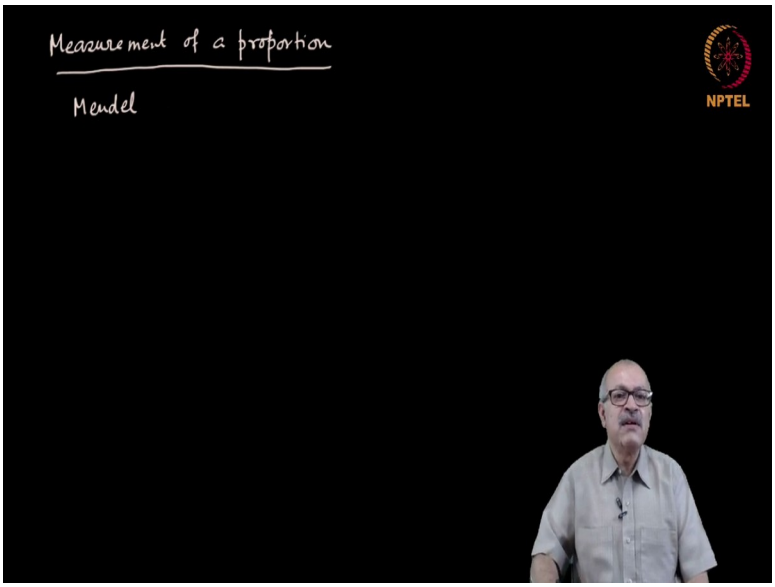
Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 35
Measurement of a Proportion, Part 01

We have learnt how to obtain measurements of a value, a value which can take any number that comes from some kind of a measurement, some kind of a reading of an instrument. We have learnt how to obtain the reliable mean value of it, how to state it to an external audience in the form of an error bar. But there is another kind of measurement that a scientist often has to do. It is the measurement of a proportion.

(Refer Slide Time: 00:55)



So, the measurement of a proportion. Just consider the experiment that Mendel did. Gregor Mendel, he crossed two varieties of pea plants: one tall, another short. It was found that there is nothing in between, tall or short. He crossed these two varieties and in the next generation he found that all the offspring are tall. Then he crossed within that generation and found that a proportion of that is tall and another proportion of that is short. So, his task then was to find out what proportion is tall.

Such kind of measurements are the measurements of a proportion. There are various situations where one can have this kind of proportion measurement problem. In any evolutionary process one encounters this kind of situation. For example, a very well

documented process of evolution that happened before human eyes was for a species of moth in England. The moth was originally whitish in color and they sat on the bark of the trees which were also whitish in color. So, they could properly camouflage, the birds could not find the moth and eat them. For millennia that was the stable situation.

But with the advent of modern civilization, with the advent of atmospheric pollution, the bark started turning more and more darker. As a result, the moths became visible and they could be eaten by birds and it was found that over a very short period of time, the whole species changed into a brown darker species, which could again be adapted to that particular colour of the bark.

Now, essentially what happened was, there was a mutation within the species which produced a coloured offspring and that coloured variety was chosen by nature. It was natural selection, the brown variety was selected by nature. So in successive generations, the coloured ones, the brown variety of the moth, their proportion increased. Finally, after some time there was no white ones left.

So, that was a speciation event, happening before our eyes. But if a scientist is studying that speciation event, then he will have to count or find out the proportion of the brown moths in each generation and that will tell you how is speciation actually happens. So, it is a problem of measurement of a proportion. Similarly in every field there are similar types of proportion measurement problem.

If a beaker has two different types of microorganisms and you are trying to find out the proportion. Suppose one is a phyto plankton, the other is a zoo plankton. One is in a vegetative kind, plant kind, the other is a animal kind that eats on the plants. As time progresses, the proportion of the two will change and that leads to the dynamics and one has to study that by studying the proportions.

So, there are various situations where one has to measure proportions. Let us attack that problem because this problem is slightly different from the measurement problem that we have encountered so far.

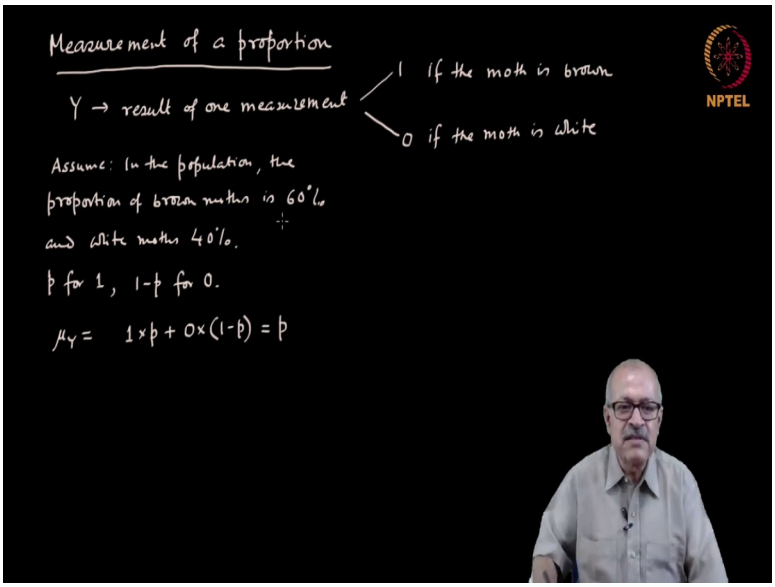
We assume that there is some entity whose property we are measuring. But that can take only two values: either a tall pea plant or a short pea plant, nothing between tall and

short. If we are trying to find out the proportion of the talls in the whole population, then if we find a tall one we will say 1 and if we find a short plant we will say 0.

So, there are two possibilities 1 or 0. Similarly, in the moth population, if there are two possible varieties: the white ones and the brown ones, then may be the brown ones will be called 1 and the white ones as 0. The point that I am making is that, in a proportion finding problem one has the entity given as either 1 or 0, that kind of a situation. Let us consider the situation of the moths, because it is better to work with an example to make things clearer.

So, we are counting moths and finally, trying to find out what is the proportion of the brown moths in the whole population. There are brown moths as well as white moths. What is the proportion of the brown moths? We catch one and we see that it is brown or white, and we assign a number 1 or 0 to it.

(Refer Slide Time: 07:01)



Let us call that measurement to be Y. Y is the result of one measurement. Now, that can be either 1 or 0. When would it be 1? If the moth is brown. And 0 if the moth is white. There are two possibilities 1 or 0. So, the Y can get two possible values, 1 or 0.

Suppose we assume, in the population the proportion of brown moths is, say, 60 percent, and white moths 40 percent. Suppose this, as we are working with an example in mind.

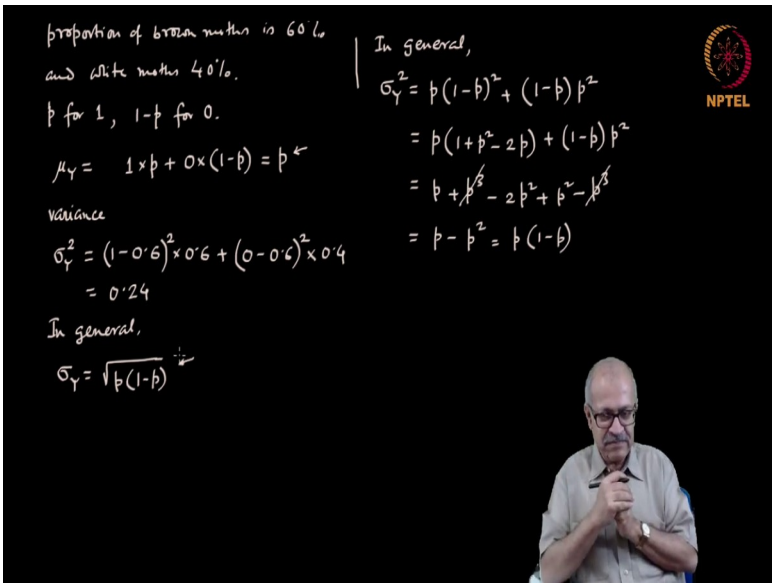
So let us assume these values. In that case, in general there will be a probability of finding 1 and another probability of finding 0.

So, p for 1 and 1 minus p for 0. In that situation, what will be the mean value of Y? I have collected many and then I have to divide by the total number of samples and thereby we obtain the mean value, the mean value will be mu: the mean value of Y.

$$\mu_Y = p \times 1 + (1 - p) \times 0 = p$$

We will do it by weighing the various possibilities. One possible value is 1, weighted by its probability, plus another value is 0 weighted by its probability. That yields p. Therefore the mean of Y should be p. In this case it would be 0.6. You would notice that 0.6 is not a value that Y can take. It can either take 1 or 0. But the mean value can be a fractional number. Now, what will be the variance of this? Let us try to work out the variance.

(Refer Slide Time: 10:37)



The variance can be obtained as a weighted sum. Well, what is variance? The difference between the value that I get and the mean, squared, and its average. There are two possible values. So we will simply do a weighted sum of the squares of the distance from the mean.

Now, there are two possible values 1 or 0. Let me write sigma Y, its variance. One possible value is value is 1 minus the mean value, this is the difference from the mean, squared, and I have to weigh it by its probability, 0.6. The other possible value is 0. So, 0 minus 0.6, square, into its probability 0.4. You see, the total variance turns out to be 0.24.

So, notice the logic: there are only two possible values 1 and 0. If it is 1 then the distance from the mean is this much, its square, that has to be weighted by the probability of getting 1. Similarly another possible value is 0, distance from it is this one, its square and you have to weight it by the probability of getting 0. Thereby you get the variance. So, let us do it in general, in terms of p.

$$\begin{aligned} \sigma_Y^2 &= p(1-p)^2 + (1-p)p^2 \\ &= p(1-2p+p^2) + (1-p)p^2 \\ &= p - 2p^2 + p^3 + p^2 - p^3 \\ &= p(1-p) \end{aligned}$$

In general, our sigma Y square will be this: p times 1 minus p square plus this was 1 minus p 1 minus p times p square. Let us just work it out: p 1 plus p square minus twice p plus 1 minus p, p square is equal to p plus p cube minus twice p square plus p square minus p cube and p cube cancels off and we are left with p minus p square, or p into 1 minus p.

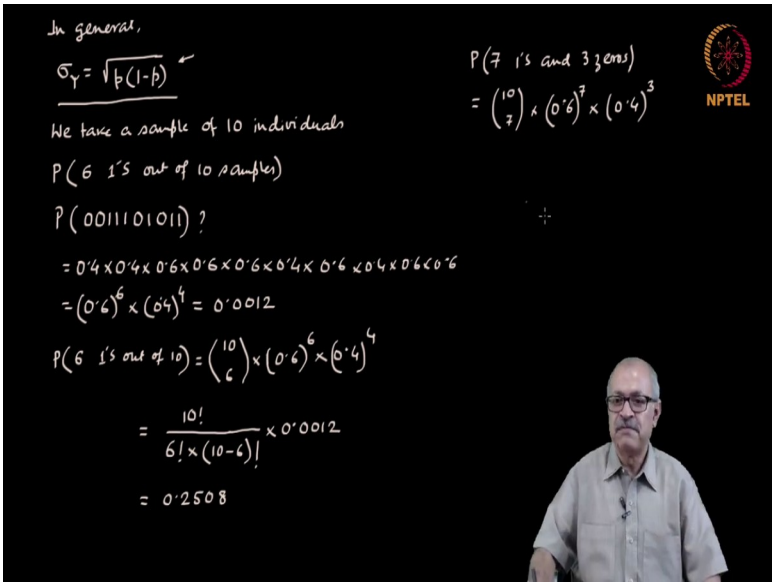
So, that then is the variance of the individual observation. The variance is this and we get the standard deviation sigma Y as the square root of p into 1 minus p.

So, this is the result we get for a single observation. That means, I make an observation, I get a value. I make another observation, I get another value. If we go on doing that for a long time, then it will get a distribution and that distribution we will have a mean p and the standard deviation this, and this distribution is called the Bernoulli distribution.

But you would notice that we actually do not make the measurement that way. How do we make the measurement? The way we actually make the measurement is that, from the population we collect a number of moths and then from that number of moths, say n, we

find out how many are brown and how many are white. So, in general we do it in a slightly different way.

(Refer Slide Time: 15:49)



We get a sample. We take a sample, say, of 10 individuals. Now, from that we will find out how many are the tall ones how many are the brown ones. How many are 1s and how manys are 0s. And, for each number, there will be a probability. Then can we find that probability. Let us try to work it out.

The point is that, we have assumed that p, the probability of finding 1 in the population, is 0.6, and probability of finding 0 in the population is 0.4. On that basis we are proceeding. We are now taking 10 samples. On an average 60 percent will be brown, on an average 40 percent will be white. And we have drawn 10 samples. Therefore, we expect that, 6 would be brown and 4 will be white. Will that really happen? No, not really, because we are ultimately doing a sampling out of a completely randomly mixed population. Therefore, there is no reason to expect that for every sample that we draw, it will reflect the population mean. That will not happen. But still we can calculate, what is the probability of 6 1's out of 10 samples? How we can do that?

Suppose we take a particular sequence. Suppose we have we collected 10 and when we find out the first one then it comes to be white so 0, second one is also white, third one is brown, 1, and fourth one is brown again, white, brown. How many 4s? Suppose we have got it in this sequence, what is the probability of this sequence? That is rather easy to

calculate, because what is the probability of getting 0? 0.4. What is the probability of getting 1? 0.6. Therefore when I have already got the first one, it had a probability of 0.4. When I got the second one what is a probability? The second one will also be 0, that means, white, it is 0.4. So, times 0.4. Effectively by the law of multiplication of probabilities the probability can be easily calculated as 0.4 for the first one times 0.4 for the second one because it is 0. Times the next three are ones, 0.6 times 0.6 times 0.6 times this one is 0 so, times 0.4 times this was 0.6. So, 0.4 times 0.6 times 0.6 and this is it is easily seen that it does not really matter in which order the 0s and 1s appeared. What really matters is how many times they appeared and so we can write this as 0.6 appeared 6 times and 0.4 appeared 4 times. So, it is 0.0012.

$$
\begin{aligned}
&P(\text{appearance of the sequence } 0011101011) \\
&= 0.4 \times 0.4 \times 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.6 \times 0.4 \times 0.6 \times 0.6 \\
&= 0.6^6 \times 0.4^4 \approx 0.0012
\end{aligned}
$$

So, this is how these are calculated. But you will notice that I had initially set out the problem as 6 1s out of 10 samples. This is not the only way you can have 6 1s out of 10 samples. It is possible that the first 6 are 1s and then the last 4 are 0s. That is also another possibility. So, there are various possibilities. But notice that all these possibilities will have the same probability. So, all these different possible ways of getting 6 1s and out of 10 samples, all of them will have the same probability.

So, all we need to do is to find out how many ways we can have that 6 1s and multiply that by this number. We know that the number of ways by which you can get 6 1s out of 10 samples is simply 10 choose 6.

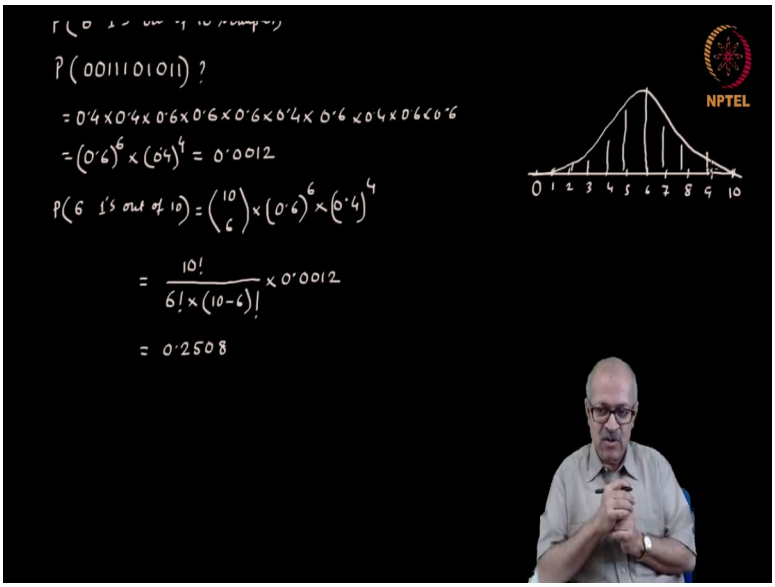
$$
\begin{aligned}
P(6\ 1\text{'s in } 10 \text{ samples}) &= \binom{10}{6} 0.6^6 \times 0.4^4 \\
&= \frac{10!}{6! \times (10-6)!} 0.6^6 \times 0.4^4 \approx 0.2508
\end{aligned}
$$

So, the probability of 6 1s out of 10 will then be 10 choose 6, this is how it is written 10 6 and that has to be multiplied by this, into 0.6 to the power 6 times 0.4 to the power 4. And we know how to write that: it is 10 factorial divided by 6 factorial times 10 minus 6 factorial times this number, which you have calculated as 0.0012, and this comes out to be 0.2508.

So, the lesson is that, even though in the population the number of the proportion of the brown moths is 0.6, if I draw a sample of 10, the probability that I will find 6 of them to be brown is only 25 percent. It is rather small number.

We can similarly find out what is the probability of having say 7 1s and 3 0s. It is not difficult to calculate that. It will be 10 choose 7, 7 1s into 0.6 to the power 7 into 0.43. That way we can calculate what is the probability of getting only 1 out of 10 samples, only 2 out of 10 samples, and so on so forth, and we can plot a graph.

(Refer Slide Time: 24:09)



The graph here: it is possible to get 0 1 2 3 4 5 6 7 8 9 10. 10 samples we have drawn and it is possible to get all of them white, it is possible to get all of them brown, it is possible to have all the intermediate ones. You will find that the largest probability will have happened at 6, because the proportion of the brown moths in the population is around 60 percent. So, this will be smaller, this will be smaller and this will be again smaller.

So, you will get a distribution. And if you now increase the number of samples, that n, if you now increase it, you will get more possibilities and therefore, each possibility will have a probability assigned to it. As you increase and increase the number of samples, it will slowly become a smooth distribution. It will tend to a normal distribution. The moment we claim that it will tend to a normal distribution, we immediately face the question, what will be the mean and what will be the standard deviation? We immediately face that question. We will take care of that.

So, what I have driven at is that, when we take samples from the population and find out the proportion within that sample, then the various possibilities will have a distribution like this, and as you increase the number of samples that distribution will tend to become a normal distribution.

The distribution of only the discrete ones is a binomial distribution, and as the number increases it tends to the normal distribution. Now, we face the problem of finding out the mean and the standard deviation of that normal distribution.