

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

**Lecture - 32**  
**The Central Limit Theorem and its Applications, Part 02**

(Refer Slide Time: 00:16)

Central limit theorem

For large sample size, the sampling distribution of the mean for samples of size  $n$  from a population with mean  $\mu$  and SD  $\sigma$  may be approximated by a normal distribution with mean  $\mu$  and SD  $\frac{\sigma}{\sqrt{n}}$ .

variance =  $\frac{\sigma^2}{n} \Rightarrow$  SD:  $\frac{\sigma}{\sqrt{n}}$

$n > 25$

The most important statement of the central limit theorem is the fact that the distribution will slowly become a normal distribution. Now, when you say a large sample size, how large should it be? You can do this numerically. Take a very badly skewed distribution; for example, something like this. If you take a very bad distribution something like this, a distribution like this as bad as this, there can be some samples here, nothing here, some samples there, nothing else. Even for that bad a distribution, if you take 10 samples and obtain the mean, it will be somewhere here, if you take another 10 samples it will be somewhere around here, another 10 samples will be somewhere here. So, you will get some distribution, but it might not be a very good approximation to a normal distribution.

But instead, if you increase the number of samples each time, you will find that as you exceed say 20, it is more and more resembling a normal distribution, and beyond 25, it is practically a normal distribution.

Because of that, we normally take it for granted that beyond 25 if you take further readings, it does not improve the approximation to a normal distribution. So, it suffices to take 25 as the number. So, 'large' essentially means  $n$  should be greater than or equal to 25, at least 25.

And I would strongly advise you to do this experiment numerically once, by taking random samples from this kind of a distribution, and obtaining the mean, and finding how the means will be distributed. You will find that it approximates a normal distribution as  $n$  exceeds 20, and after 25 there is no further significant improvement in the approximation. So, how large? 25. Now, on that basis, let us do an example. With the help of that, you will realize the advantage of the central limit theorem.

(Refer Slide Time: 03:01)

Example  
Suppose there is a population with mean 2 and SD 0.7. Suppose you take 20 samples, what will be the probability that the sample mean will lie beyond 2.2.

Central limit theorem says

$\frac{0.7}{\sqrt{20}} = 0.156$

2

NPTEL

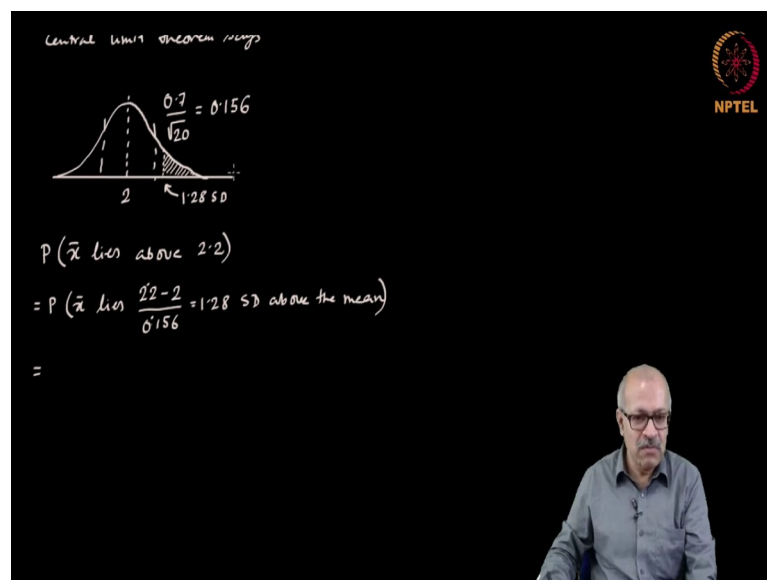
Example: Suppose there is a population with mean 2; I will not specify 2 in units, because it could be any unit. If you are measuring the height of a population it will be in meters, if you measure the weight of a population it will be in kilograms. So, whatever it is, I will specify that as a unit independent number, and the standard deviation is 0.7.

The distribution within this population is unknown. Suppose you do not know that. Now, suppose you take 20 samples out of that population and then again you take 20 samples out of the population. We are trying to assess what will be the probability that the sample mean will lie beyond 2.2.

How do you proceed to attack a problem like this? It says that the population has mean this, standard deviation this. So, if you taken 20 samples you get some  $\bar{x}$ , sample mean. If you again take another 20 samples you will get a different  $\bar{x}$ , the sample mean. If you keep on taking such 20 samples again and again, you will get a distribution of the means and the central limit theorem asserts that, since 20 is close to the necessary number 25, it can be approximated by a normal distribution.

It can be approximated by a normal distribution with the same mean as the population mean 2, and the standard deviation will be the population standard deviation 0.7 divided by square root of the number of samples, which you have taken 20. This comes to be 0.156. So, that will be the standard deviation of the distribution of the means.

(Refer Slide Time: 07:31)



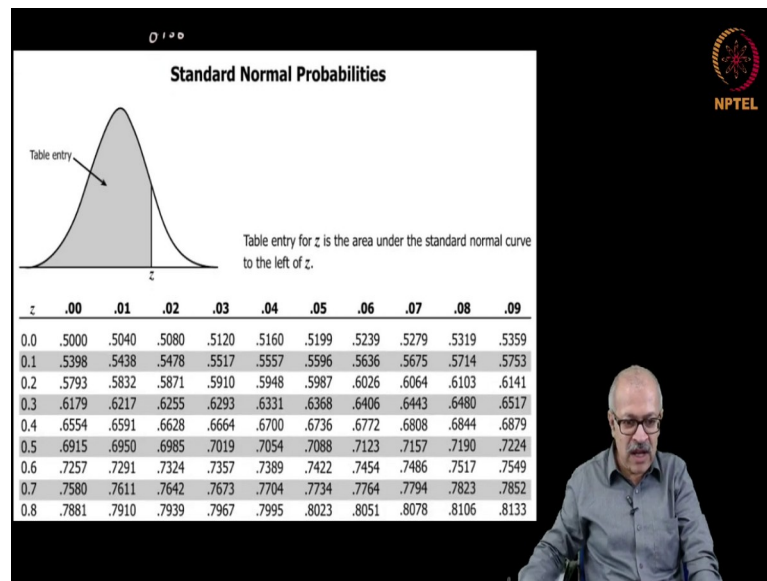
Now, we have to face the question: what will be the probability P that  $\bar{x}$  lies above 2.2? This is question asked. This is equal to probability that  $\bar{x}$  lies above—I will write it a bit differently—2.2: how different is it from the mean 2? This is 0.2. Divide it by the standard deviation, which is 0.156. This will become 1.28 standard divisions. I will write this many standard deviations above the mean.

So, 0.156 is the standard deviation, and 1.28 standard deviations would be somewhere here. So, this is 1.28 standard deviation and we are trying to find out the area under the curve in this range. This is what we have to find out.

Now, this value, the multiplier of the standard deviation, is called the z value. In American pronunciation it will be zee value, but let me go by the British pronunciation or Indian pronunciation: the z value. The z value is the multiplier of this.

Now, people have integrated the normal distribution curve up to certain values and have tabulated that. That is available as the z table and we have to read off the value from there. Let us do that.

(Refer Slide Time: 10:01)



Here I have plotted the z table and this is the normal distribution and here is the z value. What the z table gives is that, the area under the whole curve is 1 and how much of that is contained to the left of that z value. So, that is what is given in the table. So, now, we have to read off the table the area to the left of 1.28, which is the z value.

(Refer Slide Time: 10:39)

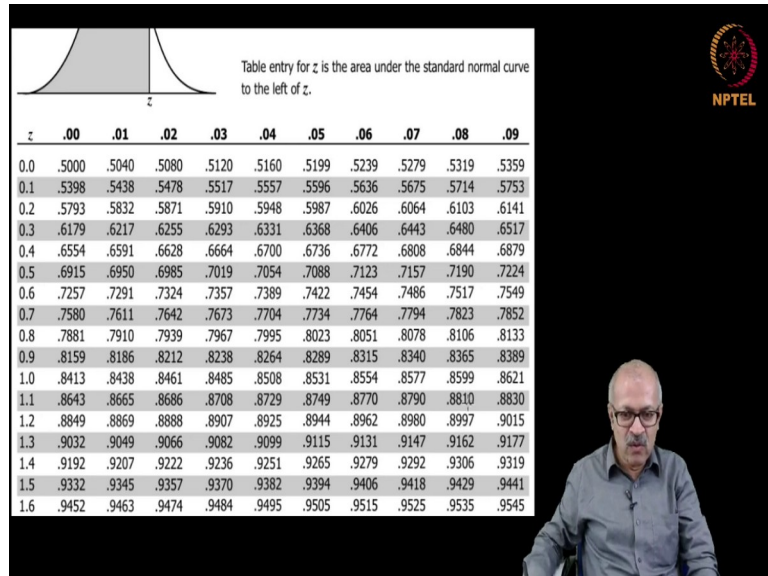
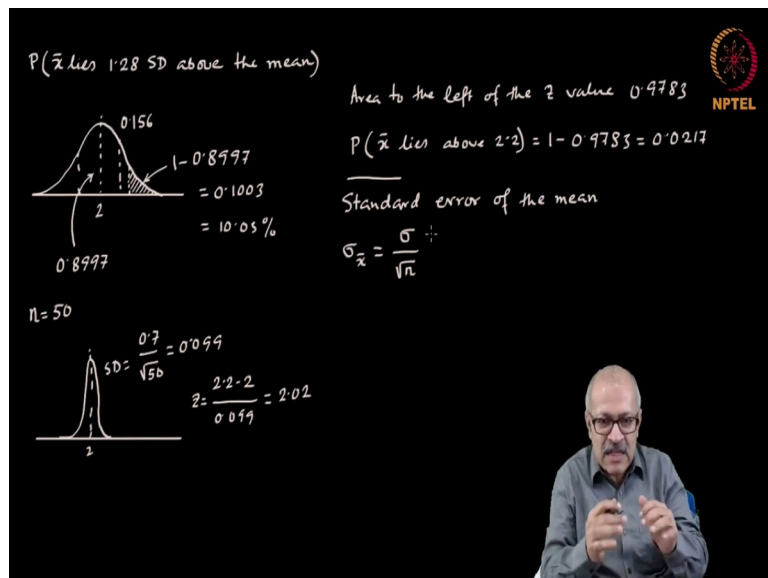


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

So, when we do that, 1.2 is here and we go further to the right, and 0.08 is here. So, we get 0.8997 as the value that we have to use. I will go ahead with that value to the rest of the calculation. So, let me write it afresh.

(Refer Slide Time: 11:09)



$P(\bar{x} \text{ lies } 1.28 \text{ SD above the mean})$   
 Area to the left of the z value 0.9783  
 $P(\bar{x} \text{ lies above } 2.2) = 1 - 0.9783 = 0.0217$   
 Standard error of the mean  
 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$   
 $n = 50$   
 $SD = \frac{0.7}{\sqrt{50}} = 0.099$   
 $z = \frac{2.2 - 2}{0.099} = 2.02$

A probability of  $\bar{x}$  lies 1.28 standard deviation above the mean, this is what we needed to calculate. It is a normal distribution because the number of samples was close to 25, 20 it was. We know that this value is the population mean, which was 2. The standard deviation was 0.156 and divided by square root of n was 1.28.

So, we had to calculate the area here, but we have actually calculated the area to the left of that. So, this area has turned out to be 0.8997. So, this area will be 1 minus 0.8997. It will be 0.1003 which is 10.03 percent or 10 percent approximately, which means that for the problem given, the actual the mean of the population was 2, but by obtaining 20 samples, I stand about 10 percent chance of obtaining a mean which is above 2.2.

So, there is a high risk of committing an error by making the measurement that way. But suppose we increase the number of samples to 50. Suppose we increase the number to 50. Then what happens?

Then we get this curve, a more narrower distribution. It will become a narrower distribution, and in this narrow distribution the mean will be 2, the standard deviation will be 0.7 the original standard deviation, divide by root over 50. Now, this is equal to 0.099.

And, using this if we calculate, if you obtain the z value. Then the z value will become: z is equal to, the way we calculated, 2.2 minus the mean 2 divided by the new standard deviation which is 0.099, then it comes to be about 2.02. For this value if we again refer to the z table, then we find that the area to the left of the z value, in this case becomes 0.9783.

Therefore, the probability of  $\bar{x}$  lies above 2.2 will now become 1 minus 0.9783, is equal to a small number, it will come to be 0.0217, around 2.17 percent. So, the chance of committing an error will be less.

Notice that, all these we are being able to calculate because the central limit theory asserts that the curve ultimately we get for the distribution of the means will be a normal distribution. So, all this could be calculated using that fact.

Now, the value that we get by dividing by square root of the n, that standard deviation of the sample means, is called the 'standard error of the mean'. So, this is a important thing that we get. A standard error of the mean, let us call it sigma. But, sigma of what? Of  $\bar{x}$ , that is our sigma divided by square root of n. This is the standard error of the mean and this gives an estimate of how much error can we commit by actually obtaining samples and talking about the means obtained from the samples. Let us illustrate that again with another example.

(Refer Slide Time: 18:06)

Ex: Suppose you have measured a quantity 36 times and have obtained a sample mean  $\bar{x} = 112.0$  and a sample SD  $s = 40$ . What is the probability that the actual mean  $\mu$  lies in the range  $[100, 124]$ ?

$P(\mu \text{ lies in the range } [\bar{x} - 12, \bar{x} + 12])$

$P(\mu \text{ lies within } 12 \text{ of } \bar{x})$

$= P(\bar{x} \text{ lies within } 12 \text{ of } \mu)$

$= P(\bar{x} \text{ lies within } \frac{12}{6.67} \text{ of } \mu)$

$\sigma_{\bar{x}} = \frac{40}{6} = 6.67$

$z = 1.8$

NPTEL

Now, another example: Suppose, you have measured some quantity 36 times (I am using a square number so that you can obtain the square root easily) and have obtained a sample mean, we are calling it  $\bar{x}$ , you have got it 112. And a sample standard deviation, how much is it?  $s$  equal to 40. Now the question is, what is the probability that the actual mean, the population mean, we called it  $\mu$ , lies in the range between 100 and 124? 100 is this mean minus 12 and 124 is this mean plus 12. So, this is how we have defined the problem.

Notice that, in this case we have only the result of the sample available to us, and we do not know the actual population mean or population standard deviation. In that case can we estimate some range in which the population mean will lie and with which probability will that happen?

Notice again the line of argument. If we repeated the experiment again and again, every time taking 36 samples, then (since 36 is bigger than 25) we will get a normal distribution of the sample means. So, the sample means will have a normal distribution something like this.

This is the mean, not of the sample, but of the population, and it will have a standard deviation which is the population standard deviation divided by root over 36, that is 6. So, this is what we can infer using the central limit theorem. Now, what we are trying to find out is the probability that the actual mean  $\mu$  lies in this range.

So, the probability that  $\mu$  lies in the range: this is 100 (100 means the actual  $\bar{x}$ ) minus 12 to  $\bar{x}$  plus 12. This is what we are trying to find out. Now this distribution we do not know really, but we have to try to somehow obtain the  $z$  value. For that, we need to find out the standard deviation of this distribution. But we do not know this number; this number is 'out there'. This number is of the population, and we have measured only the sample.

But the best we can do under the situation is to use this sample standard deviation as an estimate of the population standard deviation. So, in place of  $\sigma$  (we do not know the  $\sigma$ ), we will substitute the value that you have actually measured, which is 40. So, we get the  $\sigma$  of the  $\bar{x}$  as 40 divided by 6 which is 6.67.

So, we now have to find out:  $\mu$  lies within what range of the standard deviation? P: probability of  $\mu$  lies within 12 of  $\bar{x}$ . Now notice, here is a value  $\mu$  and here is a value  $\bar{x}$ . We have found  $\bar{x}$ , and we are trying to find out how far can  $\mu$  be from  $\bar{x}$ .

Now, this is the same as how far can  $\bar{x}$  be from  $\mu$ . Therefore, we can also write this as the probability that  $\bar{x}$  lies within 12 of  $\mu$ . Now, this is equal to probability that  $\bar{x}$  lies within 6.67 of  $\mu$  and this number is the  $z$  value. In this case it comes to be approximately 1.8.

Again if you refer to the  $z$  table and read out for this number 1.8, the area to the left of that. 1.8 standard deviation will be somewhere here. So, we want to find this area and this side and the other side. We want to find this area, this total area actually. This is in the positive side, this is the negative side.

(Refer Slide Time: 26:21)



$P(\mu \text{ lies in the range } [\bar{x}-12, \bar{x}+12])$   
 $P(\mu \text{ lies within } 12 \text{ of } \bar{x})$   
 $= P(\bar{x} \text{ lies within } 12 \text{ of } \mu)$   
 $= P(\bar{x} \text{ lies within } \frac{12}{6.67} \text{ of } \mu)$   
 $z = 1.8$   
 Area to the left of  $z = 1.8$   
 $= 0.9641$   
 $(0.9641 - 0.5) \times 2$   
 $= 0.9282$

$SE = \frac{SD \text{ of the readings obtained}}{\sqrt{\text{no. of samples}}}$

NPTEL

This area we need to find out. But we have actually found, if we read out from the z table, then the area to the left of z equal to 1.8 is equal to 0.9641. We have to refer to this z table and you have to read this out.

Now notice: here we have read to the left of it. There are two ways of calculating this whole thing. One is 1 minus this thing, you will get this area, times 2 is one possibility. The other possibility is that we have got this area. If you subtract 0.5 from this whole area, then you get only this area. Then you just get the twice of that. Then it will be within this.

So, we need to find out how much area is contained, excluding these two tails, within this area. That can be found this way. I will do it the second way. So, 0.9641 minus 0.5, this whole area minus this half area is this area. Now I have to twice: this into 2, this comes out to be 0.9282. So, the probability that the mu, the actual mean, will lie within 12 of x bar is this 92 percent 93 percent approximately.

This way, even without actually obtaining an infinite number of readings, by taking a finite number of readings, we can infer something of interest regarding the character of the population mean. This method, we will actually carry through in much of the measurement process that we will come across. I will come to that later.

So you see, we have obtained the answers to two questions that we asked right in the beginning of the course. The first question was, how many readings do we need to take

in order to make a confident estimate of the mean? The answer is 25. Now, who gives this 25 number? It is actually empirically obtained.

We have tried it out and found that, that number works in the sense that, by increasing beyond that number we do not gain much. We have ultimately obtained a value which, if repeated many times, will give a distribution of the means which is a normal distribution and from that normal distribution we can extract all the information.

And the most important information we extract is the standard error of the mean. I will just call it SE. Standard error of the mean is the standard deviation of the readings obtained divided by square root of the number of samples.

In case of a physics experiment, the number of samples means how many times you have conducted the experiment and how many data points you have got. So, this is a very important quantifier that we will use many times in the next few lectures. Just keep this in mind. This is the standard error of the mean.

With that we will close today and we will continue with that in the next class.