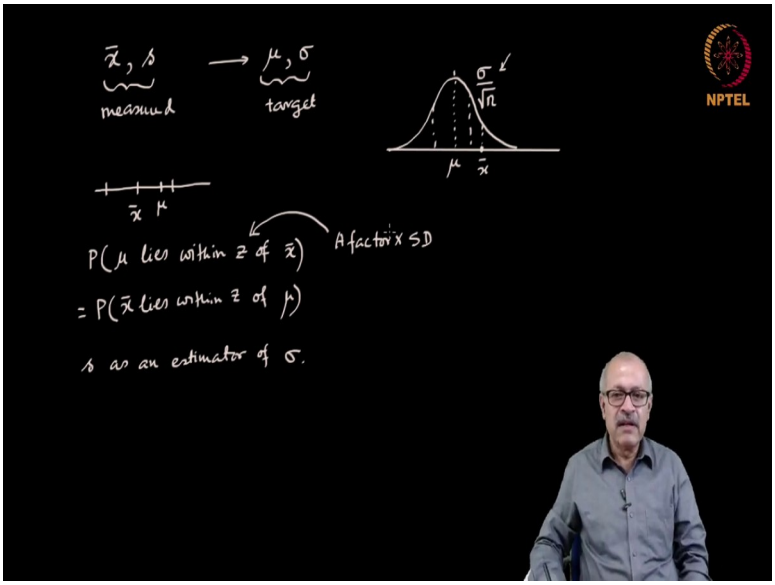
Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 33
Errors Bars and Confidence Interval, Part 01

In the last class we saw that, whenever we make a measurement, we always take a sample of some quantity and then we obtain the mean of that sample we get the sample mean and the sample standard deviation. From there we try to estimate the character of the population mean and the population standard deviation. So, we have a smaller number of data and we are somehow trying to assess the character of the quantity 'out there'.

(Refer Slide Time: 01:15)



And in doing so, we had reversed the argument. For example, we have the measured quantity x bar which is the average from the sample, and the standard deviation obtained from the sample. These are measured. And we are trying to find out the population mean and the population standard deviation. So, this is the target.

Then the central limit theorem stated that, if the number of data points is sufficiently large, then the means will be distributed as a normal distribution something like this. The mean of that normal distribution will be the population mean mu, and the standard

deviation will be the population standard deviation divided by this square root of the number of data points taken. So, that is the claim of the central limit theorem.

Now, what had we done in the last class in the example that we worked out? In that, we were given these and we were trying to figure out if we have the a particular value, suppose this is the x scale and here is my x bar, we are trying to figure out if I can define a range around x bar and can claim that the actual mu will lie somewhere in this range. And in doing so we argued that, if mu is here, then the distance from x to mu is the same as the distance from mu to x.

Therefore, the probability that mu lies within z (we had written it in the form of a multiplier times the standard deviation, z is the multiplier) of x bar, and then we said that this is the same as P of x bar lies within the z of mu.

This z is a factor times the SD. The factor is z. So, that is the factor and this is the multiplier of the standard deviation. Now, if we express it this way, then it is possible to obtain it because, for each value of z—for example, if the z is here—the area under the curve to the left of z can be found from the z table. From there we can solve the problem.

So, this z value is a factor. This is a multiplier of the standard deviation, something times the standard deviation. So, if x bar is at a distance z from mu, then it is possible to find out what is the area under the curve to the left of that value of z. From that, we were able to calculate the probabilities, and hence the confidence with which we can state that mu will lie within a certain range of x bar.

You will notice one thing: what is the standard deviation here? It is the standard deviation of the distribution of the means, which is this. In that, we do not know sigma because that is a target. That is something that we do not know. That is of the population out there.

So, we argued that in the absence of the value of sigma, we estimate it by the measured value of the standard deviation, which is s. So, we argued that, we will use s as an estimator of sigma. You might ask how logical will that be? Will not that incur errors in the calculation, in the measurement? Yes it will. But there is a logic behind this substitution.

The logic is that, if the number of samples is reasonably large, then we have seen that the central limit theorem claims that the distribution of the means will be almost a normal distribution. And if you increase the number of samples even more, it will not improve the approximation any further. So, we know that around 25 is a good number of readings to take.

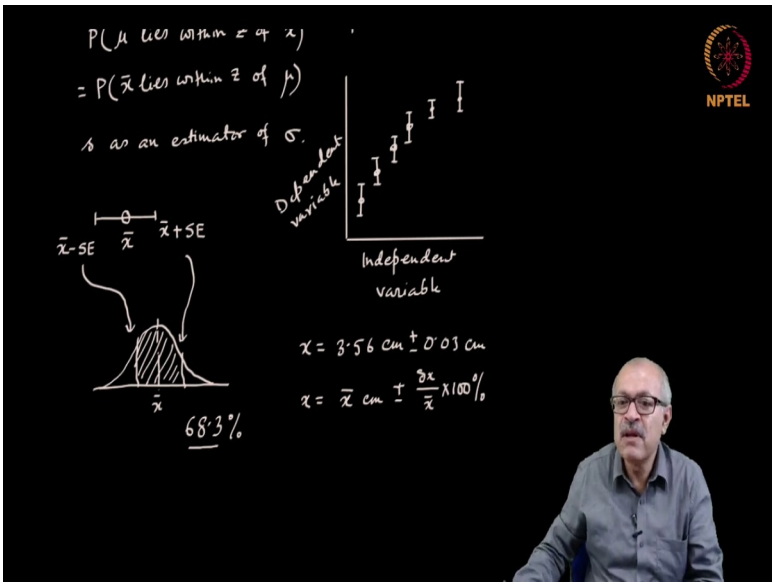
Now, if we take that minimum number of readings, then the theory in statistics (which will not get into the details of) shows that if the number of samples is reasonably large, then the difference between s and sigma will be really small.

Since the actual value is sigma divided by square root of n, and n will be a reasonably large number, the difference of sigma by square root of n and s by square root of n will not be significant. That is why it justifies the use of s as an estimator of sigma.

But, one thing is clear: that we cannot do anything otherwise. We do not have a handle on sigma. We only have s and therefore, we have to use s as a estimator of sigma. So, we will then substitute sigma by s which is known, n is known and therefore, we have a handle on the standard deviation of this curve.

If we know that, then what factor needs to be multiplied with the standard deviation to get the value, that is also known. Therefore, we can then refer to the z table to extract the value of the probability.

(Refer Slide Time: 10:12)



Now, I plot that as a number here. This is the value that I have measured, which is x bar, and suppose here is x bar minus the standard error of the mean and this is x bar plus the standard error of the mean.

Then I know that in the curve that we had already drawn here, I have this normal curve, and we are talking about a range which is x bar and this value is here and this value is here. So, it is basically within one standard deviation, and we know that the area under this curve here is 68.3 percent of the whole. What does that imply? It implies that if I define a range which is x bar minus the standard error (standard error means the standard deviation divide by square root n) and x bar plus the standard error.

If this range is defined, then I can be certain that the actual mean will lie within this range, and I can state that with a confidence of 68.3 percent. This is important. It is important to understand what is the actual meaning of the statement. This, in many cases, is written as the error bar.

You we will see graphs something like this. These are the data points. If you add the error bars these will look like this. You normally put the extremities like this. So, this is how the graphs are actually drawn. Here is a parameter or an independent variable, I would rather say a variable, and here is a dependent variable.

Now, for each independent variable, you measure the dependent variable and you may get these values. But you will also specify an error bar. What does that error bar signify? It signifies that you are 68.3 percent confident that the actual value of the dependent variable for that independent variable will lie within this range.

So, with 68 percent confidence you will state. That is the meaning of the error bar. Now, you might say that 68.3 percent is not a very large extent of confidence. Yes, surely, we will have to deal with that. We would see how to represent a higher level of confidence. But in general, the meaning of the error bar is that. So, whenever you will read an error bar, you see an error bar in a paper, you have to interpret that accordingly.

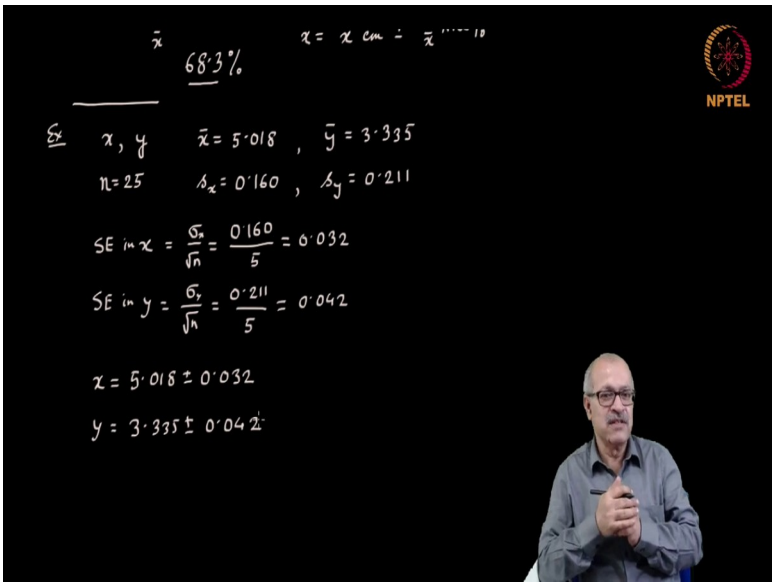
It does not mean that the actual value will be in this range. The actual value can as well be outside, because this confidence level is not very large. That means, essentially you are stating that there is a 68 percent probability that the actual value that we are trying to measure, which is out there, will lie somewhere in this range.

Sometimes we state the value, for example, suppose we have measured something and we have got x equal to, say, 3.56 centimeters. Then we will have to state it with an error bar, which is plus minus, say, 0.03 centimeters. We always state it like that, because we can never state the value exactly. Because we know that it is subject to some random errors.

Sometimes these are also expressed a little bit differently, as a percentage error. That can also be expressed as a percentage error, something like this: I will write x as the average value that we have measured. This is in centimeters. And then plus minus, now you have to put the standard error, which I can write it as delta x, the change in x divided by your x bar, which is the extent of error in fraction into 100. This is the percentage error calculation. You might state the result in both these ways. This is in absolute value and this is as a percentage error.

So, the point that I am making is that, whenever you make an observation in measurement, you state the result of the measurement always this way, and the extent of this error that you state is also objectively calculated. Let me just illustrate that by means of an example and then I hope that will be clearer.

(Refer Slide Time: 17:43)



Suppose you have made a measurement of two variables x and y, and you have made a large number of data points, say n equal to 25. 25 data points you have calculated and then from there you have found by calculating that x bar is 5.018 and y bar is 3.335.

But, the data that you have used: this 25 data points for x, 25 data points for y, from there you can also calculate the s of the x variable; s is the standard deviation of the x variable. Suppose you know how to do that. I have already said how to do that. You have found that to be 0.16. And the s of y is, say, 0.21.

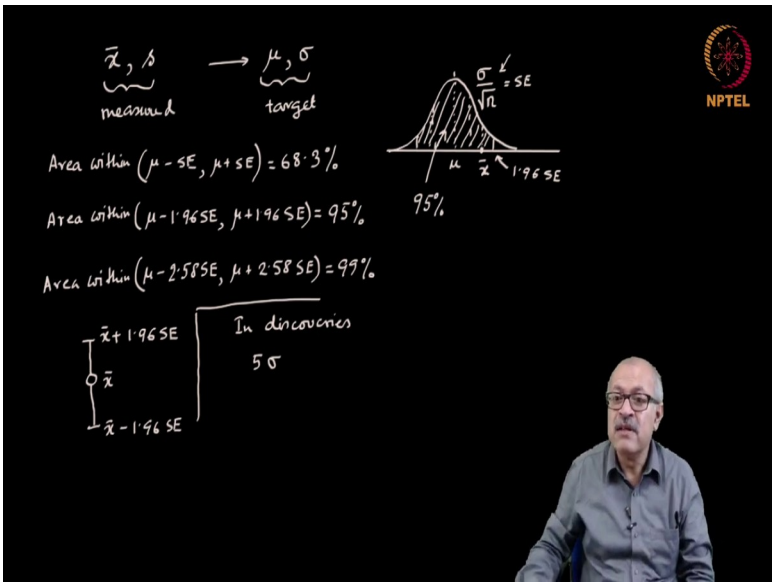
Then how do I specify these values? We need to calculate the standard error. So, standard error in x will be sigma by square root of n. Now, sigma we do not know, but we know the s. Therefore, we substitute by that. So, 0.16 divided by square root of n, is 5, is equal to 0.032. The standard error in y is this sigma. This would be x and this would be y root over n.

And we again substitute that by this standard deviation in y. So, 0.211 by, again, 5 because the number of data points were the same. Then this comes to be 0.042. And therefore, having done this calculation, we will state that I have measured x as 5.018 plus minus 0.032 and I have measured y as 3.335 plus minus 0.042.

Notice one thing, that after having these values that you have actually measured, when you calculate the mean, you could have calculated up to a larger number of decimal points. Similarly, for the standard deviation you could have calculated to a large number of decimal points. If you have a calculator with 8 digits, you can do that. But it will make no sense because the measurement has been taken with some kind of apparatus which has a least count and it makes no sense to specify something to a least count that is below that least count.

So, if the measuring apparatus has an accuracy which is meaningful to the third decimal place, you should specify everything only up to third decimal place. There is no point going beyond that. That is why, in this case I have expressed everything up to the third decimal place. But remember to what decimal place you will specify it. There is no pre-assigned prescription for that. You have to do that depending on the instrument that you use. Depending on the accuracy of that instrument.

(Refer Slide Time: 22:12)



Notice that everything hinges on the idea that in the distribution of the means, the standard error of the mean or the standard deviation of this graph is sigma by square root of n. And I have already told you that you can easily integrate the normal distribution curve and find out how much area is contained within some specific ranges, specific limits.

If it is up to the standard division of this graph, then it is 68% of the area under the curve. So, area within mu minus standard error to mu plus standard error, this range, is 68.3 percent. If you can calculate this, then you can also calculate how much does it have to be taken so that the area under the curve is 95 percent, and that has been calculated.

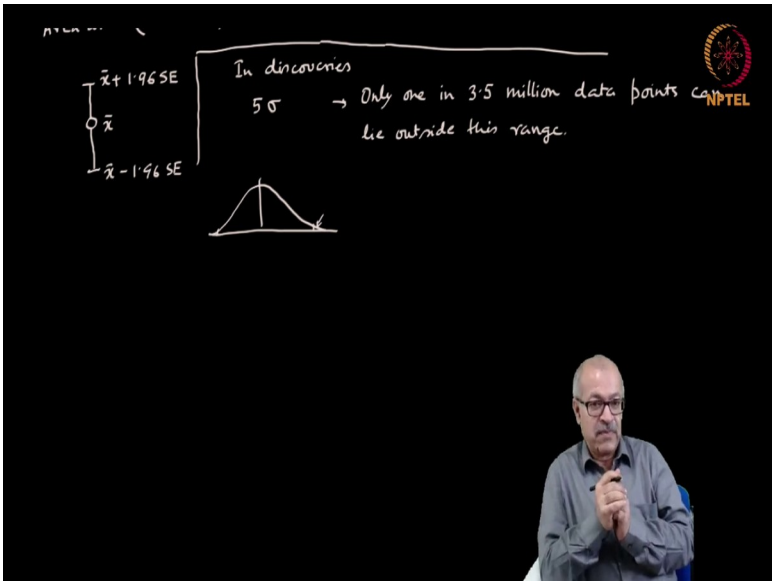
Area within 95 percent is mu minus 1.96 standard error to mu plus 1.96 standard error. This range is actually 95 percent. And if you want to have 99 percent, then mu minus 2.58 SE and mu plus 2.58 SE. So, if you consider a larger range, which is 1.96 standard error, then the area that is contained is 95 percent. This has great importance.

In some fields the confidence level of 68.3 percent is considered too low and there the demand is that you have to state it with a confidence level of 95 percent. And if that is so, your error bar will have: this is the mean value that you have calculated and this is x bar plus 1.96 standard error and this is x bar minus 1.96 standard error.

In that case you have to specify that way and this is actually true for many fields. But there are some fields that demand even more level of confidence when you state the result, for example, 99 percent. In that case you have to put 2.58 here.

There are some fields, like the discovery of a new particle in particle physics, the discovery of gravitational wave, etc., in case of such discoveries the demand is far larger. The demand is something that is stated as 5 sigma. What does 5 sigma mean? In discoveries, when we will say something has been discovered, the demand is 5 sigma. What is the 5 sigma? It is this number. This number will then become 5. That means, almost the whole area is enclosed within 5 sigma: 5 times the standard error of the mean and practically the whole area is enclosed.

(Refer Slide Time: 28:11)



What does this physically mean? It physically means, 5 sigma will mean, that only one in 3.5 million data points can lie outside this range. It means that if the result that is been obtained, for example, the discovery of Higgs boson was an observation where there was a kink in a graph, if that kink happened due to random chance event, then the possibility of such a random event being detected is one in 3.5 million. Only when we reach that level of confidence we say that we have made a discovery. So, it is very very exacting demand for the scientific community. They have to make measurements and repeat it a large number of times and then only one can reach that 5 sigma level of confidence.

That means, if the observation that has been observed is caused by something that is not what is intended to be, then the probability of that happening is only one in 3.5 million. That means, the area under the curve within 5 sigma is almost the whole area. The amount that is remaining is only this much. So, that is a kind of demand in case of discoveries.