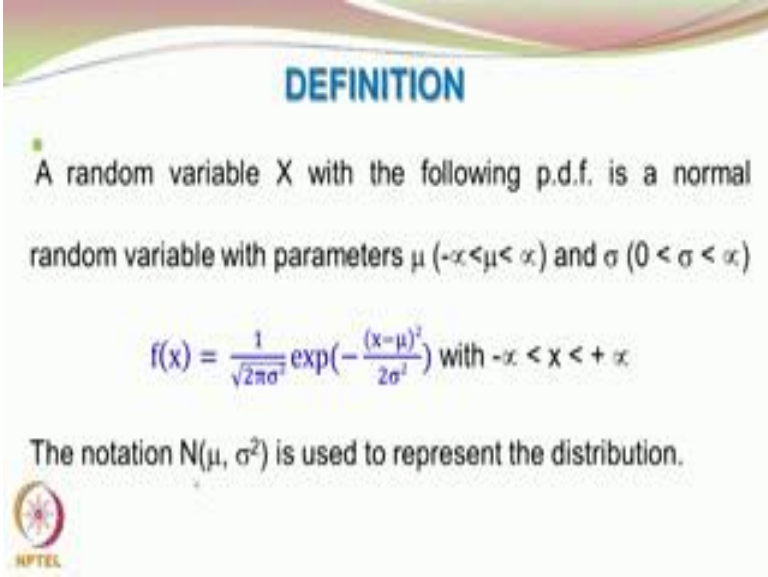
(Refer Slide Time: 18:40)



So how does a normal distribution look like? This is the shape of the normal distribution. As you can see, this is nicely symmetrical and it is centered at \bar{x} , \bar{x} is also the mean μ . Why it is called \bar{x} ? Again, we have to wait a bit; that represents the sample mean, but normally we put μ here. So, you can see that the area of the curve would be 50 percent below the mean and 50 percent above the mean. So, that the total area under the curve is 100 percent, and if you look at it from a probability point of view, it would be 1. So, you are going from minus infinity to plus infinity, but the tail becomes very thin when you go further and further away from the mean value.

So you can say that 68 percent of the area is within one standard deviation from the mean. **So** suppose this is the mean value, **so** you have one standard deviation on either side, so $\bar{x} + \sigma$ and $\bar{x} - \sigma$, where \bar{x} is the mean of the distribution. We can call it as μ also, but here it is shown as \bar{x} . And **so** 68 percent of the area is tucked between these two limits, and if you go to $\bar{x} - 2\sigma$, that means a two standard deviations in the mean value are below the mean value, and then, you go two standard deviations ahead of the mean value, then you pack something like about 95 percent of the total area **okay**. **So** when you go three standard deviations from the mean on either side, then you are packed something like 99 percent of the area.

(Refer Slide Time: 20:49).




DEFINITION

A random variable X with the following p.d.f. is a normal random variable with parameters μ ($-\infty < \mu < \infty$) and σ ($0 < \sigma < \infty$)

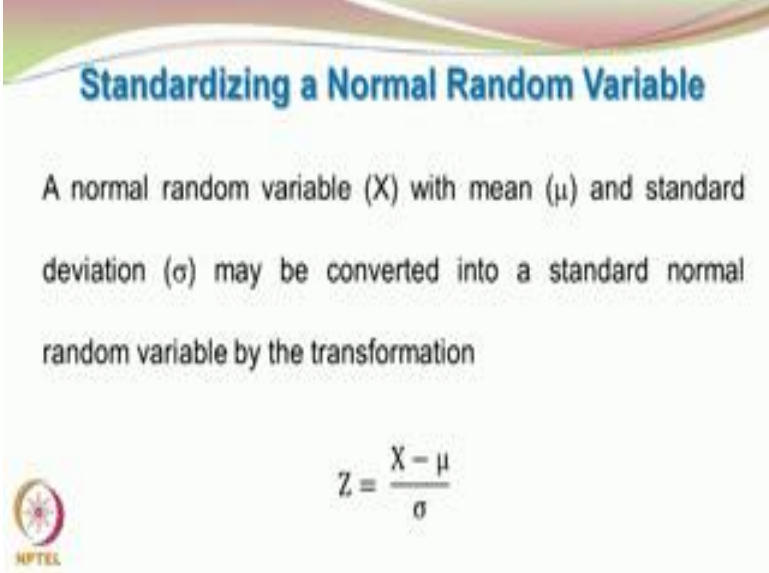
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ with } -\infty < x < +\infty$$

The notation $N(\mu, \sigma^2)$ is used to represent the distribution.



The notation for the normal distribution is capital N mu comma sigma squared, where mu is the mean and sigma squared is the variance. This is used to represent the distribution.


(Refer Slide Time: 21:03)



Standardizing a Normal Random Variable

A normal random variable (X) with mean (μ) and standard deviation (σ) may be converted into a standard normal random variable by the transformation

$$Z = \frac{X - \mu}{\sigma}$$



Now, depending upon the value of mu and sigma squared you can have infinite number of distributions. Sometimes, the mean value may be negative, the distribution may be centered at a negative value, but the variance which is sigma squared is always obviously positive and so is the standard deviation - square root of sigma squared which we call as sigma or the standard deviation is also positive. The mean may be negative, but the standard deviation is always positive. **So** we want to create a single standard normal distribution irrespective of what the mean mu and the standard deviation sigma are. **So** suppose you have any arbitrary normal distribution with any arbitrary value of mu and arbitrary value of a sigma, then you can standardize it in a very simple way.

So, when you redefine the random variable x by subtracting it with the mean mu and dividing this difference by the standard deviation sigma, we get a standard normal random variable called as a capital Z, and then, we have x minus mu by sigma is equal to e Z. When you make this transformation, the normal distribution gets magically transformed, and it is centered at a 0, and has a unit standard deviation or the variances equal to 1. **And** once you transform any normal distribution in to the standard form, and you want to look at the probabilities, all you need to do is to look at a single table


containing the probability distributions of the normal distribution with mean 0 and standard deviation 1.

(Refer Slide Time: 22:57).

Standardizing a Normal Random Variable

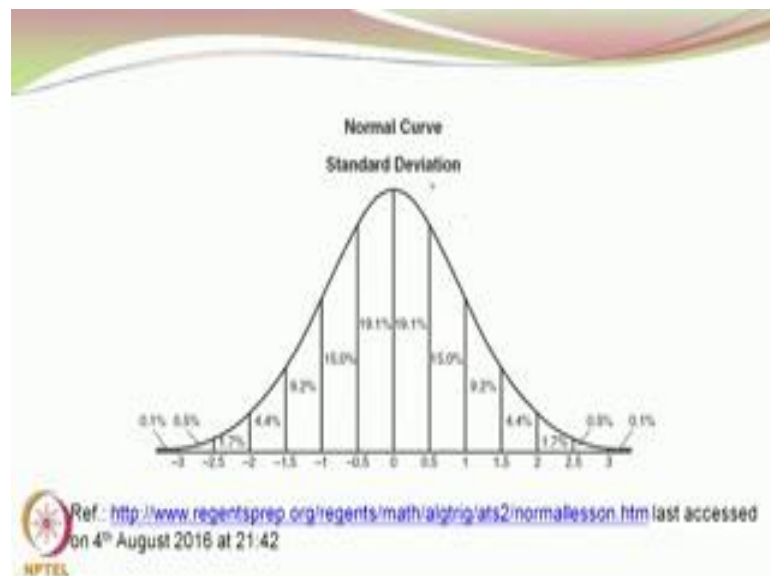
Hence this transformed random variable X has zero mean and variance 1 and hence is a standard normal random variable.

The cumulative distribution of a standard normal random variable is denoted as $\Phi(z) = P(Z \leq z)$



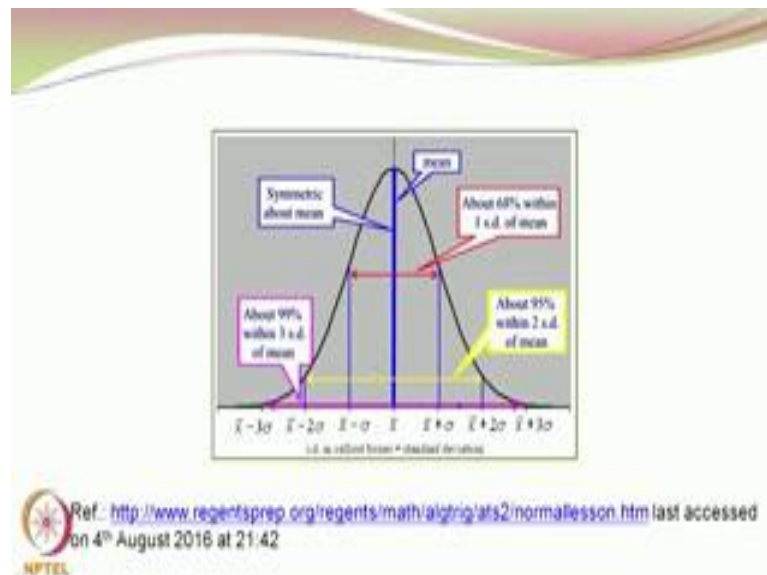
So as I said earlier, once you standardize a normal random variable, it has 0 mean and variance 1; so this is called as a Standard Normal Random Variable. The cumulative distribution of a standard normal random variable is defined as $\Phi(z)$ is equal to the probability of capital Z less than or equal to z ; remember that capital Z represents the random variable z and z is any numerical value.

(Refer Slide Time: 23:28).



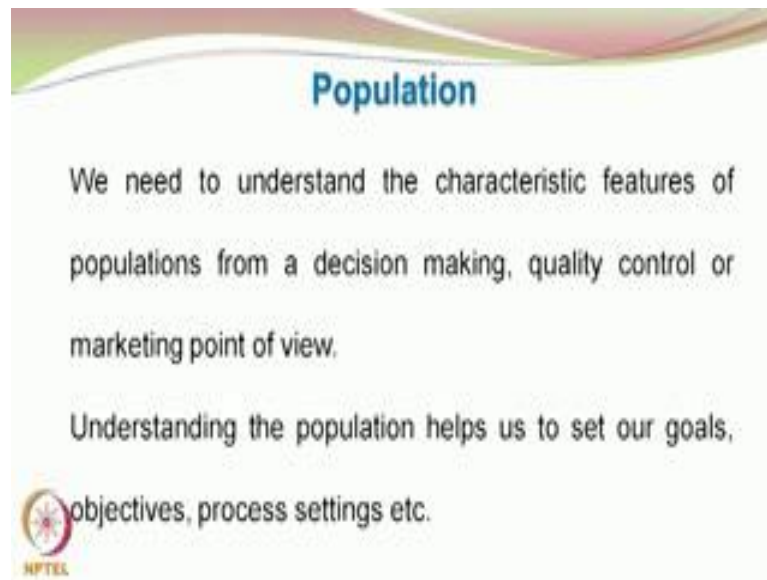
So, when you look at the standard normal curve, you can see that it is centered at mean 0, this minus 0.5 lies 0.5 times the standard deviation value from the mean. **And** obviously, the standard deviation value for the standard normal will be equal to 1. So, this represents half a standard deviation below the mean and this 0.5 represents half a standard deviation above the mean; the mean being 0. **And** you can see that the area under the curve corresponding to random variables lying one standard deviation among either side of the mean comes to 15 plus 19.1 is 34.1; and so, again, you have 34.1 here, that is 68.2 percent. So, 68.2 percent of the area under the curve is covered within one standard deviation on either side of the mean and **that's** what we saw in the previous slide.

(Refer Slide Time: 24:30).



So about 68 percent within one standard deviation of the mean, and when you say one standard deviation of the mean, we mean on the either side of the mean.


(Refer Slide Time: 24:42).



Population

We need to understand the characteristic features of populations from a decision making, quality control or marketing point of view.

Understanding the population helps us to set our goals, objectives, process settings etc.



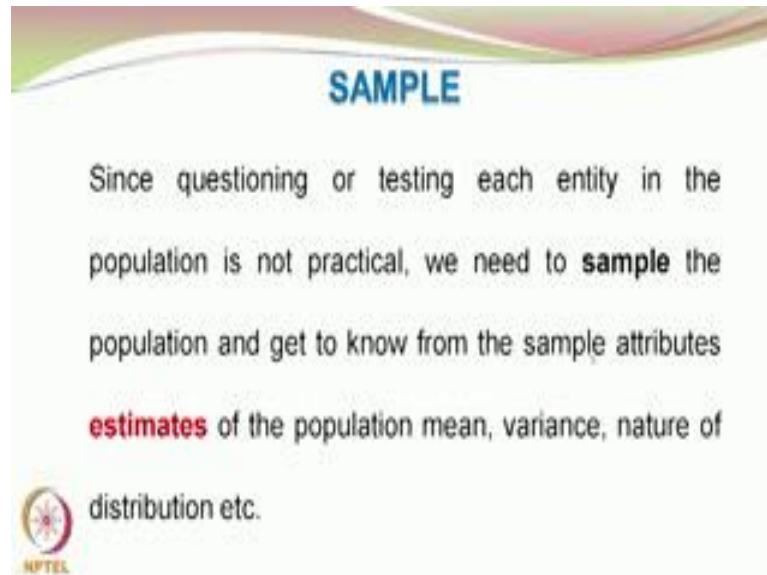
Why do we need all this? It would be ideal if we are working with the normal distribution because it is nice and symmetric, and you can also show that the mean is equal to median is equal to mode, and then the properties of the normal distribution are well understood, and very elegant to work with. **So** it would be nice if the data we are dealing with is described by the normal distribution, but it need not be the case; when we are dealing with the large data set, we really do not know how the data is distributed, and what kind of shape it has, and what would be the probability distribution it is following; so, these things are not known to us. **So** this essentially defines the population.

Why should we deal with population? Because, we want to understand the behavior of large number of entities which come under the collective term of population; for example, you can have the student population or you can have the population of waters or population of people who are preferring the particular soft drink and so on. We need to estimate the characteristic features of the population, so that we have a good idea about the population's preferences or habits or abilities and performances. So, these are required from a decision-making, quality control, and marketing points of view.

So once you understand the population we can set our goals, objectives, process, and settings etcetera. Obviously, if you want to understand the population you cannot go to each and every entity of the population and collect data from that entity, it is practically


not possible. So, we need to have samples taken from the population. And in order to get a liable estimate on the population the sample should be done carefully.

(Refer Slide Time: 26:55).



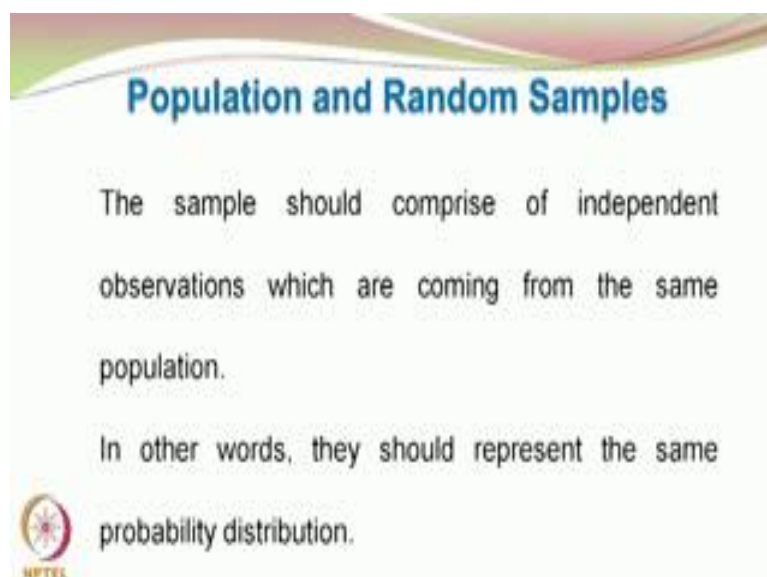
SAMPLE

Since questioning or testing each entity in the population is not practical, we need to **sample** the population and get to know from the sample attributes **estimates** of the population mean, variance, nature of distribution etc.

 NPTEL

So, once you get the sample, you can estimate the population mean, variance, what kind of distribution it may be having and so on.


(Refer Slide Time: 27:04)



Population and Random Samples

The sample should comprise of independent observations which are coming from the same population.

In other words, they should represent the same probability distribution.

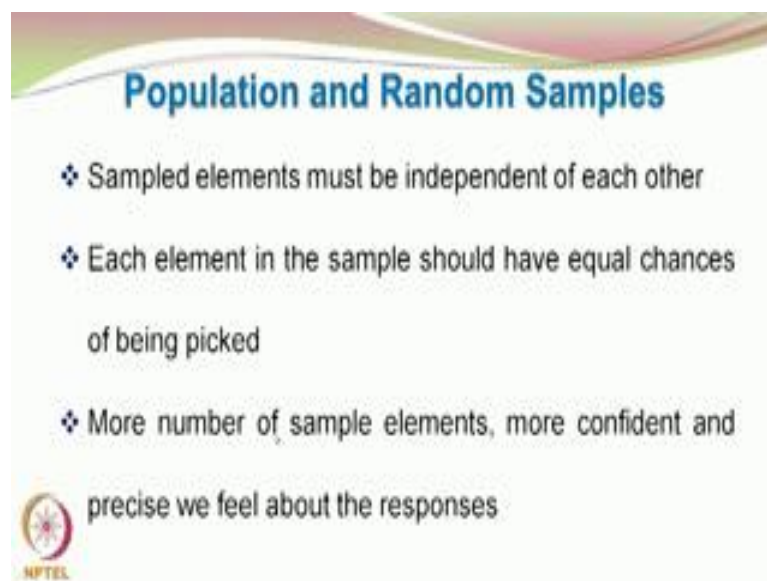
 NPTEL

So how should we sample? Whenever we sample, the observations which are considering the sample should be independent of each other and each entity which we

sample should have equal probabilities of getting picked. Further, they should also represent the same probability distribution. They should all come under same distribution. **So** now, you have a population and random samples, the sampled element must be independent of each other. Each element in the sample should have equal chances of being.


Let us say that you have a collection of 1000 balls, and you want to know an idea about the distribution of the different colored balls in this box, and if all the blue balls are packed in the top most layer, and you pick up all the blue balls, you may conclude that all the balls in this box are blue in color. You may be taking a sample of 10 balls, and if all of them happened to be blue, then that is the conclusion you will get. That means the red balls and the white balls were not given a chance of being picked, but if you randomly distribute all the balls, and then you start picking from the box, if you take a sample of 10 balls, then you will get a reasonable distribution of the balls as they are in the entire box.

(Refer Slide Time: 28:26)



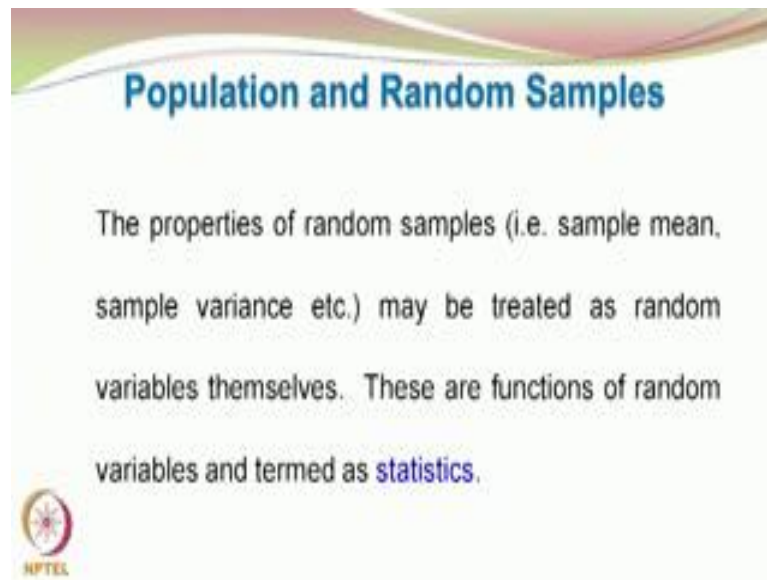
Population and Random Samples

- ❖ Sampled elements must be independent of each other
- ❖ Each element in the sample should have equal chances of being picked
- ❖ More number of sample elements, more confident and precise we feel about the responses

 NPTEL


And again, the third criterion also makes **a** lot of sense. If you have more number of sample elements, more confident and precise we feel about the responses. The more number of balls you pick from the box, better idea you will have about the distribution of the balls in the box.

(Refer Slide Time: 28:42)



Population and Random Samples

The properties of random samples (i.e. sample mean, sample variance etc.) may be treated as random variables themselves. These are functions of random variables and termed as **statistics**.



The properties of the random samples we are interested in are - the sample mean, sample variance and they are also treated as random variables. If the entities of a sample are random variables, then any mathematical combination of them will also be a random variable; and these functions of random variables are called as Statistics.