Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 31
The Central Limit Theorem and its Applications, Part 01

In the last class, before I ended, I talked about the essential problem of any measurement process. The problem is that, there is an objective value out there. It can be the mass of an electron which you are trying to measure. An objective value is there and we are trying to measure it by certain number of experimental runs.

Or a situation where a biologist has discovered a new type of insect, and has to specify that by measuring the average body weight. Now, there is an average body weight of that particular insect 'out there', an objective value and we are trying to measure it. Obviously, if we can really collect all the members of that species, measure them, and take the average, it will be the right value, the objective value. But always we are able to collect a smaller number of samples from that species, measure it, and from there we are trying to infer something about the whole species.

In case of a physics measurement also, if we can run the experiment a million times, maybe an infinite number of times, then we will get the mean value as the mass of the electron, but it is not possible to do so. We always collect samples.

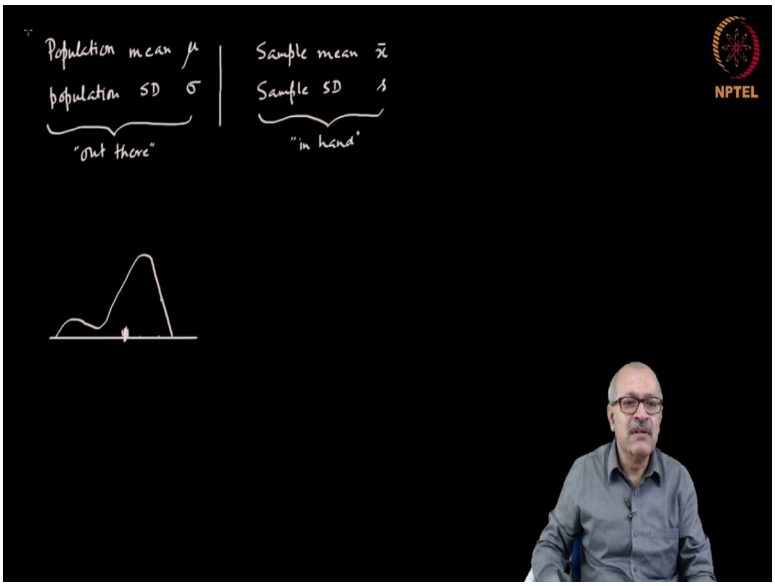
That means, whenever we are doing, say, 10 experimental runs, getting 10 values, and obtaining the mean of that, we are essentially sampling from an infinitely many possible number of readings. We are sampling. So, both in case of a physics measurement as well as in case of a biology measurement there is the issue of sampling. But then the essential question is that, how large should be the sample, so that you can be fairly confident when you are stating the result?

And then, suppose you have taken, say, 10 readings and obtained a mean value. If you now take another 10 readings and obtain another mean value, again you repeat that experiment by taking another 10 readings and obtaining a mean value, all those mean values will not be the same.

And therefore, the question comes: which value do you state as your measured value and how reliable will that measured value be for a third person? Since experiments necessarily need to be reproducible and therefore, we have to state it in a form, so that it can be reproduced by anybody anywhere else in the world.

So, that is another issue. You might want to specify a range within which you are fairly confident that the actual value will lie in that range. But how do you specify that range? You are, after all, taking a finite number of measurements. And if you do specify a range then with what level of confidence are you specifying that range? These are the natural questions that come whenever you are making any measurement.

(Refer Slide Time: 04:26)



So, the effective point is that, there is a population mean mu, and a population standard deviation sigma. These are ‘out there’, existing, and we are trying to measure that. But what we actually have in hand is this sample mean, we had called it x bar, and the sample standard deviation, we have called it s. This is something that we have in hand. And we are trying to make some kind of a confident assessment of these using these. How do we do that?

So, that is the question we are now facing. So, we always take samples, and we have the sample mean and the sample standard deviation in hand.

There is another issue. the issue is that the distribution of values in a population need not always be a normal distribution. For example, if you are talking about a particular species of organisms and you are talking about their average weight, then the distribution of weights might not be a normal distribution. It will depend on the character of the species. It might be something like this. That means, it might peak at some place, there would be nothing, no organism, found in a particular weight, and all that are possible.

But still we would like to obtain some kind of a handle on the population mean and the population standard deviation using the sample mean and the sample standard deviation.

Now consider the issue: Suppose I have taken 10 of these samples and have obtained the sample mean and the sample standard deviation. Then I might get something, say, the mean is here. But the individuals would be distributed all over the place. That means, one individual may have this value of the weight, another individual may have this value of the weight, another, all that taken together when you take the mean, it would be somewhere here.

So, the mean value we have got here. Suppose you take another 10 samples and again obtain the mean, it will again have the individuals distributed all over the place, but the mean will come somewhere close here. And again if you do the same experiment, collecting another 10 samples, it will again come somewhere close here, but not exactly the same. It will be distributed in some way.

What kind of distribution will that be? In case of the population it can have its own inherent species-specific distribution. Or, for example, if you are a geologist trying to measure the average density of the Earths crust, then you might collect samples from various places. In some place the sample might be more dense, in some place it might be less dense. So, it will have a distribution. The distribution represents the characteristic of the material with which the earth’s crust is formed. But every time you make the measurement, that means, take sample from different places, take the average, you will get a value. And every time you repeat that process you will get every time slightly different values.

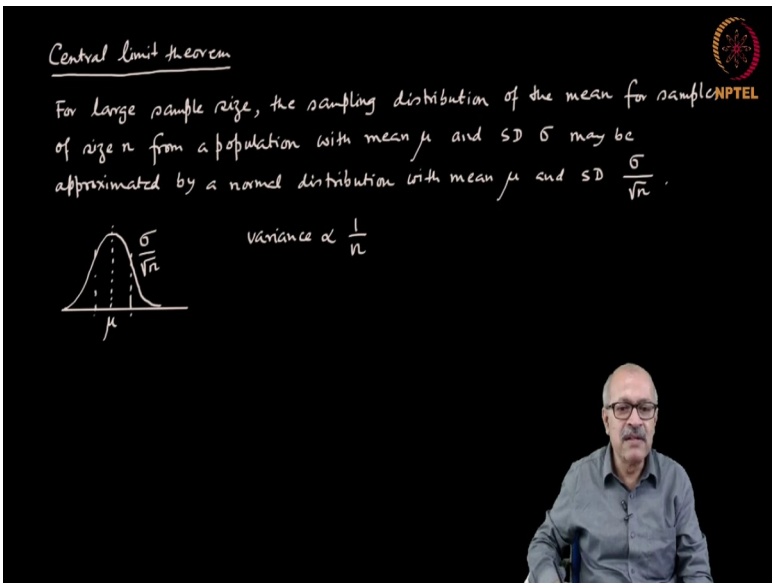
The question is that, if I now talk about how will the *means* be distributed? What will be the distribution like? In case of the measurement of the mass of an electron, every time

you measure, you will get a different value. You cannot help it. There will be some fluctuations from the normal value due to various random errors.

So, there will be different values. Suppose you take 10 measurements and you take the average. You get a value. Again you make another 10 measurements, take the average, you will get a value. Keep on repeating that, you will get a distribution of the mean values. What distribution will it be?

Now, there is a theorem in statistics that asserts that, if the number of samples is bigger than a certain minimum number, then that distribution of the means will always converge on to a normal distribution. Let me state the theorem; It is called central limit theorem.

(Refer Slide Time: 10:29)



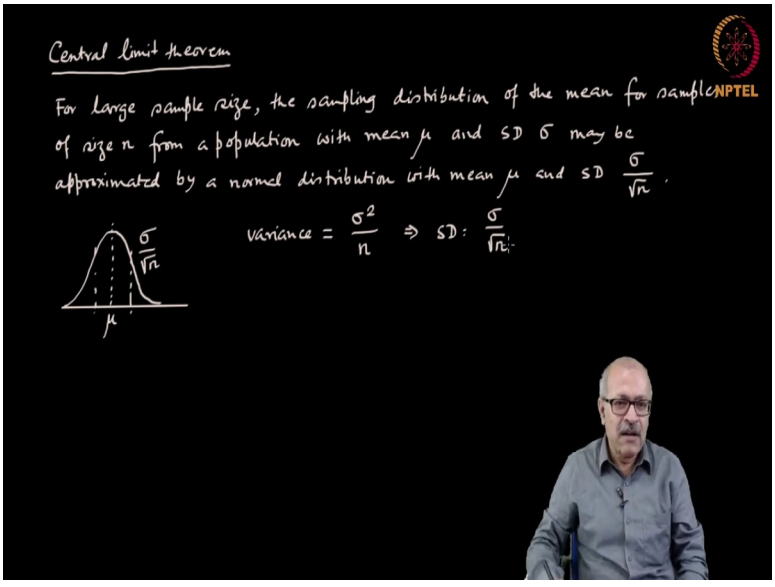
It asserts that the distribution of the means will be a normal distribution. It says that for large sample size, the sampling distribution of the means, for samples of size n from a population with mean mu and standard deviation sigma, may be approximated by a normal distribution. Then if I say it is a normal distribution, I have to specify its mean and standard deviation. It will have the same mean as the population mean mu. The standard deviation will be sigma, the population standard deviation, by square root of n. This is the statement of the theorem.

Now, let us try to understand what it means. It means that, if I repeatedly conduct the experiment, each experiment comprises taking a number of samples and then obtaining

the mean, then if we do that repeatedly and talk about the distribution of the means, the means will be distributed as a normal distribution. It will be distributed as a normal distribution like this with the mean at mu and this standard deviation, if the population standard deviation was sigma, then the standard deviation would be sigma divided by square root of n. Therefore, it depends on the number of samples we take. Stands to reason. The point is that, if instead of 10 samples I had taken 50 samples, it will stand to reason that I will get a more accurate estimate of the population mean. More accurate estimate essentially means that its deviation from the population mean will be small. Now, if I keep on repeating that experiment a number of times, each time I will get a value that will be very close to the population mean mu.

And therefore, it will be a normal distribution, but the normal distribution will have a lesser spread, which means that a lesser standard deviation. So, the assertion is that the variance will be inversely proportional to n. And so or variance will be the variance of the population divided by n, inversely proportional to n, and therefore, this gives the standard deviation as the population standard deviation divided by square root of n. Stands to reason.

(Refer Slide Time: 15:24)



If you take a larger number of samples, it will be more or more accurate representation, or more closely resemble the population mean value, and therefore, the spread of that

normal distribution will be less. This theorem asserts that the variance will be inversely proportional to the number of samples you take, and therefore, the standard deviation would be sigma by root n.

So, that is the assertion of the central limit theorem, what does it give us? Actually the central limit theorem allows us to estimate the confidence with which we can state a value as the measured value.

That means if I state something as a measured value, I need to give some kind of a bound and I need to talk about with what kind of confidence I am talking about that. All that can be obtained using the central limit theorem. I will illustrate that with the help of an example.