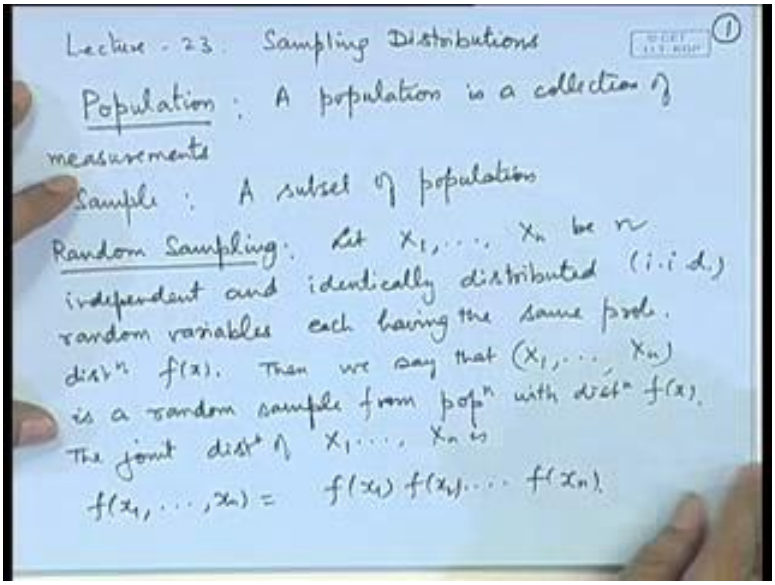
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute Technology, Kharagpur

Lecture - 45
Basic Concepts

So, today we will introduce sampling distributions.

(Refer Slide Time: 00:26)



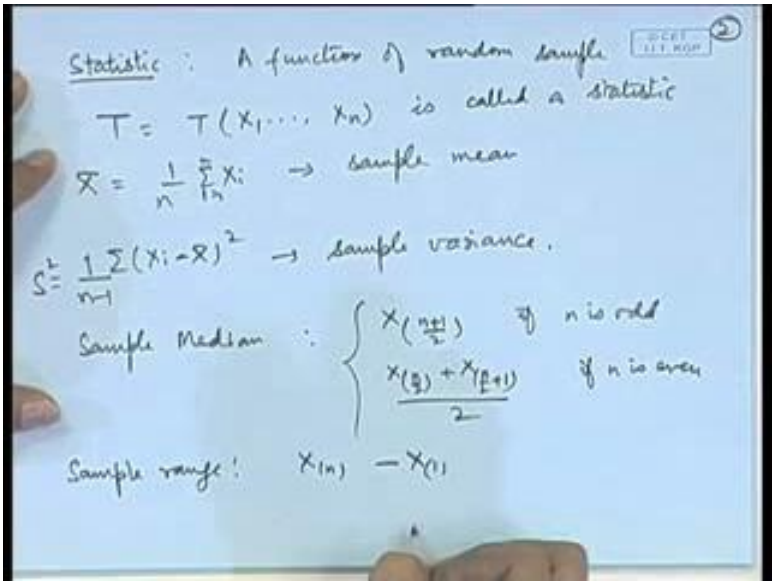
So first of all we introduce what do we mean by a sample; what is a population, so we introduce the term population. So, a population is a collection of measurements on certain characteristic for example, if you are studying heights of people then the measurements of the heights of our desired target population that will be the statistical population. If we are interested in the lives of the people or longevity of people, then if we consider say the total life; total age at death of a set of people then that is our target population. If we are interested in say the number of smokers in a population, then the characteristic of recording that is whether a person is a smoker or not a smoker for a certain set that is our target population. So, a statistical population is a collection of measurements whether it is numerical or a qualitative measurements.

A sample is a subset of population, so since it may not be possible to have the complete enumeration of the population in various studies it is enough if we consider a certain sample of the population. So, a general random sample which we consider in statistics is

taken in such a way that the probability of selecting each observation is same; however, this is the methods of doing sampling it is a part of another topic called sampling theory or sampling techniques. In this particular course we are assuming that we already have random sample and then we proceed with that, so what is a random sample in the contest of distribution theory.

So, we say that let X 1, X 2, X n be n independent and identically distributed that is i.i.d, random variables each having the same probability distribution f x. Then we say that X 1, X 2, X n is a random sample from population with distribution f x and the joint distribution of X 1, X 2, X n is defined as f of x 1, x 2, x n is equal to product of f of x 1, f of x 2, f of x n, any characteristic of the sample we call it as a statistic.

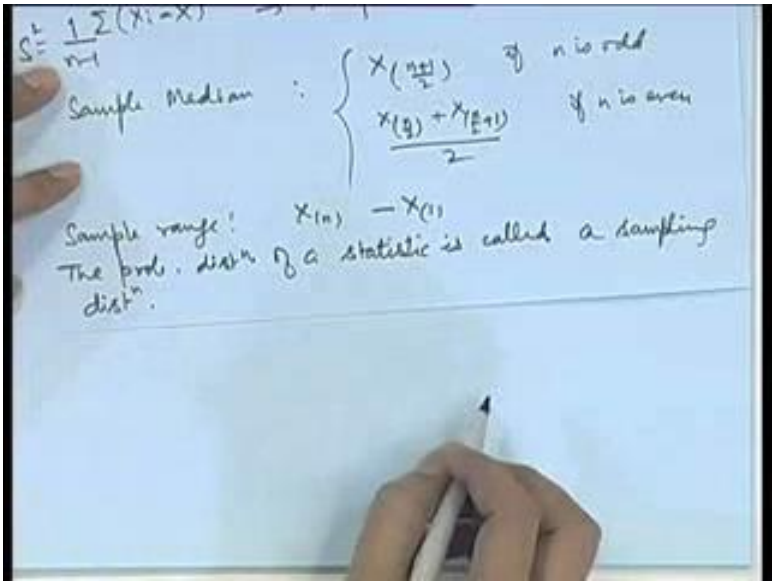
(Refer Slide Time: 04:21)



So, a function of random sample let us say T; that is T of X 1, X 2, X n this is called a statistic. For example, we may consider X bar that is 1 by n sigma x i; that is the sample mean, we may consider sample sum of a squares from the deviation from the mean, we may consider 1 by n minus 1 of these which we usually denote by s square; that is sample variance, we may interested in say sample median. We already talked about order statistics, so that is also a statistics; sample median we may define to be the X of n plus 1 by 2 that is n plus 1 by two-th order statistics; if n is odd; that means, the middle order statistics or if n is even; then we may take the mean of the middle 2 that is X n by 2 plus X n by 2 plus 1 by 2.

We may consider say sample range that is the difference between the largest and the smallest. So these are examples of certain statistics and when we are dealing with a sample; we are interested in these characteristics and therefore, we will be interested in their distributions. So, the distributions of the statistics they are known as sampling distributions.

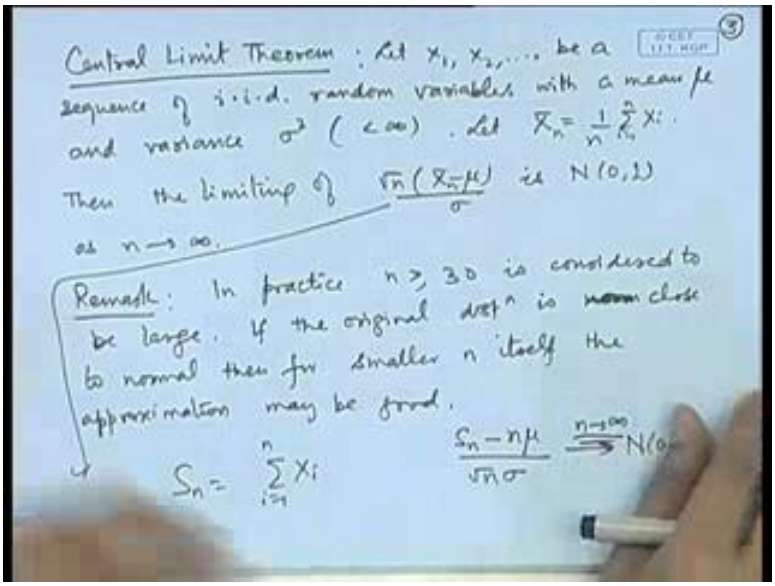
(Refer Slide Time: 06:10)



So, we may formally define a sampling distribution is the probability distribution of a statistic is called a sampling distribution. Now as such here the distribution of X 1, X 2, X n is known, so the joint distribution of the sample is known to us. So, if we consider any function of that T of X 1, X 2, X n; the derivation of the distribution relates to the technique which we have defined in the previous lecture that is for transportation of random vectors; that means, we may considers a one variable as T of X 1, X 2, X n and we may define some other variable say u 1, u 2, u 1 minus 1. So, that we have end to end transformation and we may determine using any techniques for determining the sampling distribution.

However there are some particular characteristic such as sample mean or the sample variance which play very important role and here we will consider the distributions of that.

(Refer Slide Time: 07:37)



So, one of the first results which is related to the distribution of the sample mean is a quit important result in the sense that, it applies to a very large number of situations; it is known as central limit theorem. So, let X 1, X 2 and so on be a sequence of independent and identically distributed random variables. So, basically what we are saying is that we are taking a sample with a large size, so i.i.d random variables with a mean mu and variance sigma square; we assume it to be finite.

So, if you are assuming that say X n bar is the mean of the n observations; then the limiting distribution of root n x bar minus mu by sigma is normal 0, 1 as n tends to infinity; that means, the standardized sample mean has a limiting standard normal distribution. Now if we carefully look at conditions of the theorem, this is pretty general we are not making any assumption on the form of distribution of X i's; all that we are assuming that mean is given and the variance is given. In that case the limiting distribution of the sample mean after a certain change of location under scale is standard normal; provided the sample size is large.

In fact, this is the result which places the normal distribution in the center of a statistical theory, what happens that in practice when we are taking observations or measurements on certain thing, we are usually not taking one observation for example, we may be measuring length of certain article. Suppose it is a physical experiment, so in place of taking one measurement, there is some measuring device and we takes a 30

measurements and will take average of those measurements to say that this is the actual estimate of the length of that equipment.

So, in that case basically what we are using is the x n bar rather than individual x i; the same thing is used at various places for example, if we are looking at average crop per field, then we are not taking individual crops rather then we are taking a sample of the fields and then we take the average; that means, the crops of the individual fields and then e take the average of that.

So, likewise in large number of practical situation; we are interested or we are actually using the mean rather than the individual observations and therefore, the distribution of the sample mean is what should be used and this particular result which is known as the central limit theorem; it says under very pretty general conditions that the distribution is actually normal. Another thing we should notice here that is here we have assumed that the distributions; that means, the random variables X 1, X 2 etcetera are from the same population; that means, it is the sample from the same population.

In fact, this central limit theorem has been further generalized; that means, we may lose the condition of say identically distributed or you may lose the condition of independence also and even then under certain condition; the central limit theorem holds. However, that is part of another study right now we are concerned with this sampling distribution in which case we take X i's to be independent and identically distributed random variables.

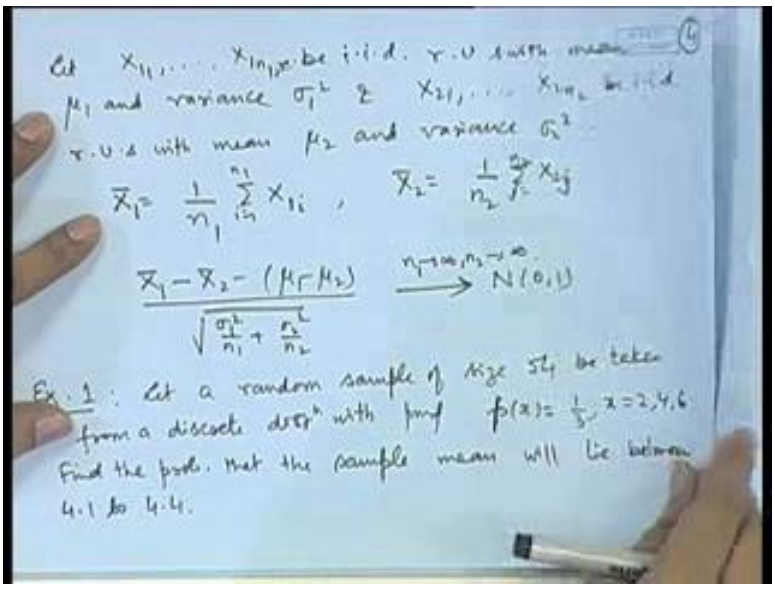
Now, one question may arise at how large n should be such that this approximation is good. So, in practice n greater than or equal to 30 is considered to be large. If the original distribution is normal or it is close to normal then for smaller n itself the approximation may be good. One more point earlier we have seen that binomial distribution was approximated to normal or the Poisson distribution was approximated to normal.

So, that is actually case special case of central limit theorem because what is a binominal random variable, it is the sum of successes in individual trail. So, if you are taking X 1, X 2, X n, so basically it becomes the distribution of the sample sum. So, actually you can write a equivalent form also; suppose I define S n is equal to sigma x i, i is equal 1 to n; then an equivalent form is that if we right S n minus n mu by root n sigma; then this will be converging to a standard normal random variable as n tends to infinity.

So, the binominal approximation to normal is actually a special case of the central limit theorem. Similarly the Poisson distribution approximation to the normal is also a special case here because a Poisson random variable is the number of arrivals. So, if we are looking at the arrivals in individuals instants, how was very small that instant we may choose; then x is denoting the number of arrivals in the full length of the time, which is becoming a sum and therefore, the sum of the observations must follow approximately normal distribution.

Now, in case of one sample we have a straight forwardly for sample mean, suppose we have two samples then the second samples mean may also may have normal and therefore, if we use the linearity property of the normal distributions then the differences etcetera may also follow a certain central limit theorem; let me give a generalization of this one.

(Refer Slide Time: 14:43)

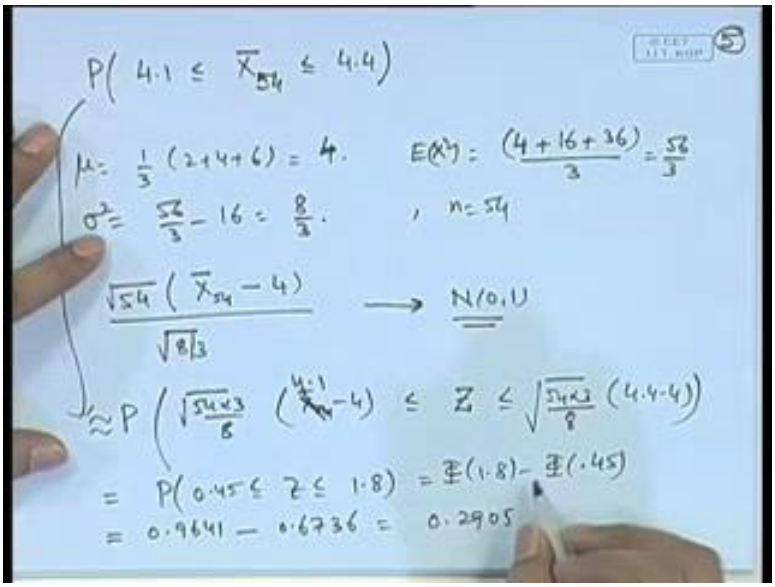


So, let say X 1 1, X 1 2, X n 1 be etcetera, so let me take n only be i.i.d random variables with say mean; mu 1 and variance sigma 1 square and say X 2 1, X 2 2, X 2 n 2; be i. i. d random variables with mean mu 2 and variance sigma 2 square. So, you consider the random variable say X 1 bar, which is actually the mean of the first sample and x 2 bar is equal to say mean of the second sample. Let me put j here and construct the random variable X 1 bar minus X 2 bar minus mu 1 minus mu 2 divided by square root sigma 1

square by n 1 plus sigma 2 square by n 2. Then this converges as n tends to infinity to a normal 0, 1 that is here n 1 tending to infinity and n 2 tending to infinity.

So, this result is quite useful the original central theorem and this results to solve variety of probability problems, where original probability distribution of the sum may be quite complicated but using this we can derive the probabilities. Let me give some examples here, let a random sample of size say 54 be taken from a discrete distribution with probability mass function say p x is equal to 1 by 3 for x equal to 2, 4, 6. Find the probability that the sample mean will lie between say 4.1 to 4.4.

(Refer Slide Time: 17:48)



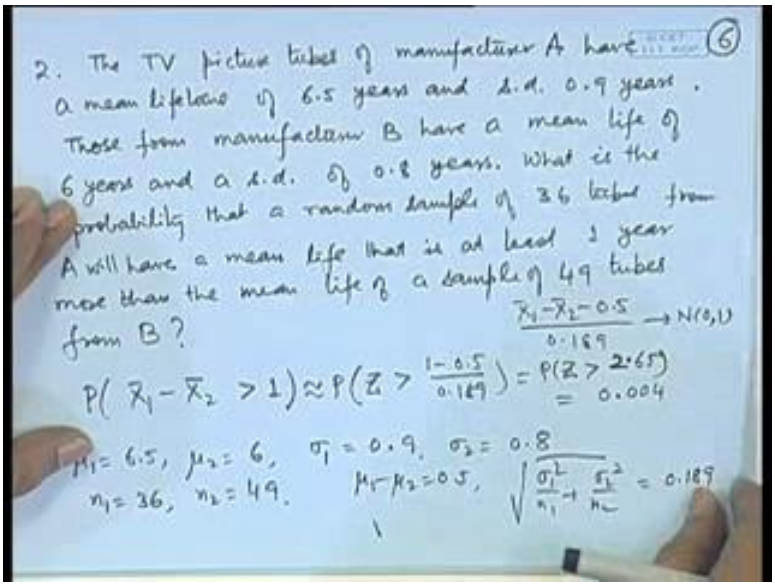
So, basically we are interested to get the probability of 4.1 less than or equal to X n. So, here X n is x 54 bar less than or equal to 4.4, now this is the discrete uniform distribution centered at 2, 4 and 6. So, if you look at the mean of this one mu is 1 by 3; 2 plus 4 plus 6 that is equal to 4 and if you look at the variance. So, we can check say expectation x square that is equal to 4 plus 16 plus 36 by 3; that is 56 by 3. So, variance of x that is equal to 56 by 3 minus 16 that is 8 by 3 and n is here 54.

So, we have the distribution of root 54 into X 54 bar minus 4r divided by root 8 by 3. This will be approximately normal 0, 1, so if you use this property here the probability of X 54 bar lying between 4.1 to 4.4 is approximately same as, so root of; now we may take it to the numerator. So, it becomes root of 54 into 3 by 8; X 54 bar minus 4, so this is 4.1

minus 4 less than or equal to Z. So, approximately root 54 by into 3 by 8 into 4.4 minus 4.

So, if you simply this terms it is probability of z lying between 0.45 to 1.8 which is approximately, so phi of 1.8 minus phi of 0.45, so form the tables of the normal distribution; these values are 0.9641 minus 0.6736 that is equal to 0.2905 that is approximately 30 percent of the time; the sample mean will lie between 4.1 to 4.4. Here we notice that the original distribution is uniform, so the distribution of X 54 bar will be very complicated. We have seen earlier that is some of two independent continuous uniform distributions is triangular distribution. If we take three of the independent continuous uniform distributions, the form is some sort of parabolic in nature. So, if we take 54 such observations and try to find out the actual distribution; that is very complicated and here using the central limit theorem easily, we are getting an approximate value for this and 54 is in fact, a large sample size and therefore, this approximation will be almost quite good.

(Refer Slide Time: 21:12)



Let us take another example the TV picture tubes of say manufacturer A; have a mean life time of 6.5 years and standard deviation say 0.9 years. Those from manufacturer B have a mean life of 6 years and a standard deviation of 0.8 years. What is the probability that a random sample of say 36 tubes from A will have a mean life that is at least 1 year more than the mean life of a sample of 49 tubes from B?

So, here we will apply the extended version of the central limit theorem because we are dealing with the two samples. So, we can consider that X 1 bar minus X 2 bar minus mu 1 minus mu 2 divided by sigma 1 square by n 1 plus sigma 2 square by n 2 will be approximately standard normal distribution. So, here we see that we are supposed to find out the probability of x 1 bar minus x 2 bar greater than 1. Now we look at the parameters here mu 1 is 6.5, mu 2; 6 sigma 1 is 0.9, sigma 2 is 0.8; n 1 is equal to 36 and n 2 is equal to 49. So, if we calculate say mu 1 minus mu 2; that is 0.5 and square root of sigma 1 square by n 1 plus sigma 2 square by n 2; that is equal to 0.189. So, x 1 bar minus x 2 bar minus 0.5 divided by 0.189; it is approximately normal distribution.

So, if you have to calculate this probability; we can approximate it by probability Z greater than 1 minus 0.5 divided by 0.189 that is equal to probability Z greater than plus 2.65. If we see from the tables of the normal distribution, this probability is only 0.004, so the probability that a random sample of size 36 from A will have a mean life at least 1 year more than the mean life of another sample of 49 from B is extremely small.

So, here you see that actually the mean life difference is only 0.5 year, so we are expecting in the sample means to have difference of 1, so the probability of that is going to be very small. Now another point that we notice here is that, if the original distribution itself is normal then square root and x bar minus mu by sigma has exactly a standard normal distribution, so approximation is exact when we have normal distribution. So, today we will stop at this point.