

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

Lecture - 36
Measurement of a Proportion, Part 02

(Refer Slide Time: 00:17)

Measurement of a proportion

n samples
 X : the no of 1's in the sample of n individuals
 Problem: μ_x, σ_x

$X = Y + Y + Y \dots n \text{ times}$
 $= \mu_y + \mu_y + \mu_y \dots n \text{ times} = n\mu_y$

$\text{Var}(X) = n \text{Var}(Y) = n p(1-p)$
 $\text{SD} = \sqrt{n p(1-p)}$

The measured proportion

$\hat{p} = \frac{X}{n}$
 $\mu_{\hat{p}} = \frac{\mu_x}{n} = \frac{n p}{n} = p$
 $\sigma_{\hat{p}} = \frac{\sigma_x}{n} = \frac{\sqrt{n p(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$

NPTEL

We have come to a situation where the problem is that, from the population we draw a certain number of samples, say n samples, and a part of that we will be found to have the value of 1, a part of that we will have the value of 0. We are trying to find out what proportion has the value of 1. We have drawn n samples. Suppose the number of 1s in that n sample is called X . So, X the number of 1's in the sample of n individuals.

X is a fraction of n and we will now face the problem of finding what is the mean of X , the standard deviation of X , and so on and so forth. If I draw that n number of samples again and again from that population, I will not get the same value of X every time. I will get different values, and those values of X will also have a distribution. And we have just concluded that, if the number of samples becomes large, that distribution will become more or less a normal distribution.

Now we are facing the problem of finding what is the mean of X , and what is the standard deviation of X . So, our problem is the mean of X and the standard deviation of X . Now, X is the number of 1's in the sample of n individuals. Therefore, in finding out

what X is, what I am doing is that, I am picking up each individual from the sample and we are asking whether it is 1 or 0? Again I am picking the second individual and I am asking: 1 or 0? This is the same as what we are doing earlier. That means, we are actually making a measurement of the Y 's. Each measurement is Y , which can have two possible values 1 or 0. We have seen that earlier.

So, the measurement of X is actually n measurements of Y , which is nothing but Y plus Y plus Y that goes on n times. If that be so, we have already found out what is the mean of Y . Then the mean of X is nothing but the mean of Y , n times.

So, then we can argue that, the mean of X can be simply expressed as the mean of Y plus the mean of Y plus the mean of Y ... n times, and we know what the mean of Y is p . Therefore, n times of p is np . So, that should be the mean of X .

Now, what will be its variance? I would argue that, we would carry out the variance calculation exactly in the same way, using the variance of Y . Since the measurement of X is nothing but n measurements of Y , and we know the variance in the measurement of Y , therefore, this is nothing but n times the variance of Y . And we have already learned that: n times the variance of Y is p into $1 - p$. So, that should be the variance of X .

So, we have a distribution whose the variance is known. Therefore, the standard deviation would be square root of $np(1 - p)$. Now, we are actually trying to find out the measured proportion. The measured proportion is the number that we have found: X is the number that we have found, divided by the total number of samples. That is the measured proportion.

Let us call it measured (it is not the proportion out there), let us call it with a hat. So, we are talking about the measured proportion: the 1's that we have measured using n number of samples is \hat{p} and that will be the measured proportion.

And if we do a large number of trials of taking n samples each time, we will get a distribution of the measured proportions. That means, we have taken n samples and we have got a measured proportion. If we do that again, taking another n samples, we will get another value of this. A third time you will get another value of this. Ultimately, the \hat{p} will have a distribution.

What will be the mean of the distribution of the p hats? Let us call it mean of p hat. That will be the mean of X divided by n. Since the p is X by n, therefore, it will be this, and the mean of X is something we know: n times p.

So, n times p by n. Therefore, it will be p. So, if I repeatedly make measurements, each time taking n samples, I will get a distribution and that distribution will have a mean which is the proportion 'out there' in the population. So, that is a good thing. We can obtain the actual proportion by making repeated measurements, every time taking n samples.

And what will be the standard deviation? The standard deviation, sigma of p hat, that will be the sigma of X divided by, there are n samples, so n. Now, we know that sigma of X was square root of np 1 minus p, and here is n. Therefore, this comes to be square root of p 1 minus p by n.

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

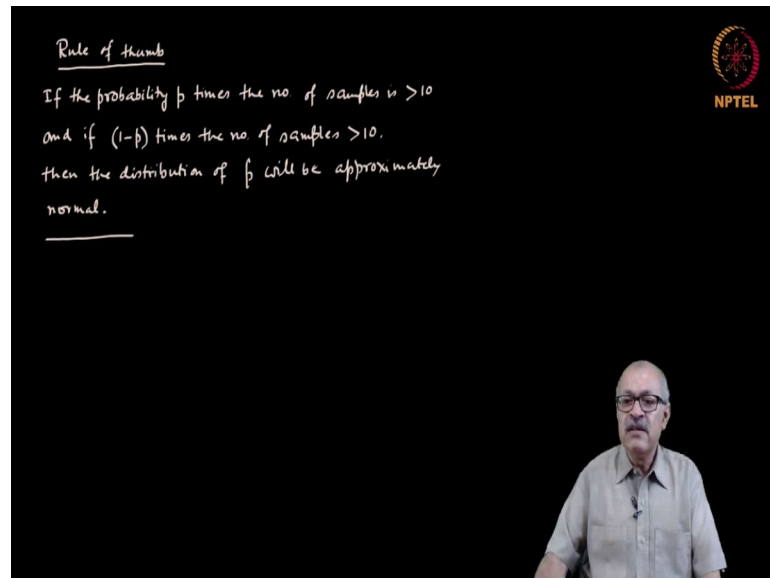
So, there is a proportion in the population out there and I am trying to find it using a sampling process. I am drawing samples, n samples each time. If I repeat the process again and again, then every time I will measure a different value of the measured proportion p hat, and that will have a distribution. That distribution will have a mean at the population mean p, and that p hat will have a distribution whose standard deviation will be this, a well defined number.

You might ask: what is the guarantee that this distribution will be a normal distribution? Earlier our argument for rooting on a normal distribution was the central limit theorem. But here we cannot apply the central limit theorem, because it is a different problem we are dealing with. So, how do we ensure that it will be a normal distribution?

The question is legitimate. It has been found that, it goes by a rule of thumb. It is not difficult to see that, if the number of 1's in the population is too small, it is like a dwindling population, something close to extinction, if that is so, then the distribution will not be a normal distribution.

So, there are situations where it will not be a normal distribution. Therefore, we need to talk about under what condition we can expect more or less a normal distribution. There is a rule of thumb for that.

(Refer Slide Time: 11:16)



The image shows a video frame with a black background. In the top right corner, there is a circular logo with a globe and the text 'NPTEL' below it. The main content is handwritten text in white: 'Rule of thumb' is underlined. Below it, the text reads: 'If the probability p times the no. of samples $n > 10$ and if $(1-p)$ times the no. of samples > 10 , then the distribution of \hat{p} will be approximately normal.' The text is underlined at the end. In the bottom right corner, there is a small inset video of a man with glasses and a mustache, wearing a light-colored shirt, looking towards the camera.


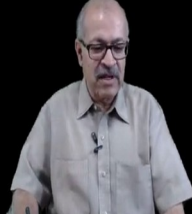
This is rule of thumb. It is not a theorem. Just people have tried out and found that it works. That is, if the probability p of getting a 1 times the number of samples is greater than 10, and if the probability of the other one; that means, $1 - p$ times the number of samples, that is also greater than 10, then the distribution of the measured \hat{p} will be approximately normal.

If it is normal, then we can apply all that we have learnt about the normal distribution. That is a major advantage. But the caveat is that, that is not applicable to all cases. For example, in the example that we have taken, p was 0.6, $1 - p$ was 0.4, and suppose we take 20 samples. n is equal to 20. Then it will become p times 20, p into n will be 12 and $1 - p$ into n will be 8.

(Refer Slide Time: 13:20)

and if $(1-p)$ times the no. of samples > 10 .
then the distribution of \hat{p} will be approximately normal.

$p = 0.6, 1-p = 0.4, n = 20$
 $p \times n = 12, (1-p)n = 8$

So, obviously, this is not bigger than ten and therefore, in this collection the normal distribution will not apply. So, if you keep on drawing samples of 20s and expect \hat{p} to be distributed in normal distribution, that will not work. You have to take more than 20 samples; then only the normal distribution will work.

(Refer Slide Time: 14:35)



Measurement of a proportion

Ex Tall plants, short plants
 $n = 50, 33$ are tall.
 $\hat{p} = \frac{33}{50} = 0.66. \quad p \in [0.64, 0.68]?$

The measured proportion

$$\hat{p} = \frac{x}{n}$$

$$\mu_{\hat{p}} = \frac{\mu_x}{n} = \frac{np}{n} = p$$

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



Before we end, let us quickly do an example. Consider the Mendel-type experiment in which there are tall plants and short plants, and you have collected a sample of 50. You have counted the number and you have found that 33 are tall. I will write.

So, your measured proportion is 33 by 50 is equal to 0.66. The question is, can you state that the actual proportion out there lies in the range 0.64 to 0.68? Can you say that? Does actual p lie in this range? This is the problem.

(Refer Slide Time: 16:22)

$n = 50$, 33 are tall.
 $\hat{p} = \frac{33}{50} = 0.66$. $p \in [0.64, 0.68]$?
 $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.66(1-0.66)}{50}} = 0.067$
 $p \times n = 33$, $(1-p)n = 17$, normal.
 $P(\mu_p \text{ is within } 0.02 \text{ of } \hat{p})$
 $= P(\hat{p} \text{ is within } 0.02 \text{ of } \mu_p)$
 $z = \frac{0.02}{0.067} = 0.3$
 The area = 0.6179
 $(0.6179 - 0.5) \times 2 = 0.2358$

So, how do we attack this problem? We know, we have calculated the mean, sample mean, to be this, and we can calculate the sample standard deviation: sigma \hat{p} to be square root of $p(1-p)$ by n .

Now, p we do not know. Therefore, we substitute by whatever we have measured, which is this. So, it comes to be 0.66 times 1 minus 0.66 divided by the number of samples. We have taken 50 and that turns out to be 0.067.

$$\sigma_{\hat{p}} = \sqrt{\frac{0.66(1-0.66)}{50}} \approx 0.067$$

Now, will this distribution be a normal distribution? Let's quickly check: p is 0.66. So, p into n is 0.66 times 50 is 33, and 1 minus p times n is equal to 17. So, both are bigger than 10. So, the distribution is normal. That is good.

So, if the distribution is normal, then we can picture the distribution something like this. We know the mean will be at p , and the standard deviation we have just calculated.

The question was the probability of this. So, probability P of the mean of p is within 0.66 and 0.64, i.e., within 0.2 of the mean: this is what we are trying to calculate. So, p hat 0.66 minus 0.2 and plus 0.2. So, our statement is: what is the probability that this is true?

$$\begin{aligned} &P(\mu_p \text{ is within } 0.02 \text{ of } \hat{p}) \\ &= P(\hat{p} \text{ is within } 0.02 \text{ of } \mu_p) \\ &= P(\hat{p} \text{ is within } 0.02/0.067 = 0.30 \text{ standard deviations of } \mu_p) \end{aligned}$$

I will write this as capital P because it is a probability. Again, like the earlier problems, we will reverse the argument. We will say p hat is within 0.2 of the mean of p, the distance between this and that is the same as distance between this and this.

So, we have the same argument. And then we have to express that as a multiplier of the standard deviation. Then we get the value of z. So, the z value is the multiplier 0.02 divided by this is 0.067, and that turns out to be 0.3. z turns out to be 0.3.

So, we are essentially saying that p hat is within 0.3 standard divisions of the mean. You now consult the z table and find out what is the area for z=0.3. Suppose 0.3 is somewhere here, may be. If you find this area, then you will find that the area is 0.6179.

So, we will subtract 0.5 from it. So, 0.6179 minus 0.5, thereby we get only this area. That times two. That turns out to be 0.2358. So, that is the probability of having the actual proportion of the two populations out there to be in this range. That is only this much.

Which means that, I can state with only 23 percent confidence that the proportion will be in this range. It is a very low confidence. That immediately tells us that we need to change our strategy. We need to do something else in order to increase the level of confidence for proportion measurements. I will stop here and continue with that in the next class.