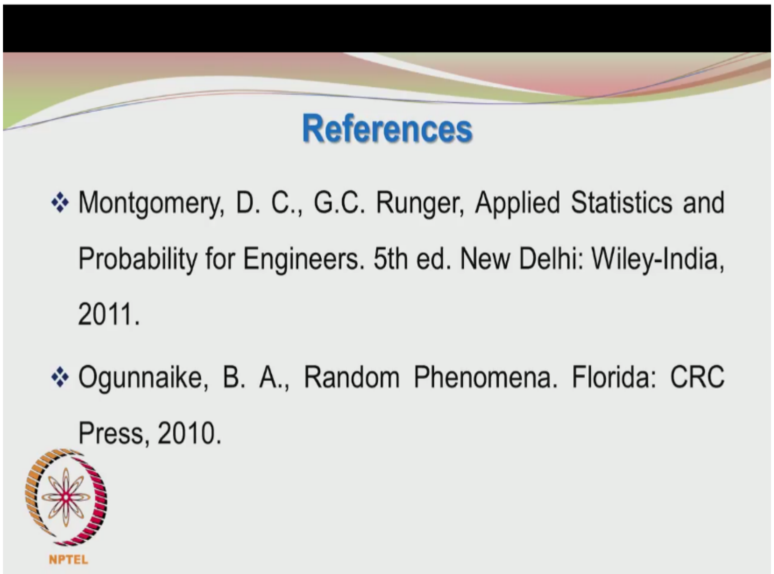
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 10
Random samples: Sampling distribution of the mean (Part A)

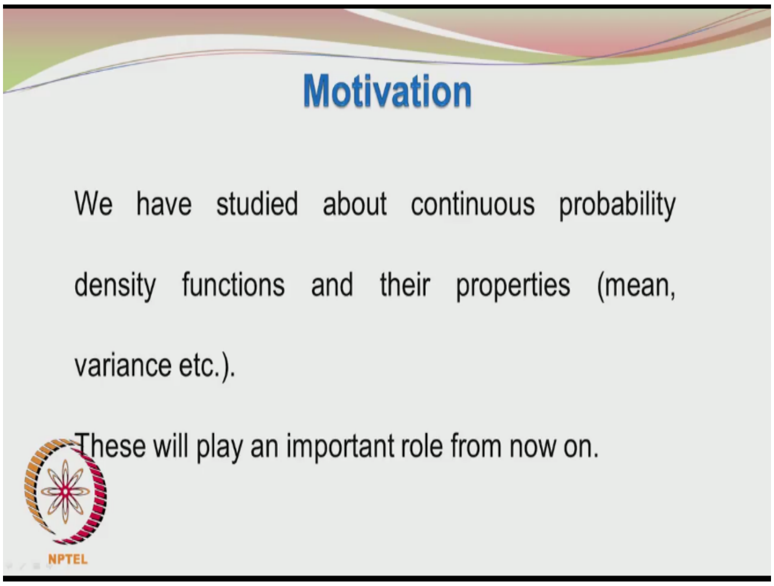
Hello, welcome back to the course on statistics for experimentalists. Today, we will be looking at the random samples and we will be looking at the sampling distribution of the mean.

(Refer Slide Time: 00:34)



As before, the prescribed text books for this course and especially the current lecture is the one written by Montgomery and Runger and I also refer to Ogunnaike's book on random phenomena. The notation I am following are from the Montgomery and Runger's book.

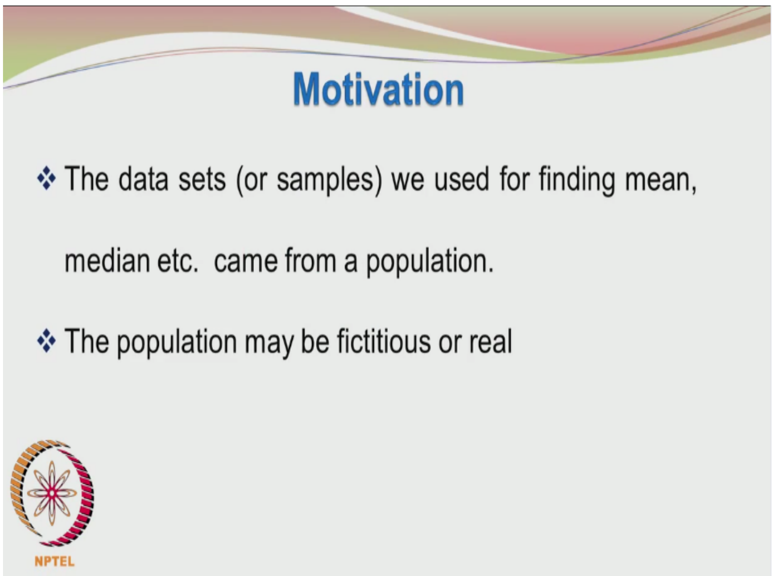
(Refer Slide Time: 01:06)



What have we done so far? We have looked at random variables, the discrete and continuous probability distributions, their properties like mean and variance. We also looked at median and also mode but more frequently we will be looking at the mean and variance. These form the basis for our analysis of experimental data. We have seen that the experimental data may be scattered.

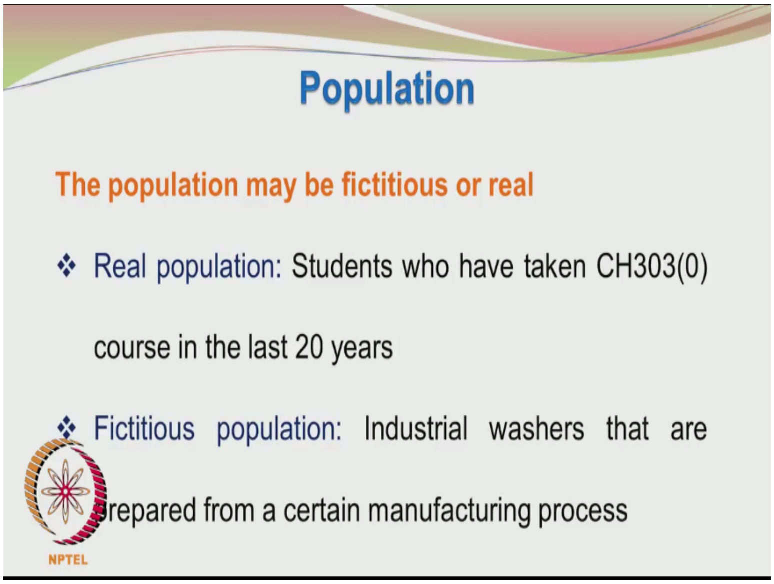
And we need to hence find the average value of the response and also quantify the scatter. We can talk of a large population from which the experimental data are sampled. Hence, we have to look at the properties of samples, what are their desirable characteristics, how we should sample and their attributes. So let us now look at the population random sampling and their properties.

(Refer Slide Time: 02:46)



The population maybe fictitious or real. We can assume that the data or the sample being collected is coming from a certain population. It maybe fictitious one, sometimes the population may also be real and we can directly relate to it.

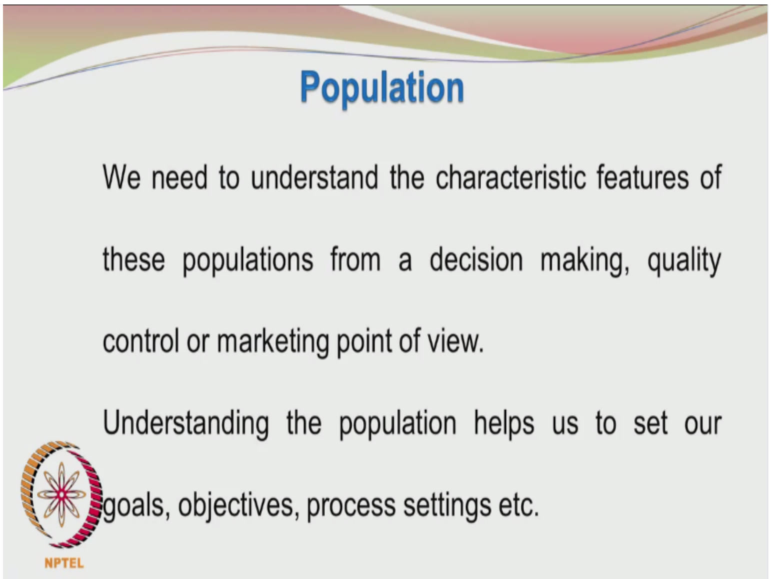
(Refer Slide Time: 03:15)



Let us look at a real population. The population may be the student's community, which have taken let us say a course in chemical engineering over the last 30 years or 35 years. Here I have put 20 years, but it is a popular mass transfer core course. So I guess even at the right at the beginning of the institute, the course would have been there. So we may be looking at the performance of students who have taken this course since it has been offered.

It can also be a fictitious population. It may represent industrial washers that are being prepared by using a particular manufacturing process.

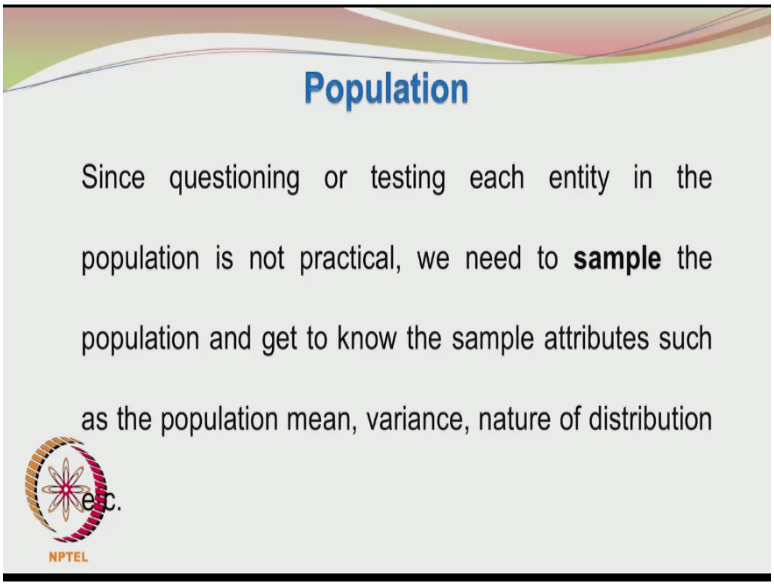
(Refer Slide Time: 04:09)



We do not really need to look at the entire population. Knowing about the entire population may not be practical because it may comprise of let us say millions of entities and trying to get or record their attributes would be a complete waste of time. So we know that there is a population and we are going to take representative elements from the population. It is important to understand the characteristic features of these populations.

So that we can make proper decisions. We can also pass judgment on the quality or make corrections or changes from a marketing point of view okay. It also helps us to set our goals, objectives and the settings at which the processes should be run.

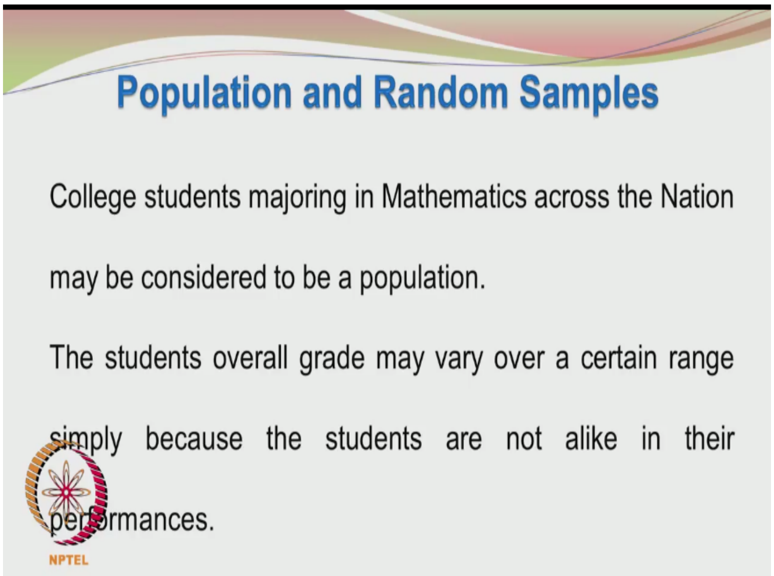
(Refer Slide Time: 05:16)



Population as I said is a very large entity okay and trying to understand the properties of the entire population is a herculean task and so we need to take a sample out of the population.

So that we can try to infer the populations characteristics by knowing the values of the sample. So from the samples attributes like the sample mean and the sample variance, we tried to get an idea or estimate about the population mean and variance.

(Refer Slide Time: 06:07)

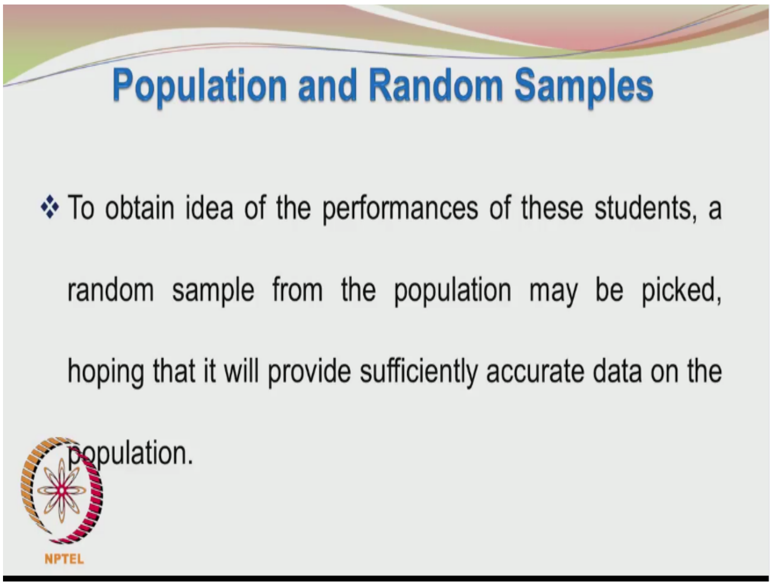


So for example college students majoring in mathematics across the nation may be considered to be a population and the students overall grade or performance may vary over a certain range because all students do not perform alike okay. So there will be a distribution in their performances. We need not assume that the distribution of the student's grades will be normal or Gaussian okay.

That is the simplest or the most direct assumption we tend to make; however, that may not be correct okay. We do not know the populations distribution unless we have prior data or historical evidence okay. So we really cannot assume about the populations probability distribution. We also do not know what is the average of the entire population. We also do not know what is the standard deviation of the entire population.

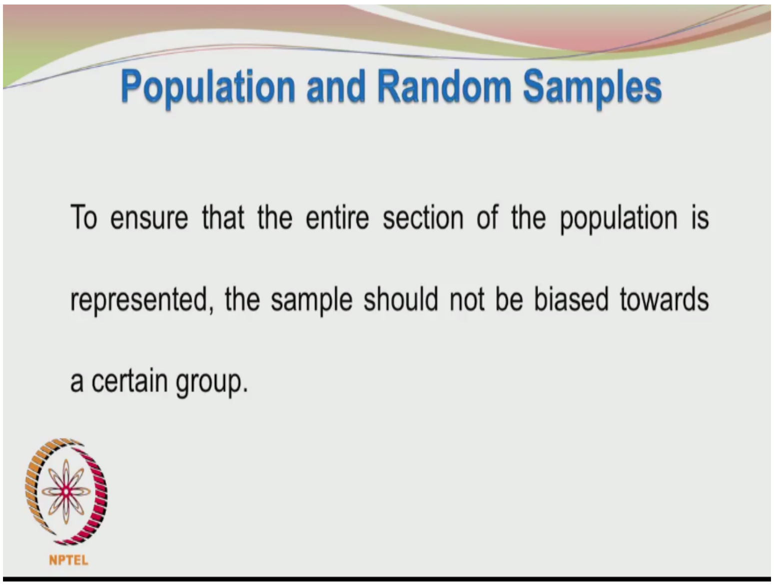
In this particular example, we do not know the average performance or the average mark of the student's population in the nation. We also do not know how their performances or grades are spread. In other words, we do not know the standard deviation or variance of the entire population.

(Refer Slide Time: 07:50)



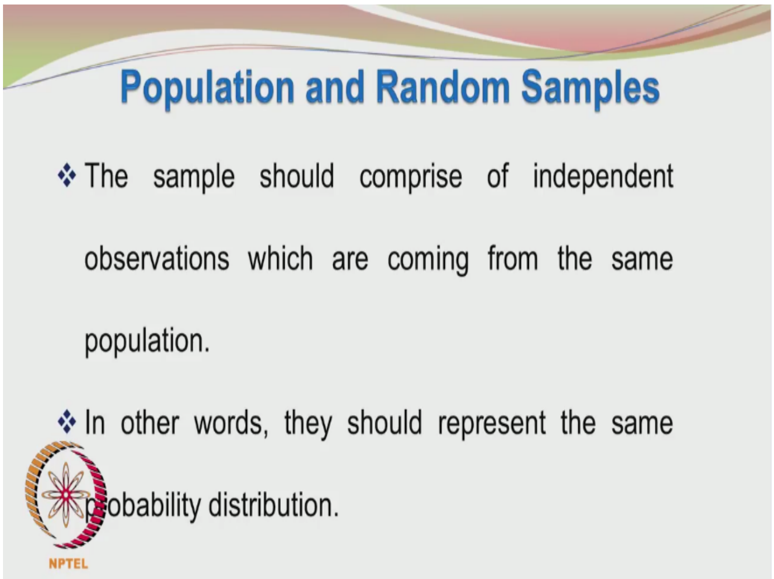
So we have to get an idea how these students majoring in mathematics are performing, we take a sample from the population okay. The sample should be a random one and we also hope that the random sample we have picked from the population is sufficiently representative of the entire population so that we can understand the population better.

(Refer Slide Time: 08:22)



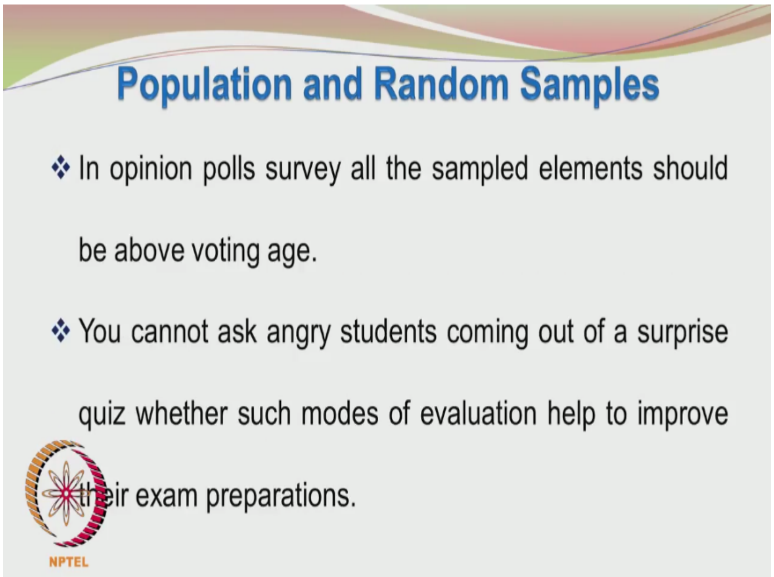
The sample should not be biased towards a certain group.

(Refer Slide Time: 08:29)



The sample should comprise of independent observations, which are coming from the same population. So they should be coming from a population, which is having the probability distribution identical.

(Refer Slide Time: 08:56)

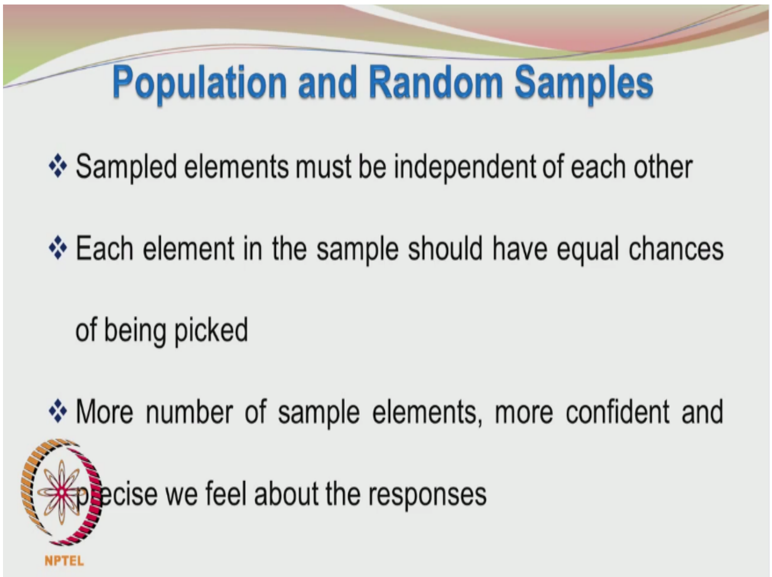


When you conduct a sample, you obviously follow certain precautions. For example, in an opinions polls survey all the sampled elements should be above the voting age. So that all are eligible for considering their voting performances and we can get the correct information. There is no point in asking a person for his preference if he is not or if he or she is not eligible for voting.

And in some colleges and institutions, surprise quizzes are conducted and if you want to know an opinion whether the surprise quizzes are useful from academic point of view then you should conduct a survey at the beginning of the course or at the end of the course where the students would have probably benefitted from the surprise quizzes and they would have been more well prepared or they would have a mixed opinion.

On the other hand, if you ask angry students coming out of a surprise quiz whether such modes of evaluation help to improve their exam preparations, most of them or all of them in fact would be biased towards a negative response to such a query. So you cannot really say that this is truly representative of the collective opinion regarding these conduct of surprise quizzes.

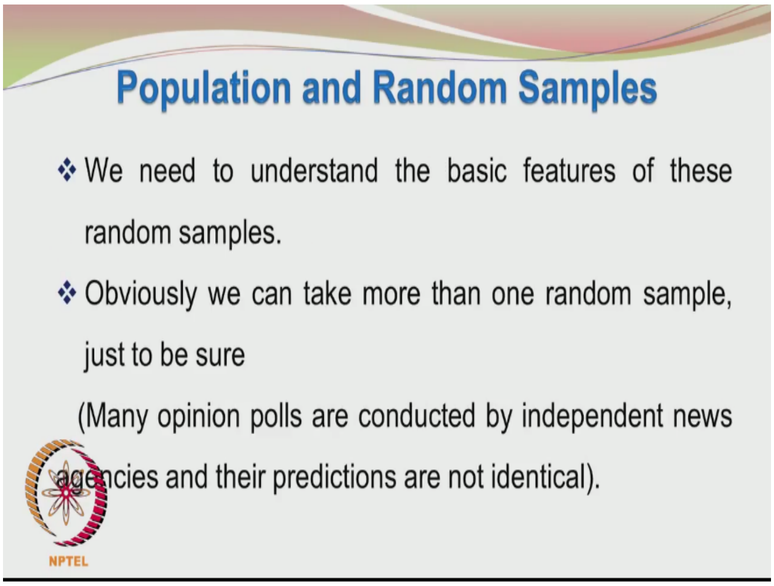
(Refer Slide Time: 10:48)



The sampled elements should be independent of each other and each element in the sample should have equal chances of being picked. Another thing is the sample size. What is the optimum sample size? Obviously, you cannot sample the entire population but if you sample a very large number obviously you are going to get the response more accurately or more precisely.

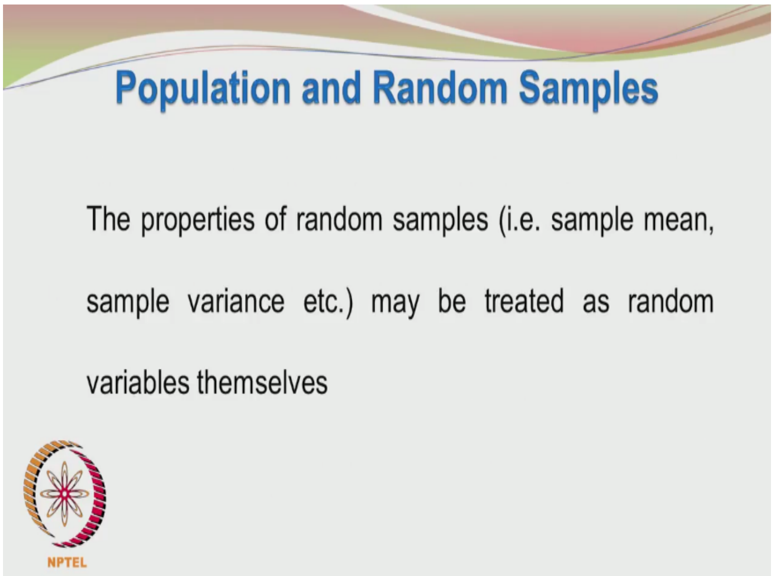
However, this is not very practical to sample a huge response. Sometimes sampling the responses from a population may also mean that you may have to conduct some destructive tests on the specimens and it is not economical to destroy a large number of specimens towards the purposes of sampling. So the sample size is very important; however, it is intuitively evident that larger the sample size more confident and more precise we feel about the responses.

(Refer Slide Time: 11:58)



Right so we have to understand the basic features of these random samples after understanding which we hope to understand the population. We can obviously take more than 1 random sample just to be on the safe side. For example, many opinion polls are conducted by independent news agencies and their predictions are not identical.

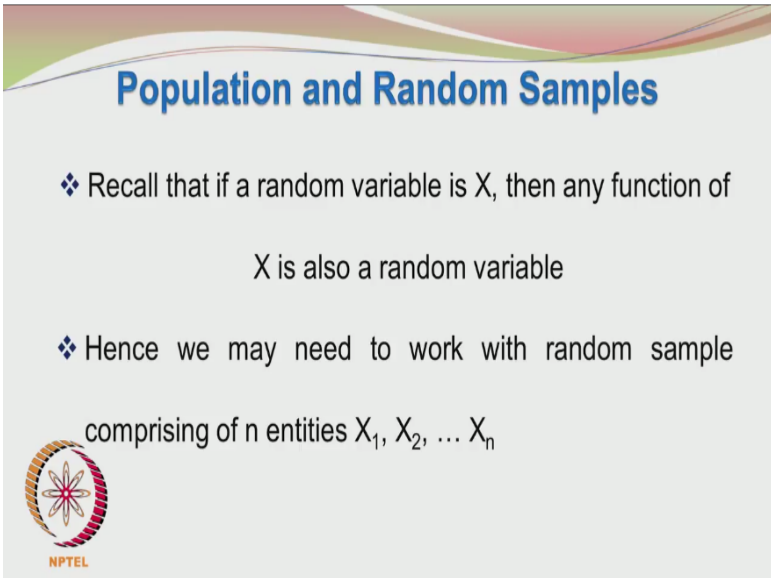
(Refer Slide Time: 12:22)



So coming back to the random variable, we know that random variable is described by a probability distribution and any mathematical combination of the random variables will also result in a random variable. So a combination of random variables may also be treated as random variables themselves.

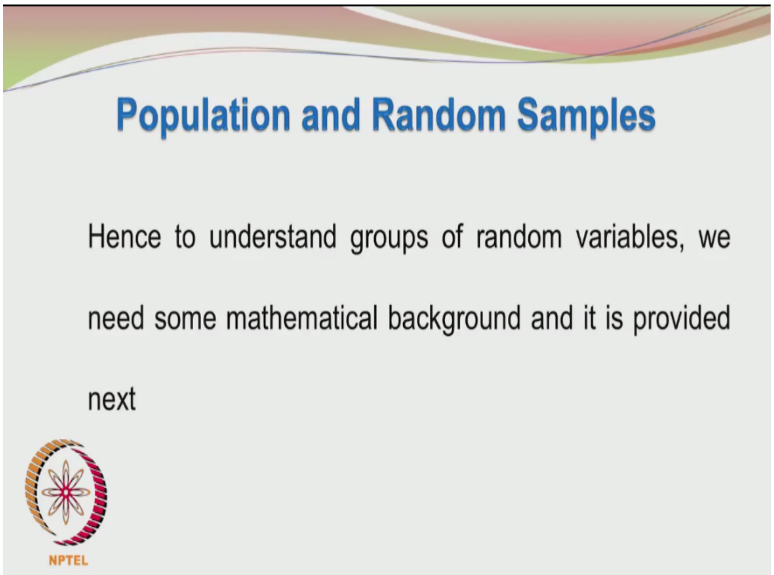
For example, the sample mean and the sample variance which are obtained by collection of random variables may be treated as random variables themselves. So the sample mean is a random variable, sample variance is also a random variable.

(Refer Slide Time: 13:08)



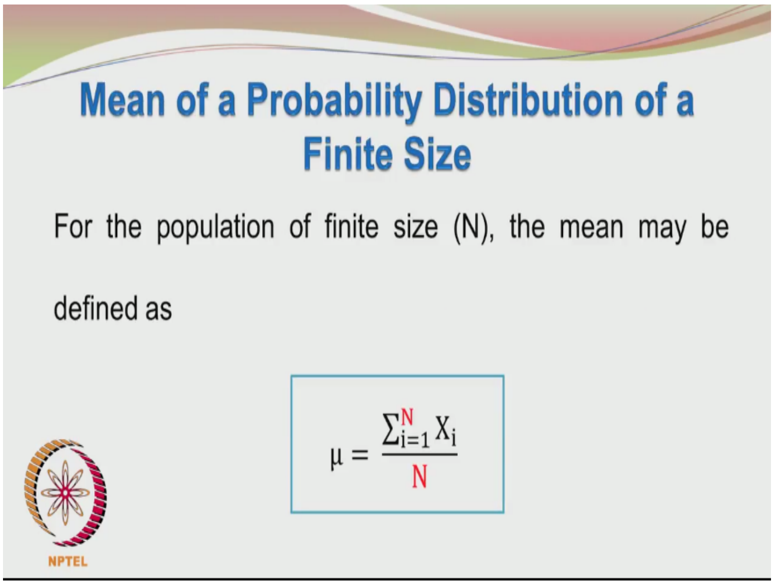
Let us say that we have conducted a survey or we have done the sampling and we have taken n entities from the population. Let us denote them by X1, X2, and so on to Xn. So these are all random variables and these are independent.

(Refer Slide Time: 13:36)



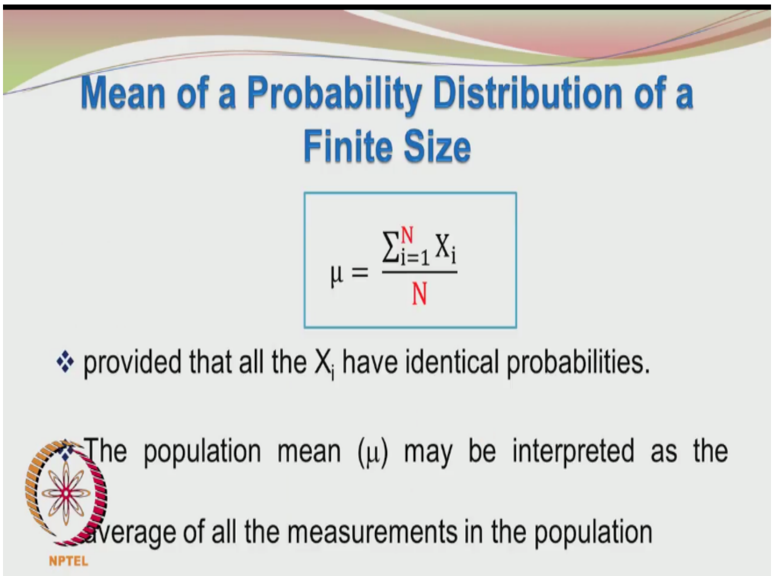
So to make further progress in random variables and random samples, we need a bit of mathematical background, which will be provided in the next few slides.

(Refer Slide Time: 13:50)



Let us first look at a few definitions. Some of these definitions we might have come across earlier. I am repeating them so that you become more familiar with them and also there are some small differences between various definitions, which you should be aware of. So if you look at the population of let us say a finite and large value. The mean may be defined as mu=sigma i=1 to N where N is the size of the population Xi/N.

(Refer Slide Time: 14:31)

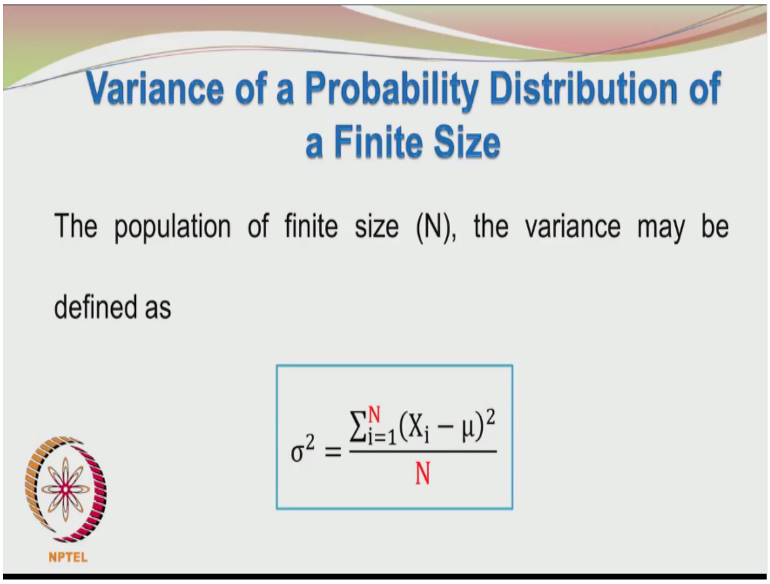


We really do not care to sample the entire population so we really do not know what the exact value of N is. Most likely it is a very large number but we do not worry too much about the population mean. Even though it is central to all our discussions, we do not really know its exact value. So we are having a discrete collection of random variables taken from the population.

And if we attribute the same probability to each and every entity in the population, we have the probability as simply 1/N. There are N entities in the population and if each is having the same chances of being selected or quit then the probability would be 1/N. So the mean becomes simply sigma i=1 to N Xi/N. We assume that the Xi random variable has identical probability from i=1 to N.

The population mean mu may be interpreted as the average of all the measurements in the population right.

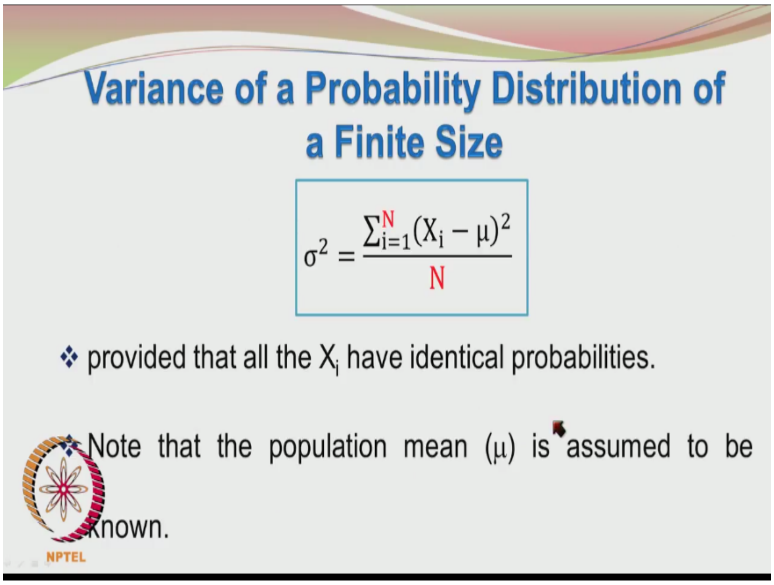
(Refer Slide Time: 15:55)



So we know the value of mu. Let us say it is known to us beforehand. This is very important okay that we have knowledge of mu somehow. For example, it may be a hypothetical stated design specification right. So that is giving us the value of mu. When mu is known already, the variance sigma squared since we are talking about the population, we are talking about sigma squared.

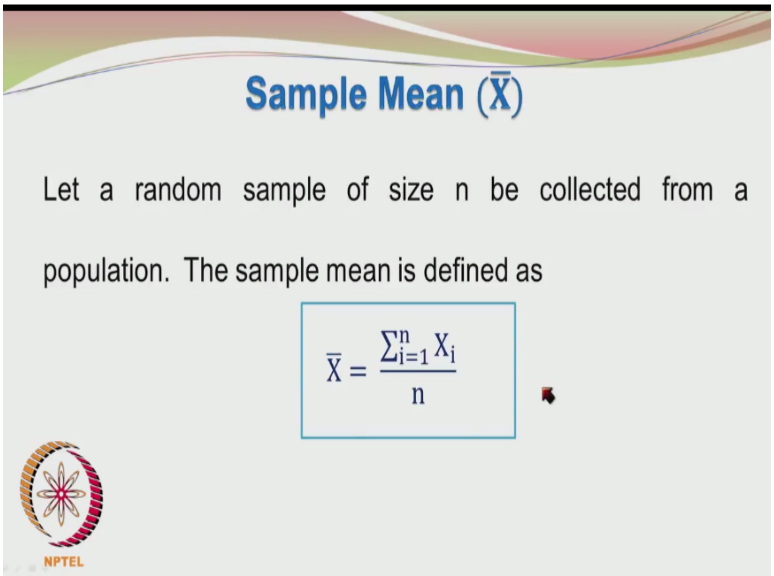
Just as we talked about the population mean and represented it with mu, the population variance is represented by sigma squared. So we have sigma squared=sigma i=1 to N Xi-mu whole squared/N. Note that we are using the entire population size N here. Earlier in the sample variance, we used the sample size-1 but here we are using capital N which is the entire size of the population.

(Refer Slide Time: 17:10)



This sigma squared is also based on the fact that all the Xi values have identical probabilities and the population mean mu is assumed to be known.

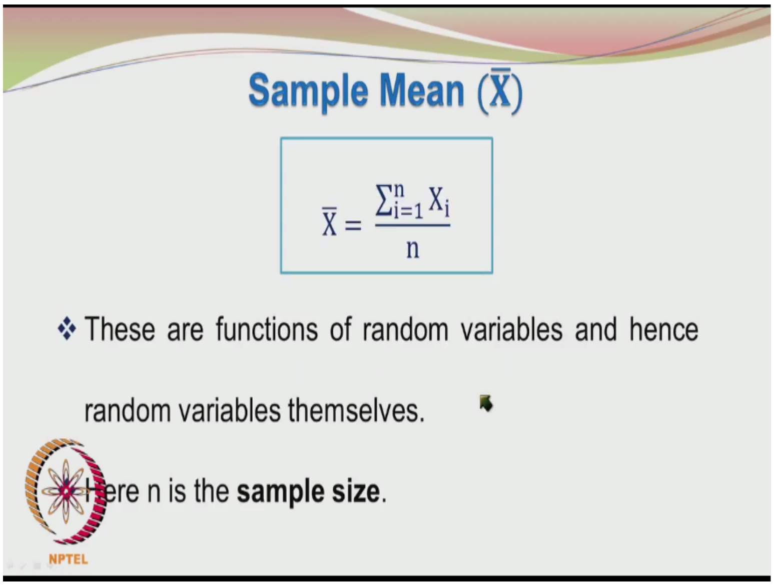
(Refer Slide Time: 17:32)



Now let us define the sample mean. The sample mean is collected from a population. Obviously, we cannot sample the entire population. So we take a sample of size small n and the small n value is much, much lower than the capital N value. In other words, the sample size is much smaller than the population size. So the sample mean is defined as X bar=sigma i=1 to n Xi/n where n is the sample size.

Xi is all the entities in the random sample collected. They are assumed to be random and identically distributed and also have equal probability of being selected.

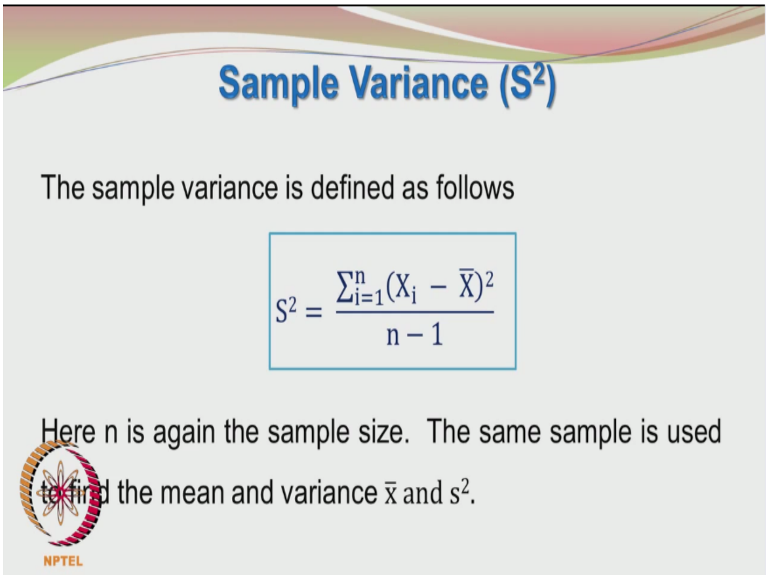
(Refer Slide Time: 18:31)



This is a sum of random variables and we are totaling them and dividing it by a constant value the sample size n and hence this is a mathematical function of the random variables resulting in the sample mean X bar okay and since these are random variables, a function of the random variables is also a random variable. So we may treat X bar as a random variable. It is a quantity derived from the random variables.

It is also a random variable and so it will have a probability distribution.

(Refer Slide Time: 19:22)



A sample variance is defined as S squared=sigma i=1 to n Xi-X bar whole squared/n-1. Note that we are using capital S squared and we are also using capital Xi and capital X bar in these definitions. Earlier in the sample mean also we used capital X bar. As of now, we are manipulating random variables, which are abstract entities until the experiment has been carried out or the sample has been collected.

So they are abstract entities until then and we are defining another abstract entity X bar in terms of these random variables. So we are using capital X here and also we are using capital X here. Similarly, before the experiment is conducted or the sample has been selected, we are talking about abstract entities and so we have Xi, capital X, capital X bar and this S squared is also having a capital S.

Another feature is we are using n-1 here. We have already seen why we should be using n-1 in this definition, n is of course the sample size and we use n-1. It is a measure of the degrees of freedom and when we refer to degrees of freedom, we always talk about independent entities. The degrees of freedom represent number of independent entities and since X bar was not known beforehand, it was calculated from the sample okay.

We are using the same sample to find X bar and then trying to find a variance of that particular sample. So the n deviations, Xi-X bar are not independent of each other. So we use n-1. Another important thing is it is a measure of the deviation from the mean. Ideally, this should have been mu but we do not know mu. It is only X bar and the Xi values are closer to the sample mean than to the population mean okay.

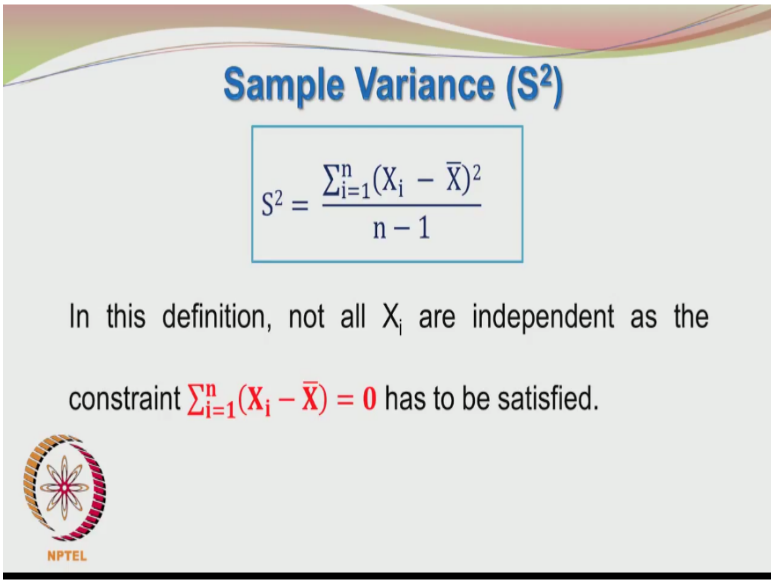
The X bar itself has been so defined that the Xi values are clustered around it and it is in the center of all these random variables but the mu need not be at the exact center in which case this Xi-X bar whole squared and you sum that for all the entities in the sample is likely to be smaller than if you had sum the squares of Xi-mu okay. So the important thing is when you are trying to calculate the spread based on the sample, the spread is likely to be smaller.

Because the Xi values are clustered around X bar and their distances from X bar are effectively smaller than the distances from the unknown population parameter mu. So you are trying to reduce the scatter by basing the variance definition on X bar. So this reduction of the scatter may give a false sense of security. You may feel that there is not much scatter in the data okay.

And to compensate for that what we do is if this deviation squared is slightly smaller because we are using X bar instead of mu, here also we reduce the degrees of freedom and put it as n-1 instead of n okay. So the numerator has decreased and the denominator has also decreased, so there is a compensating effect. So the S squared is more reliably estimated and it is a true reflection of the population variance sigma squared.

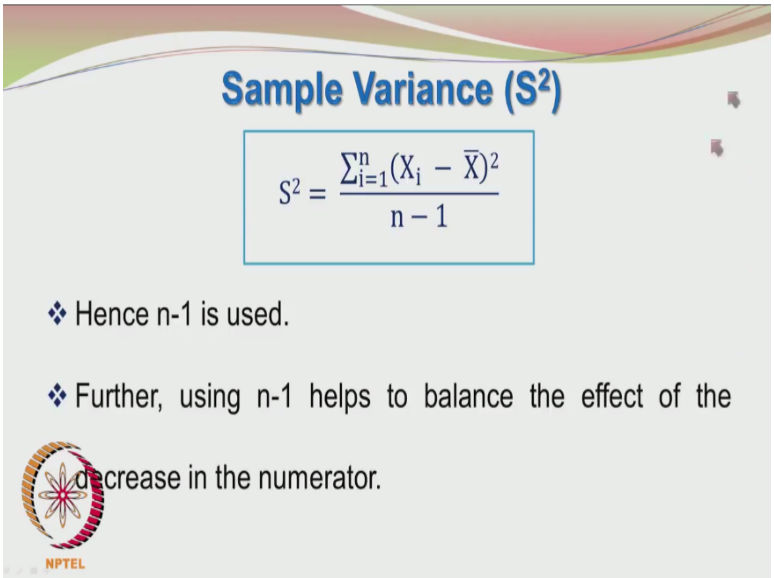
Remember the population is having parameters mu and after sampling we get X bar and we hope that X bar is suitably representative or sufficiently representative of mu. We also get the sample variance based on the sample and we hope that S squared is sufficiently representative of sigma squared.

(Refer Slide Time: 25:08)



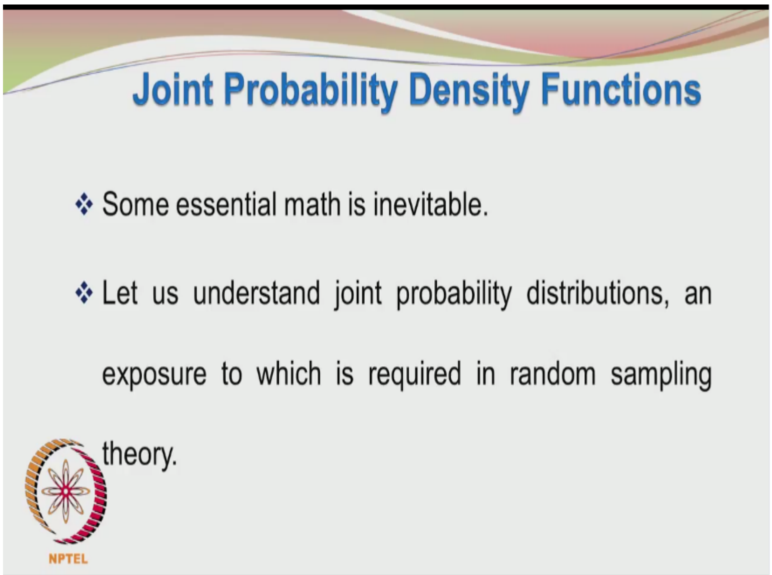
So there are 2 reasons for using n-1, the first one is sigma=1 to n Xi-X bar=0. So this is a constraint which tells that only n-1 Xi-X bar are really independent. If we know n-1 Xi-X bar values using the fact that the sum of all the deviations=0, we can find the nth deviation. We have n-1 deviations, we also have this particular constraint that the sum of the deviations=0, so using the n-1 deviations and this constraint, we can find the nth deviation.

(Refer Slide Time: 26:03)



So the nth deviation is truly not independent one and using n-1 also helps to balance the effect of the decrease in the numerator. This is also what I discussed a couple of slides back.

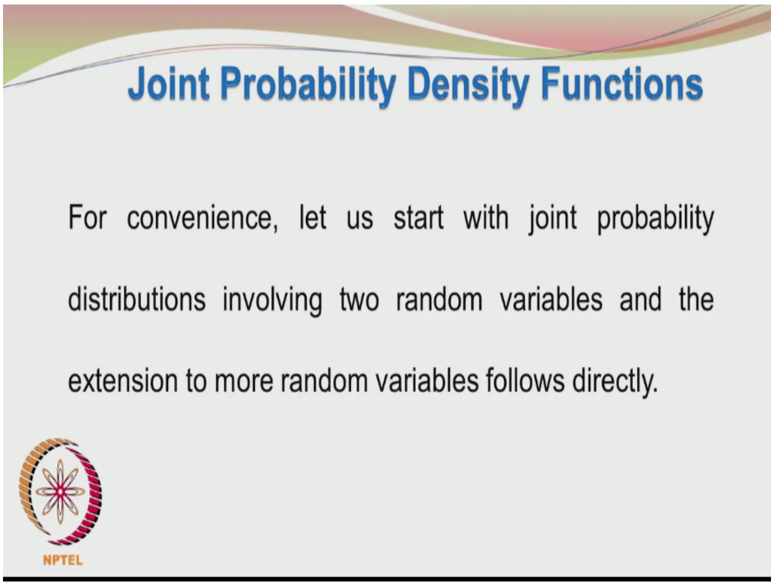
(Refer Slide Time: 26:16)



Now we come to joint probability density functions. Some of you may not really follow the detailed math which I am going to discuss shortly; however, there are 2 ways of doing it, one you look up your favorite book on calculus, try to understand about multiple integration. Multiple integration is very straight forward. It is direct extension of the simple single integration.

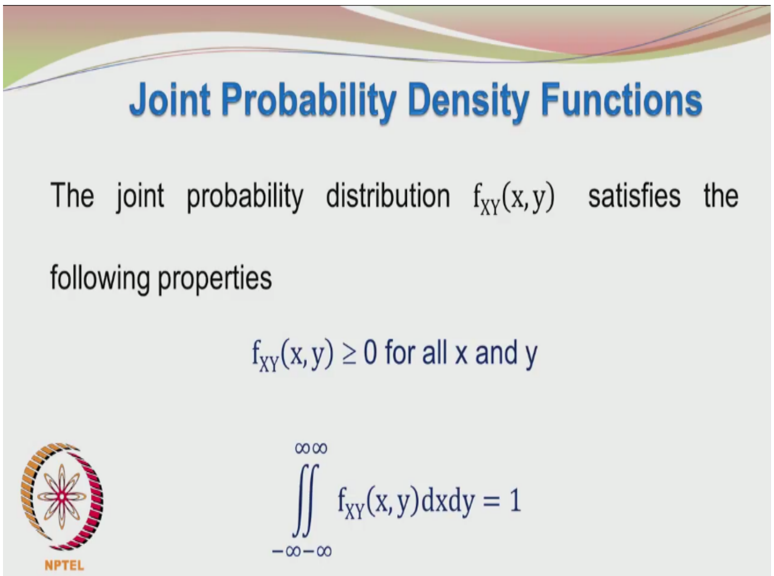
Another way is you do not have to understand the derivation but please understand the final proof okay that is very important. Even if you did not follow the mathematical derivations, you can also try to understand the conclusion, the main conclusion which is coming at the end of the derivation.

(Refer Slide Time: 27:27)



A joint probability distribution involves 2 or more random variables. They are described as the name implies in a joint fashion okay. There are 2 random variables, which are occurring together. So we will be first starting with the joint probability distribution of 2 random variables and then we will take up the extension of more general case where n random variables are jointly described.

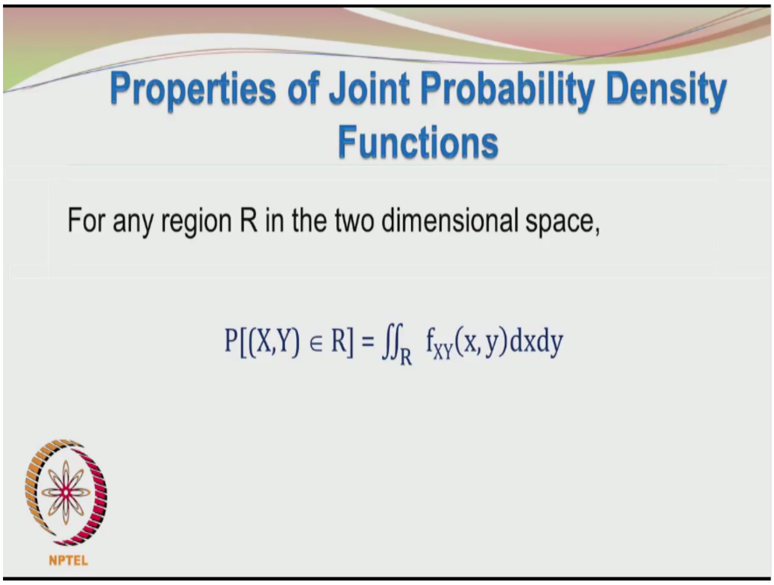
(Refer Slide Time: 28:17)



If you look at the joint probability distribution functions, there will be lot of similarities with the single random variable probability distribution function. So use that as the basis to understand the multiple random variables case. The first property is f of XY x, y should be>=0 for all x and y. So this is a probability distribution function involving random variables, capital X, capital Y and it is a function of small x and small y.

And it should be positive. This small x and small y represent the values of the random variables X and Y after the sample has been carried out or the experiment has been carried out. Just as you see that the sum of all the probabilities should be = 1, we are looking at continuous probability distribution functions and so we have f of XY x, y dxdy = 1. Earlier in the case of a single random variable, we had the integral f of X dx=1. Now we are having f of x, y dxdy=1.

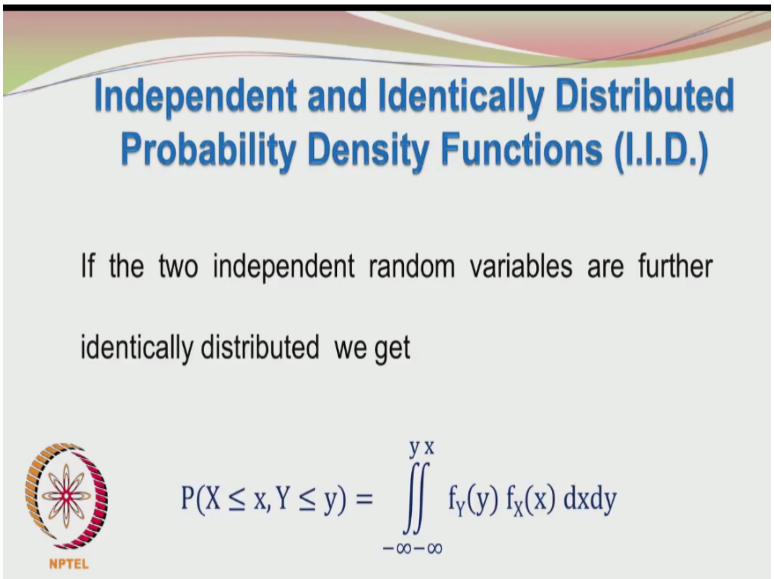
(Refer Slide Time: 30:04)



The f of x, y is defined such that probability of the random variable X and Y belonging to region R in a 2-dimensional space is given by double integral of R f of XY x, y dxdy. So it is just a statement defining the probability of the 2 random variables belonging to a particular region and that is given by the double integral.

Double integral represents the area okay. So we are talking about a random variable X which is covering a certain length and we can think of the random variable Y which is covering a certain width, so a joint distribution of both X and Y would cover a rectangular region in the 2-dimensional space or a rectangular surface and that is given by the double integral f of XY x, y dxdy.

(Refer Slide Time: 31:24)

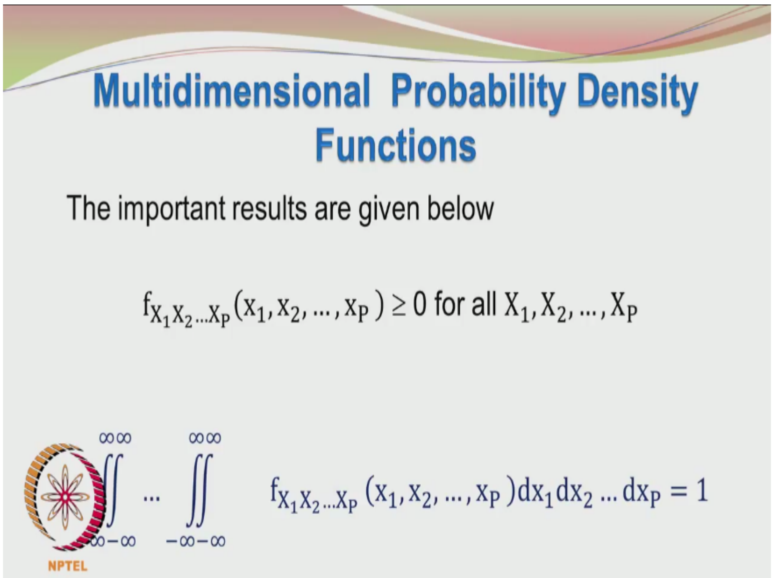


Now we come to a very useful and interesting property of the joint probability distribution function. If the 2 random variables, which are defined together are in fact independent of one another and they are also identically distributed okay. They are independent and identically distributed, we get probability of X<=x Y<=y. We have again the double integral going from –infinity to y –infinity to x f of Y*y*f of X*x dxdy.

So earlier it was f of XY x, y and now it has become f of X*x*f of Y*y. The fact that we have split them into 2 functions implies that the random variables X and Y are independent and they are also having the same functionality, only the variable is different. Here it is x and here it is y but the functional form is identical. This shows that the 2 random variables are independent but they are identically distributed.

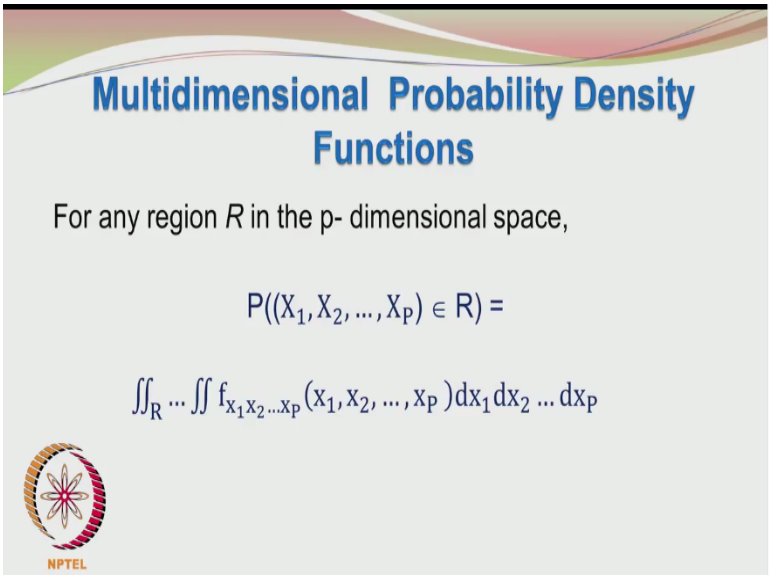
One can say that equations are also like pictures. What can be said in many words can be represented in one photograph or in one equation.

(Refer Slide Time: 33:11)



So when we are considering the joint probability density function involving p non-independent random variables, the f of x1, x2 so on to xp should be>=0 for all values of X1, X2, so on to Xp and then the area under the curve should be=1. So earlier we were talking about 2 random variables, now we are going in for a general case where we have p random variables and the functional form is given in this manner.

(Refer Slide Time: 34:02)



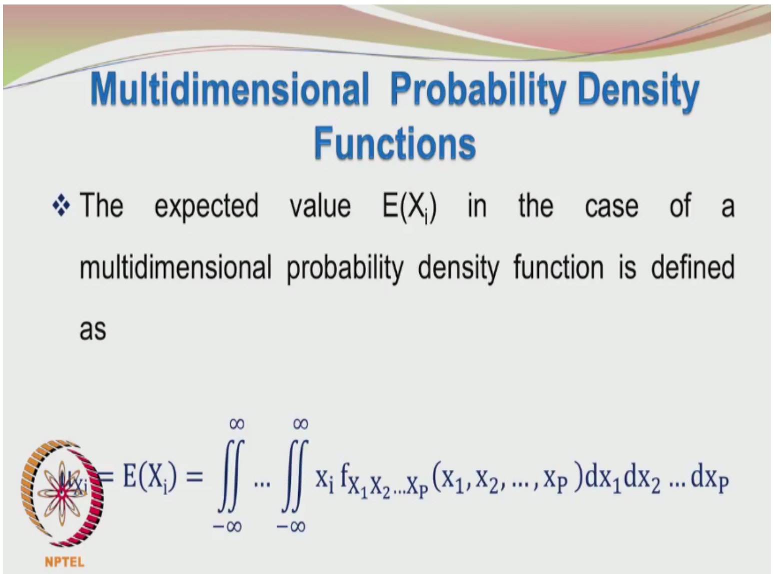
So if you look at a region in the p dimensional space because there are p random variables, you have the multiple integral encompassing or surrounding the region R, the p dimensional region and that probability value is obtained by calculating this multiple integral by using the joint probability distribution function f of x1, x2 so on to xp*X1 this is not into this is actually a function of X1, X2, so on to XP dx1 dx2 so on to dxp okay.

So what is the real meaning of this argument? We know that the random variable will take a particular value but when we are talking about continuous probability density functions, the probability at a particular value whether it is a single random variable or many random variables, the probability at a particular point will be=0 just as the weight of the conical block of wood is 0 at a point.

Only when we consider the thickness of the conical block of wood and its surface area, we can talk about volume and then we know the density and then we can talk about its weight. So we are also saying that in the entire space we are talking about a region R and so what is the probability of the random variable X1, X2 so on to XP falling within that region R okay and that would be represented by coordinates for X1, X2, so on to XP.

If this lower limit may be –infinity or it may be another lower boundary so we are talking about the n-dimensional space formed by all these random variables and in this n-dimensional space, you can have the lower boundary and the upper boundary. This is very interesting; however, we may not be really doing some problems with these multiple integrals. We had to just understand the concepts.

(Refer Slide Time: 36:52)

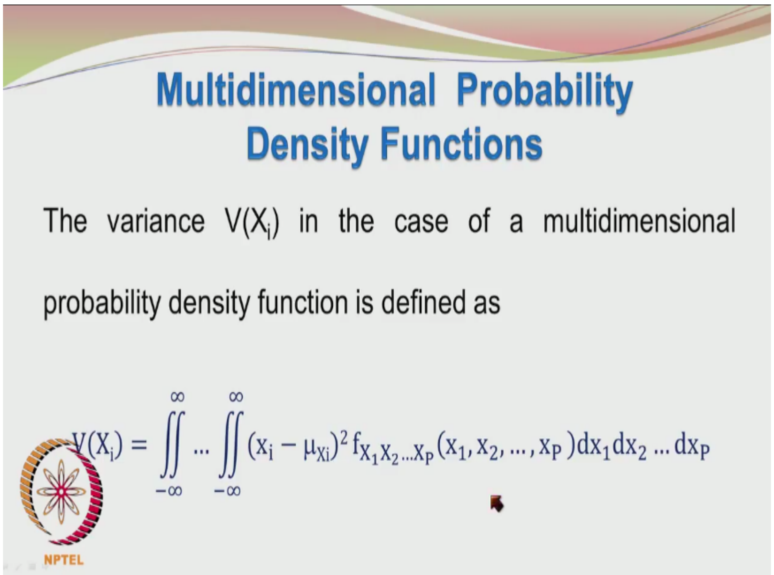


So we can talk about the expected value of a particular random variable Xi and that is given by mu of Xi=expected value of Xi and that is given by the multiple integral ranging from –infinity to +infinity. That particular random variable representative or the sampled value or the exponential value Xi*f of X1, X2 so on to XP as a function of all these random variables X1, X2 so on to XP dx1 dx2 so on to dxp.

So we do not put any numerical value here that is an important thing. We do not put 5 or 0.5 or anything. We are just saying that the random variable corresponding to Xi, Xi maybe X1, X2, or XP so for that particular random variable what is the expected value? Given this joint probability density function, the expected value of that particular random variable obtained from the joint probability density function.

So we plug in Xi, the variable Xi corresponding to the capital Xith random variable and then we multiply with the joint probability distribution function, carry out the integration and the limits of the integration are from –infinity to +infinity. Since we have carried out the integration from –infinity to +infinity as a result of this exercise will get a numerical answer provided this function is well defined.

(Refer Slide Time: 38:49)

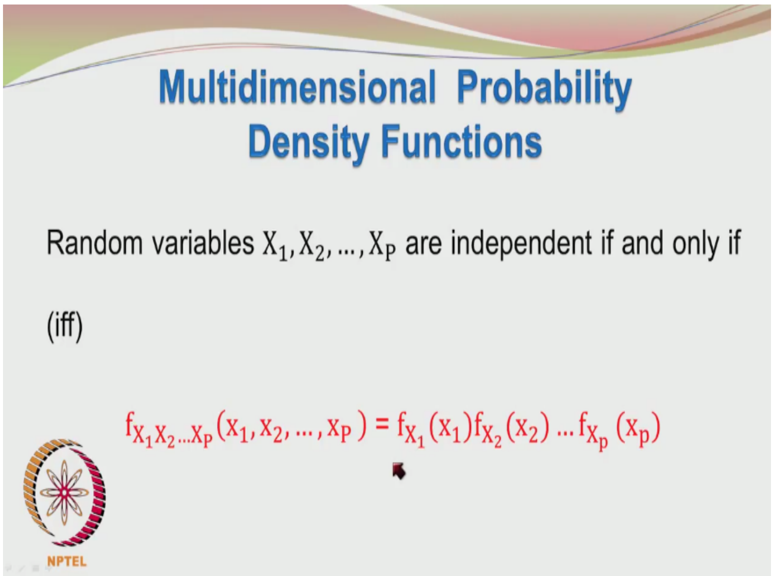


Similarly, we can find the variance of Xi. Whatever we did for the single random variable, we are now doing it for again a single random variable but now described in terms of its association or combination with other random variables X1, X2 so on to XP. So variance of a particular random variable Xi is obtained by again carrying out the multiple integration between the lower boundary to the upper boundary.

Here it is –infinity to +infinity as a general case. Xi is the value of the random variable Xi we are looking at. Here Xi is not a numerical value, we write only simply as small xi and then we put –mu of Xi which is the expected value of the capital Xi whole squared into the probability distribution function. In the simplest case where we had only one random variable, where we had only one probability distribution, we had written as f of x dx x-mu whole squared okay.

Now we are writing in terms of x-mu whole squared, mu was obtained from the expected value of Xi and then we are multiplying with the joint probability distribution function. We get the variance V of Xi. Again, this will be a value because you are integrating it from –infinity to +infinity. It will be a numerical value.

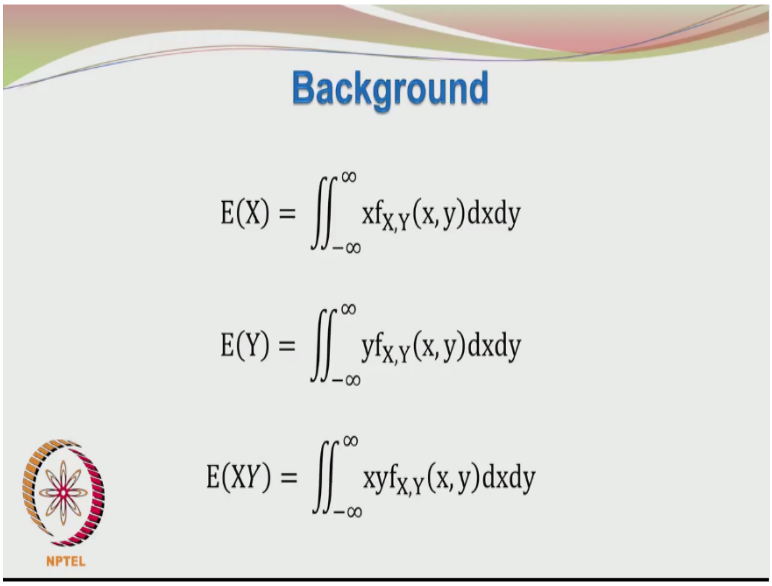
(Refer Slide Time: 40:52)



So the independence of P random variables is going to be discussed in the current slide. The random variables X1, X2 so on to XP are independent if and only if f of X1 X2 so on to XP x1, x2 so on to xp this is small x and this is capital X is obtained from the product of the individual probability distribution functions of x1 x2 so on to xp okay and important thing is they are all identically distributed.

In other words, the functional forms are the same, only the arguments of these functions are different. So this is a joint probability distribution function and if the random variables forming the joint probability distribution function are independent of one another, we can multiply with the individual probability density functions. This is a very useful result.

(Refer Slide Time: 42:06)



Now we will be looking at product of 2 random variables. So just to make yourself familiar you may look at these equations. Expected value of X is the particular representative of capital X that is small x f of X, Y x, y dxdy. We are talking about a joint probability density function involving 2 random variables x and y. Similarly, we have expected value of Y=-infinity to +infinity y f of X, Y x, y dxdy.

Now this is the important thing. Here we are writing it as a product of 2 random variables XY, there is nothing to worry or panic. We simply put it as the representative of capital X which is small x, representative of capital Y which is small y and then we put the same joint probability distribution function and carry out the integration and report the final answer okay.

So we are having E of X here, E of Y here and E of XY here. So we will take a small break here okay. We have looked at population, we have looked at sample, we have also seen that by looking at the properties of the sample, we tried to infer the properties of the population. So the samples properties to be a good representation of the populations properties, we need that the sample is indeed random.

Each element in the sample should have equal probability of being picked and they should also be identically distributed. Then we saw that a sample involves multiple random variables, a function of multiple random variables is also a random variable. Since we are talking about the collection of random variables, we also need a brief mathematical background on the joint distributions of the random variables.

Fortunately for us, the sample comprises of independent random variables. So the probability density function for the joint random variables case gets considerably simplified. We also assume that they are identically distributed. The random variables are identically distributed in which case the individual probability distribution functions are also identical. So considerable simplification has been enabled by the virtue of our requirements.

And then we looked at the expected value of the random variables in the joint probability distribution conditions. We also found that just as we can define expected value of X, expected value of Y, we can also define expected value of a combination of random variables X and Y. We will proceed after a small break at this point.