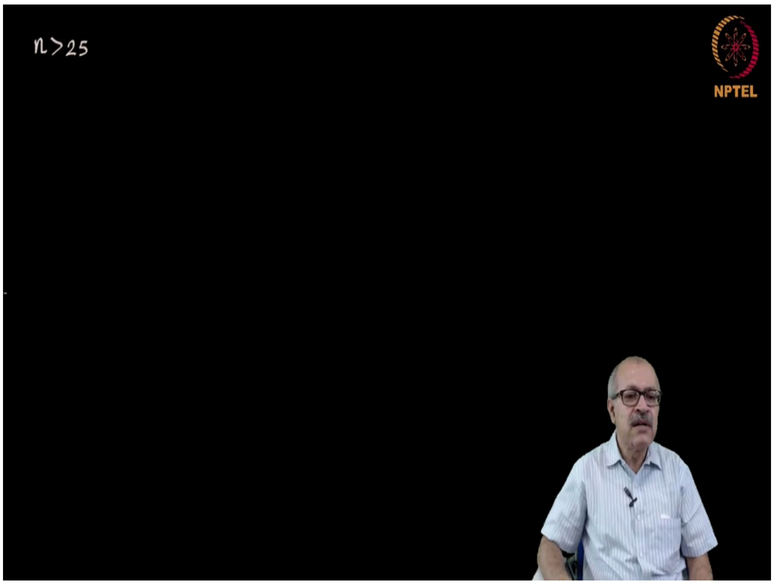
Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 37
Examples of Proportion Measurement

In the last class you may have noticed that, the processes of making a measurement of a value and the process of making a measurement of a proportion are somewhat different. At least to the extent of how to record the result, how to state an error bar, how to ensure that the distribution of the means will be a normal distribution – these are somewhat different.

Let me just enumerate the differences. In both cases our argument is that, I have taken a number of readings and then I have obtained a mean, I have obtained a standard deviation. And then from there, can I estimate the character of the population? That was the essential logical structure. We argued that if I keep on making similar sets of measurement, say 25 measurements in each set, each time I will get a different mean value. But over a large number of such repetitions of the experiment, these will be distributed as a normal distribution provided n, the number of readings, is greater than 25.

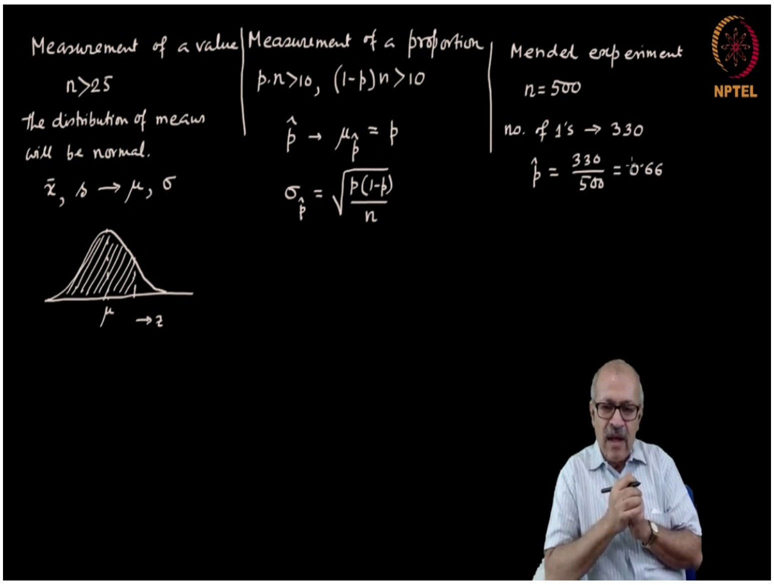
(Refer Slide Time: 01:41)



$n > 25$

I will write the two separate cases. First is the measurement of a value. And then we have talked about measurement of a proportion.

(Refer Slide Time: 01:51)



Measurement of a value

$n>25$

The distribution of means will be normal.

$\bar{x}, s \rightarrow \mu, \sigma$

Measurement of a proportion

$p.n>10, \ (1-p)n>10$

$\hat{p} \rightarrow \mu_{\hat{p}} = p$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Mendel experiment

$n=500$

no. of 1's → 330

$\hat{p} = \frac{330}{500} = 0.66$

In this case, so long as n is greater than 25, the distribution of sample means would be normal. Then, from the data, we have the x bar and we have the value of s, the sample standard deviation. From there we can estimate the value of the mu and the sigma, because the means will be distributed as a normal distribution.

The means will be distributed as a normal distribution, and here is the mu. Since it is a normal distribution, we know the area under the curve up to any particular value of z. Up to a particular z value we know the area. These are tabulated in the z tables: the area to the left of that particular value of z, and from there we have done various problems.

In case of the measurement of a proportion, we have seen that if the proportion out there is p, then the requirement is that p times n, the number of measurements, should be greater than 10, and the other one 1 minus p times n should be also greater than 10. Now this does not follow from the centre limit theorem. This follows from actual empirical observation that, if the proportion in the population is small, for example, say 0.1, then in the distribution you will get the peak at 0.1. And in order for you to get more or less a normal distribution around that value of 0.1, you will need to have a larger number n. So, if p is small or if 1 minus p is small, then you will need to take a large number of readings, a large number of samples. But if it is somewhere in the middle, evenly distributed between the two possibilities, 1 and 0, then you might get away with a relatively smaller number.

In the last class, we did a problem that showed us that the number is much larger than what you might get an idea from the measurement of a value. It is not 25. It is much more than that.

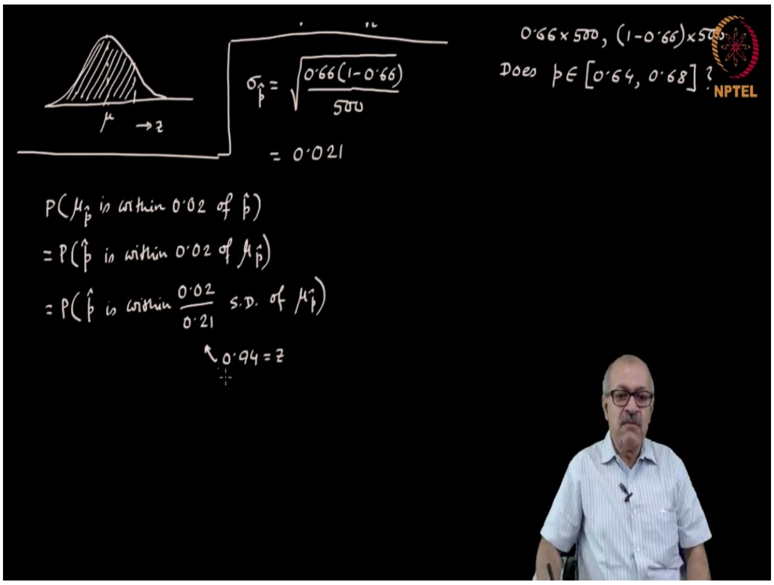
In fact, the problem that we do did in the last class, there we saw that we had taken 50 readings and from there we estimated the value. The confidence with which we can estimate and state that value was of the order of just 23 per cent which is unacceptable. We have also learnt that, if we make the measurement of a proportion, and if we get a value, say the we get a value of p hat, the measured proportion, if this is that, then the mean value of various such measurements will become the actual population mean of the proportion. And the standard deviation of these measurements of p hat, we found that would be square root of p times 1 minus p divided by n. So, using that, we were trying to calculate. These three are the major results that we have used.

Now in the last class we used a number of n, 50, and we found that it is insufficient for our purpose. Now suppose we increase the number to 500, i.e., 10 times. Let me now state the problem, so that it becomes easier for us to actually solve it.

We are basically doing the Mendel experiment. And this time I am stating that we have taken n is equal to 500, we have taken 500 readings, and in that 500 readings the number of ones, I would assume that to be exactly the same, which is 330. So that, the estimated value of p is 330 by 500 is equal to 0.66.

So, the measured value of p is the same as in the last problem that we did. We are assuming that. But let us see if we can now state the result with a larger level of confidence.

(Refer Slide Time: 09:20)



$0.66 \times 500, (1-0.66) \times 500$

Does $p \in [0.64, 0.68]$?

$$\sigma_{\hat{p}} = \sqrt{\frac{0.66(1-0.66)}{500}}$$

$$= 0.021$$

P($\mu_{\hat{p}}$ is within 0.02 of $\hat{p}$)

= P($\hat{p}$ is within 0.02 of $\mu_{\hat{p}}$)

= P($\hat{p}$ is within $\frac{0.02}{0.21}$ s.D. of $\mu_{\hat{p}}$)

↖ $0.94 = z$

In the present case, what is the sigma of the measured value? I have made a measurement and I have got a specific value of the proportion as measured from the samples. Now if I keep on taking 500 samples again and again from that population, that means, I have cultivated these tall and short pea plants and I have randomly selected 500 of them. I have measured, and found that 330 are tall.

And now I have repeated the experiment by choosing another 500 samples and counted. I may not get the same number. A third time I may not get the same number. So, it will be different numbers, which will have some kind of a distribution. That distribution will become a normal distribution provided these conditions are met.

Let us check quickly check whether these conditions are met or not. The measured value of p is 0.66 times the number n is 500 is definitely much higher than 10. Both these numbers are much higher then 10. Definitely it will be a normal distribution. So, we can go ahead with a normal distribution curve like this. We know that this number will be, in the present case p is something we do not know, that is the proportion in the population out there.

But we have made a measurement, which indicates it will be of the order of 0.66. So, in absence of a measurement of the actual value of p, we estimate it by 0.66. So, it will be 0.66 times 1 minus 0.66. This was the same as in the last days problem. But now in the denominator it will be 500. So, as a result this will turn out to be 0.021, which is different from what we got in the last days problem, because the denominator is larger now.

Now, what do we need to do? We are trying to find out: does p belong to the range 0.64 to 0.68? That was the problem we set the last time. So, then p is the mean. Does mu, the mean value of p hat lie within 0.02 of this mean value? So, within 0.02 of the mean value p hat: this is what we are trying to figure out. What is a probability that we will find the actual population mean, which is the mean value of repeated measurements of 500 samples, will it be within this range? What is the probability of that? Now, that we will write as the probability. We will now reverse the argument. We were asking if the mean is within this range of p hat. Now we will reverse the argument and say that the p hat is within 0.02 of the mean.

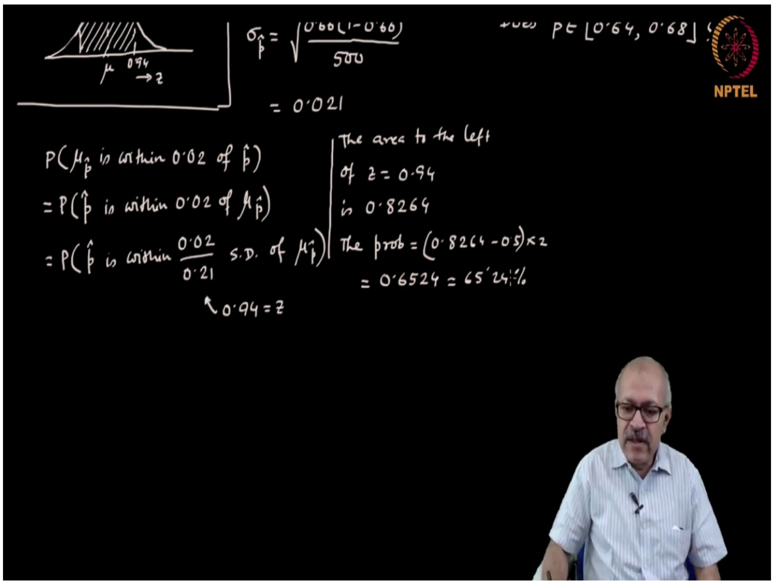
$$\begin{aligned}
&P(\mu_p \text{ is within 0.02 of } \hat{p})\\
&= P(\hat{p} \text{ is within 0.02 of } \mu_p)\\
&= P(\hat{p} \text{ is within } 0.02/0.021 = 0.94 \text{ standard deviations of } \mu_p)
\end{aligned}$$

Now, this is something which is the midpoint here. We know the distribution because we know the standard deviation of the distribution. So, we can calculate that, provided this is expressed as a z value. So, let us do that. Probability of p hat is within ... now we have to say how many standard deviations, so we have to divide this by this standard deviation 0.21 standard deviations of mu p hat, and this turns out to be 0.94.

So, this is the z value. We now have to consult the z table with this value in mind, and then we will have to calculate whether it gives any better results.

So, here the z value is 0.94 and we will calculate the area to the left of it. Now the area to the left of it for 0.94 is, the area to the left of z equal to 0.94 is 0.8264. Now we need to calculate this area. But we have calculated this whole area.

(Refer Slide Time: 15:44)



$$\sigma_{\hat{p}} = \sqrt{\frac{0.60(1-0.60)}{500}}$$

$$= 0.021$$

$p \in [0.64, 0.68]$

$P(\mu_{\hat{p}}$ is within 0.02 of $\hat{p})$

$= P(\hat{p}$ is within 0.02 of $\mu_{\hat{p}})$

$= P(\hat{p}$ is within $\frac{0.02}{0.21}$ s.D. of $\mu_{\hat{p}})$

$\nwarrow 0.94 = z$

The area to the left of $z = 0.94$ is 0.8264

The prob $= (0.8264 - 0.5) \times 2$

$= 0.6524 = 65.24\%$

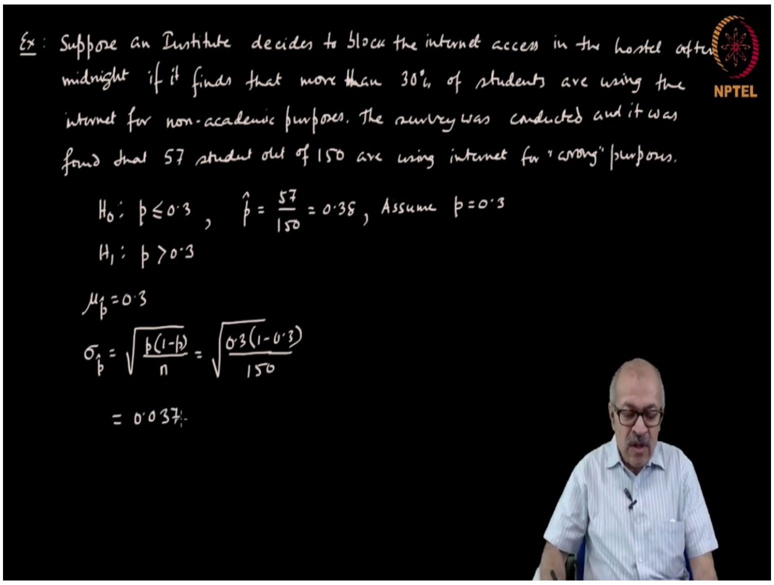
So, how do we calculate this area? We will subtract 0.8264 minus 0.5, then we have got this area minus this area. So, we have got only half this area, times 2, and that comes to be 0.6524, which is around 65.24 percent.

So, notice that we still do not have the level of confidence that we wanted to reach. We wanted to reach a level of confidence at least 95 percent. It is still not that, even though we have taken 500 readings. Now this is a lesson that, whenever we are trying to make a proportion measurement, the number of samples that we have to take is much more. Definitely not 25, even 500 does not suffice, and this is important.

So, just being able to approximate it by a normal distribution does not suffice. Still we get a low level of confidence and the way to improve the level of confidence is to take a larger number of samples and from there obtain the proportion.

In order to get you somewhat comfortable, let us do one more problem so that we can then go ahead with it.

(Refer Slide Time: 18:49)



Ex: Suppose an Institute decides to block the internet access in the hostel after midnight if it finds that more than 30% of students are using the internet for non-academic purposes. The survey was conducted and it was found that 57 student out of 150 are using internet for "wrong" purposes.

$H_0: p \leq 0.3$, $\hat{p} = \frac{57}{150} = 0.38$, Assume $p = 0.3$

$H_1: p > 0.3$

$\mu_{\hat{p}} = 0.3$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(1-0.3)}{150}}$

$= 0.037$

This time I will do a little faster. The question is, suppose an institute decides to block the internet access, that means the Wi-Fi connection, in the hotel after midnight, if it finds that more than 30 percent of students are using the internet for non-academic purposes and are losing sleep.

This was actually observed in our institute, so that is why I am posing this problem. Many students do various things on the internet, I would not like to explain what these are, but in any case, they were losing sleep and the next morning they were not attending classes because they are asleep. So, in order to avoid the problem, suppose the Institute decides that, if more than 30 percent students are doing that, then they will block the internet connection after midnight.

So, how do you decide 30 percent? You go to a number of students, you sample a number of students, and then find out how many are engaging in this kind of activities. Suppose the survey was conducted and it was found 57 students out of 150 are using internet for wrong purposes.

The question is: does the data provide sufficient evidence on the basis of which internet access can be blocked? A sufficient evidence means 95 percent confidence. Can we state with 95 percent confidence that the number of students engaging in this kind of practices is more than 30 percent? That is the question.

So, in this case we are making a statement that more than 30 percent students are using internet for a wrong purpose. Let that be the hypothesis and the null hypothesis would be the less than 30 percent are doing that. So, the null hypothesis is that the probability is less than 0.3 and the alternative hypothesis is that probability is greater than 0.3.

Whenever we encounter a problem like this, we always root ourselves in the null hypothesis and then check whether the data provided provides sufficient reason to reject the null hypothesis. Always we go that way. So, in that case we believe in the null hypothesis: p, the probability, is less than or equal to 30 percent or 0.3.

So, if this is true, then what is the probability that I will get a result like this? What I have got, the proportion measured, is 57 by 150 is equal to 0.38. I have got this, and what is the probability that I can get this even if the null hypothesis is true? It is possible to get because every time you sample you will get a different number, so there will be a distribution and what is the probability that I can get this?

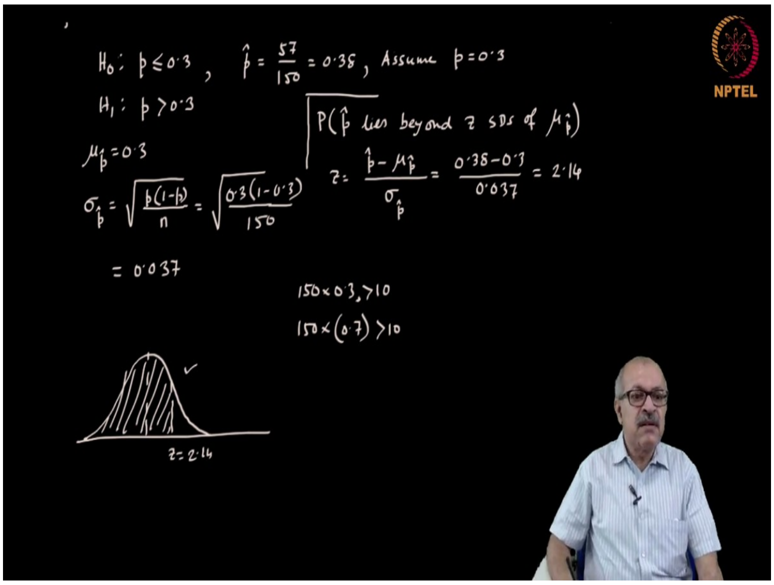
If that probability is less than 5 percent then there is enough evidence that this is probably true; that means, the alternative hypothesis is true. So, in order to proceed we argue that we take the maximum value that is allowed by the null hypothesis, which is 0.3. So, if the mean is 0.3, then what is the probability of getting 0.38? If the mean is less than that, it will be even larger.

So, what is the probability? We assume that the p is equal to 0.3 and then we ask what is the probability of finding a value of p hat 0.38? In order to do that, we have to calculate the mean, we have already found the mean.

So, mu p hat is 0.3 and the standard division of p hat will be square root of p 1 minus p divided by the number of samples n. Now p is something that we have assumed to be this. So, 0.3 1 minus 0.3 by 150 and this turns out to be 0.037.

We are now asking the probability that p hat lies beyond. In that case, let me calculate this: z standard deviations of the mean. So, what will z be? z equal to, in that case it will be, the p hat minus the mean divided by the sigma, this is 0.38 minus 0.3 divided by this is 0.037 is equal to 2.14.

(Refer Slide Time: 26:22)



$H_0: p \leq 0.3$, $\hat{p} = \frac{57}{150} = 0.38$, Assume $p = 0.3$

$H_1: p > 0.3$

$\mu_{\hat{p}} = 0.3$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(1-0.3)}{150}}$

$= 0.037$

$P(\hat{p}$ lies beyond z SDs of $\mu_{\hat{p}})$

$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.38 - 0.3}{0.037} = 2.14$

$150 \times 0.3 > 10$

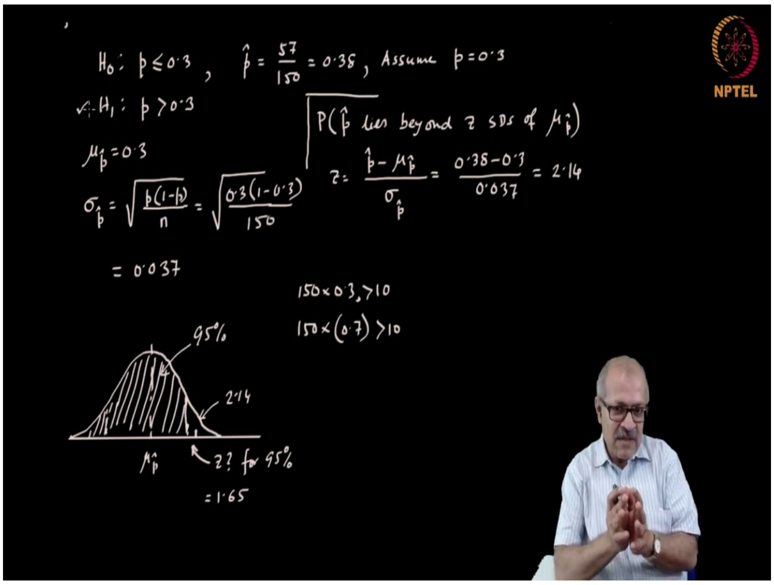
$150 \times (0.7) > 10$

So, p hat lies beyond 2.14 standard deviations of this. Suppose this is my z value, is equal to 2.14. The normal distribution curve will give me this area. The question now is: will it be a normal distribution? Again we have to check: n 150 times the p into 0.3 this is definitely greater than 10, and 150 times 1 minus p 0.7 is definitely greater than 10. Therefore it is a normal distribution.

So, if it is a normal distribution, then we are trying to check whether we can state with 95 percent confidence that the null hypothesis is true or the alternate hypothesis is true.

I will slightly change the line of argument. This is our value of mu. Now we will say: what should be the value of z, so that the area enclosed is 95 percent? What would the value of p, so that the area to the left of it is 95 percent? So, we are asking z for 95 percent.

(Refer Slide Time: 29:14)



$H_0: p \leq 0.3$, $\hat{p} = \frac{57}{150} = 0.38$, Assume $p = 0.3$

$H_1: p > 0.3$

$\mu_{\hat{p}} = 0.3$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(1-0.3)}{150}}$

$= 0.037$

$P(\hat{p}$ lies beyond z SDs of $\mu_{\hat{p}})$

$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.38 - 0.3}{0.037} = 2.14$

$150 \times 0.3 \geq 10$

$150 \times (0.7) > 10$

Now, that value of z turns out to be 1.65. This is what comes from the z table. You just consult the z table for 95 percent and you get 1.65. So, to the left of 1.65 is the area which is 95 percent. And the value that we have calculated is beyond that: this is 2.14, which means the area to the left of it is even more and the area to the right of it is even less.

So, had the z been 1.65 then the confidence level would be 95 percent. The odds, the probability, of finding a value as big as 0.38 would be less than 5 percent if the z value were 1.65. Now the z value here is even more, which means the probability of finding this proportion is even less.

So, we have sufficient evidence to reject the null hypothesis and converge onto the alternative hypothesis. Remember, you cannot say that the set that you have collected, that 150 students, that I have sampled—this is not the right sample, I need to change the sample—that you cannot say, because if you have taken a sample really randomly. Then if you take it again, it will be a different value.

One very common mistakes that one does is that, if you are inclined to believe something, then if you get something at odds with that belief, you might say that, no, this sample is not good, I will take another sample, something that, sort of, conforms to my belief, and then we will go ahead. That is the wrong approach.

I have randomly done it, blindly done it. Whatever value I get from there, I have sufficient evidence that the null hypothesis can be rejected. The alternative hypothesis is true and then we have to block the internet. I will come to this hypothesis testing issue a little later in more detail, but presently I just wanted to show you how to use this proportion measurement issue.