Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Sampling Theory with R Software
Lecture - 26
Sampling for Proportions and Percentages
Mean and Variance of Sample Proportion

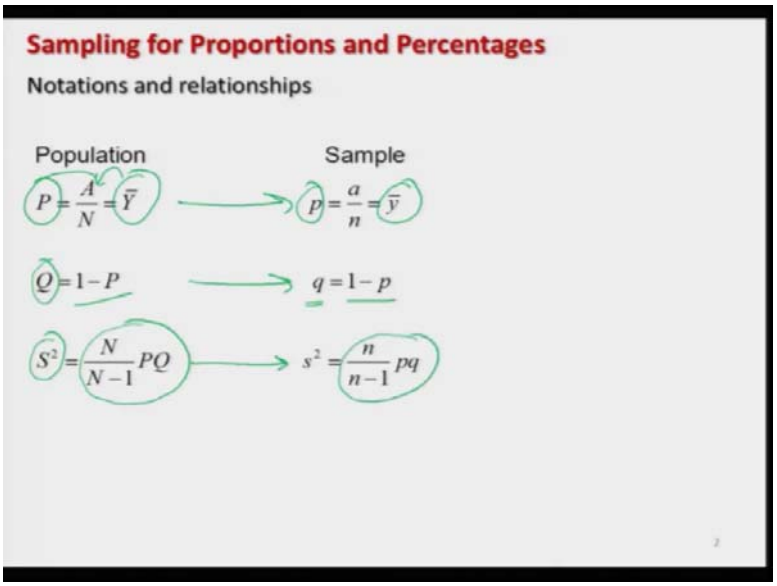
Hello friends, welcome to the course Essentials of Data Science with R Software-2 where we are trying to understand the topics of Sampling Theory and Linear Regression Analysis. In this module, we are going to continue with the Sampling Theory with R Software and we are going to discuss the topic of Sampling theory for Proportions and Percentages.

So, you may recall that in the earlier lecture, I started a discussion on the sampling for proportion and percentages and I proposed that we use the simple random sampling. And what I did? I just made a one to one correspondence between the tools in the setup of proportion with the tools in the setup of a quantitative variable.

Now in this lecture I will try to continue my discussion and I will be using those results in establishing several things. First question how to estimate the population proportion, what are the properties of the estimator of population proportion like as their bias, variance, standard error and confidence interval.

So, you will see that although we took a long time in simple random sampling in understanding all these concepts, but in this case that will be quick and that may not take long time. Why? Because all the basic fundamental and concept in that chapter. Now, I am going to use it. So, once again I assume before the start of the lecture that you have taken a good revision of the earlier concept and they are in your mind while I am doing this lecture ok. So, let us start, let us begin our lecture with the quick revision.

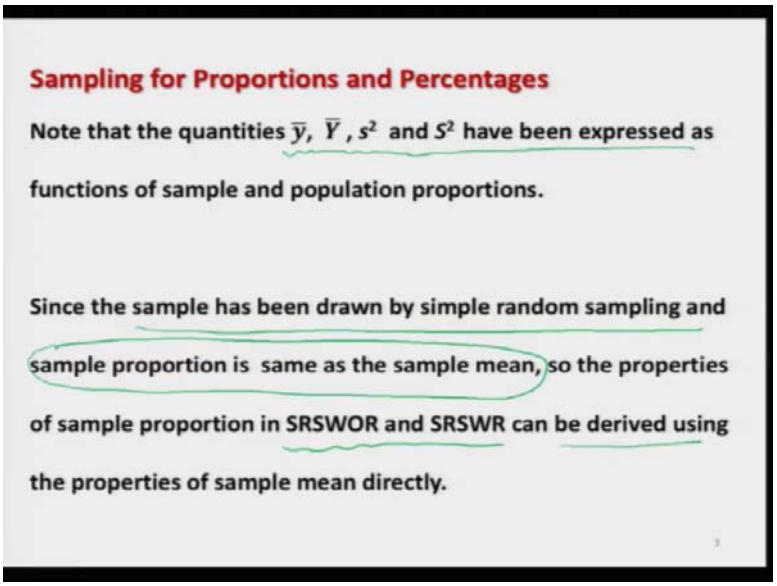
(Refer Slide Time: 02:22)



So, you may recall that in the earlier lecture, we concluded it with this slide where we had denoted the population mean $\bar{Y}$ which translated to A/N and that was indicated by population proportion P and the corresponding value was the sample proportion that is miserable with sample mean $\bar{y}$.

And similarly Q was the proportion of the units in the complementary class which is 1 - P and the sample version of Q is a q which is 1 - p which that is 1 - sample proportion in the class c. And similarly we had a converted $S^2 = \frac{N}{N-1} PQ$ and this thing is translated to a s^2 in the setup of a sample and $s^2 = \frac{n}{n-1} pq$, right.

(Refer Slide Time: 03:37)



So, with this notation what I have done? That I have converted these quantities $\bar{y}, \bar{Y}$, s^2, S^2 in terms of population proportion which is the setup of simple random sampling when the variable under study is qualitative right. Now you have to think on one lines on these lines and you have to concentrate what I am trying to say. This is going to create the basic fundamental for the development of all the results in this situation.

Now, you may recall that in the case of simple random sampling, we had used the sample mean as an estimator of population mean and I had given you the justification that why do we prefer to use sample mean rather than any other estimator. And then I established that sample mean as an unbiased estimator in case of SRSWR as well as WOR and after that I had found the variance of sample mean under SRSWOR and SRSWR.

Now, once I have made a one to one relationship between the quantities like $\bar{y}$, $\bar{Y}$ s^2 and S^2 and when I am going to follow the same sampling procedure what I had followed earlier; then all the properties whatever we have found in the case of simple random sampling for a quantitative variable that will simply be translated to simple random sampling for a qualitative variable.

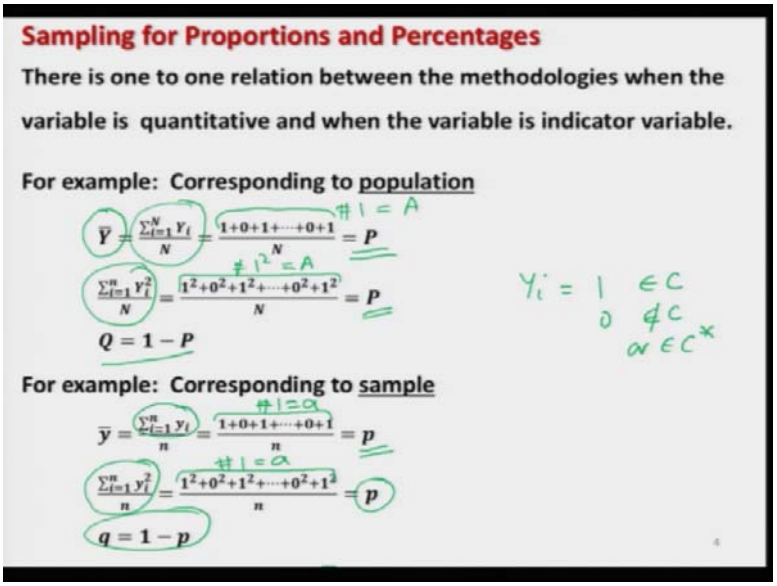
So, now you think take a pause, try to think about it what I am trying to say. In case if use this fundamental or if you have understood this fundamental, whatever are the

properties of sample proportion they can be obtained directly without doing any algebra without doing any mathematical calculations. All those calculation, they will be followed from the past from the earlier lectures on simple random sampling for quantitative variable, ok.

So, let us try to begin this discussion here. So, since the sample has been drawn by simple random sampling and we have noticed that sample proportion is the same as sample mean. So, whatever are the properties of sample mean they will continue to hold for the sample proportion also in the case of simple random sampling with and without replacement.

And if you want to find them mathematically or statistically you can follow the same steps and you can find them. So, whatever are the properties of sample mean, they will be translated to sample proportion.

(Refer Slide Time: 07:02)



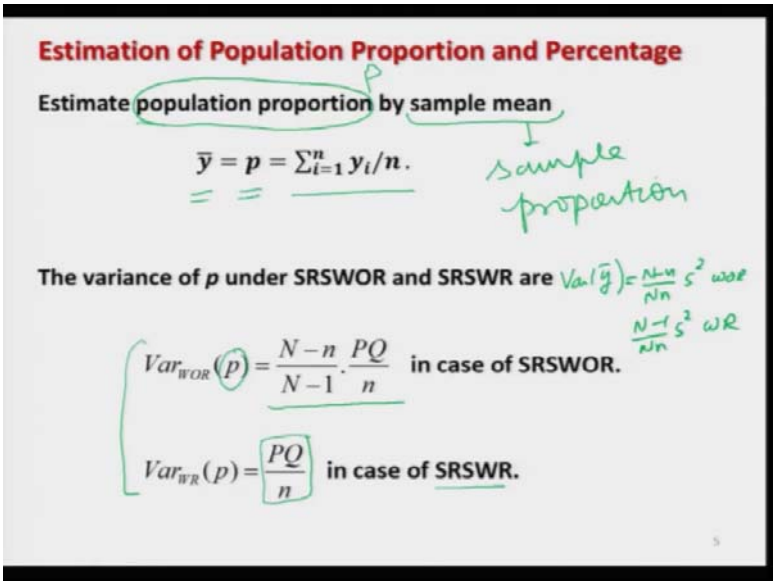
Now, but before that let me try to explain you what do I mean. For example, if you try to consider the quantity $\bar{Y}$ which was earlier defined for the quantitative variable then you have means of; that means, I already had shown you in the earlier lecture this is going to be $\frac{1}{N}\sum_{i=1}^{N} Y_i$. So, this is going to be the sum of 1 0 1 0 and so on divided by N.

So, this the number of 1's in this is A. So, this will become $\bar{Y}$ is same as P and $\frac{\sum_{i=1}^{N} Y_i^2}{N}$, this is same as $1^2 + 0^2 + 1^2$ and the number of here 1^2s are here once again A. So, this is here. Once again P right and Q is here 1 - P. And similarly for the sample also sample mean is defined as $\frac{1}{n}\sum_{i=1}^{n} y_i$.

So, y_i takes value here 0 or 1 because you remember that we had defined Y_i equal to 1 if the unit belongs to the class C or if the unit does not belong to C or it belongs to class C star right. So, this summation y_i will be will be something like 1 +0 +1 and so on and the total number of 1's in this sum will be a.

So, this a/n will be same as p and similarly $\frac{\sum_{i=1}^{n} y_i^2}{n}$ that will become, once again here sum of 1^2 and 0^2 and the number of here 1's are a. So, that will once again become a/n which is here p and q is 1 - p. So, now with this reasoning with this explanation, now we are we have completed our basic fundamental to understand the theory and fundamental for the case of qualitative variable.

(Refer Slide Time: 08:59)



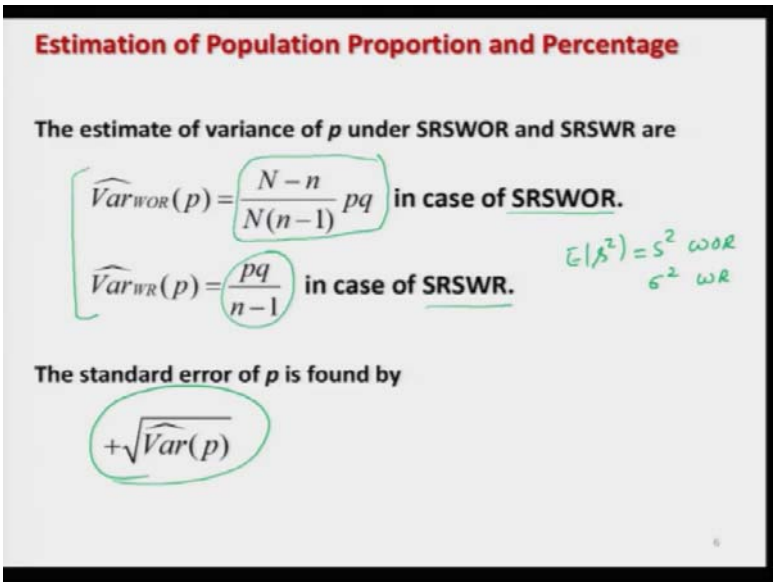
So, now I can propose without going into the detail without giving you a long lecture that let me estimate the population proportion p by the sample mean which will translate to the sample proportion. Because if I can estimate the sample if I can use the sample mean to estimate the population means so, simply so, on the same lines I can extend and can say that sample proportion is being used to estimate the population proportion.

So, now, this $\bar{y}$ which is here the same as p that will be given by $\frac{1}{n}\sum_{i=1}^{n} y_i$ . . Now, we already had found the variance of $\bar{y}$ as $\frac{N-n}{Nn}S^2$ in case of WOR and $\frac{N-1}{Nn}S^2$ in case of WR, right.

So, this relationship can be used here correctly and just by using the value of S^2 , S^2 which we had defined you see here in this slide as (N/(N – 1)) PQ, I can substitute these values and I obtain the variance of sample proportion p under WOR that is SRSWOR is used as $\frac{N-n}{N-1}.\frac{PQ}{n}$.

And in case of SRSWR, this quantity will become here PQ/n. So, these two expression will give us the variance of sample proportion just like we had obtained the variance of sample mean in the case of quantitative variable, right.

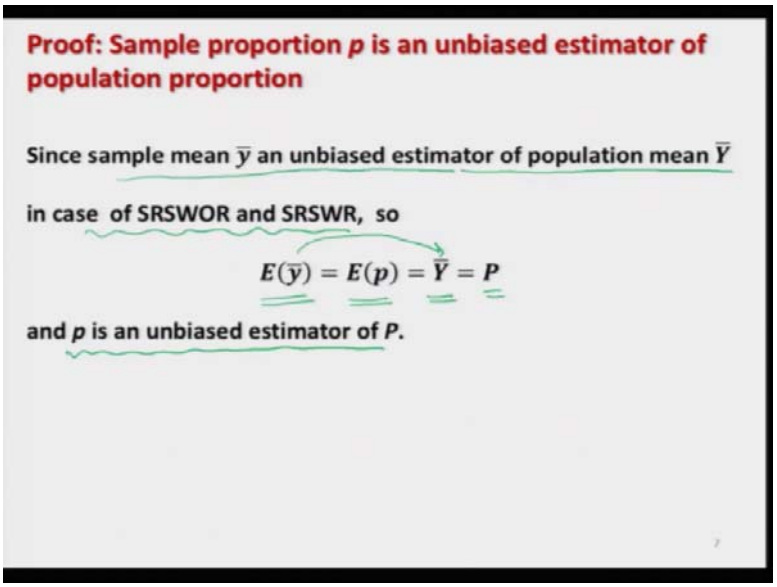
(Refer Slide Time: 11:12)



And after that if you remember, we had obtained the expected value of s^2 which was coming out to be S^2 and σ^2 in case of without replacement and with replacement. And we had substituted these values in the variance of $\bar{y}$ to obtain the unbiased estimators of variance of $\bar{y}$ under SRSWOR and SRSWR in case of simple random sampling for quantitative variable.

Similarly, I can replace these quantities over here in the estimate and I can obtain the estimate of sample proportion p under SRSWOR. Here I say $\frac{N-n}{N(n-1)}pq$ and under the case of SRSWR that will be a $\frac{pq}{n-1}$. And in case if you want to find out the standard error of this sample proportion that will be positive square root of the variance the corresponding variance under SRSWR or WOR, right.

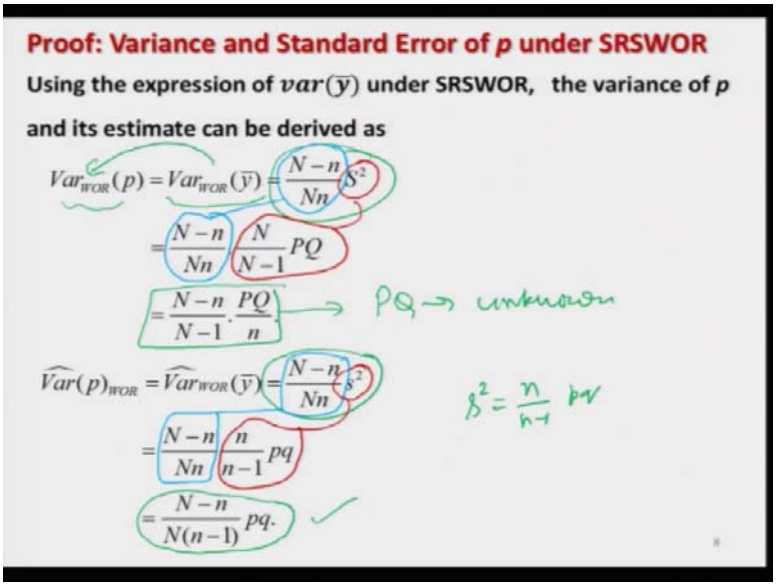
(Refer Slide Time: 12:32)



So, this is how I can do everything in case of when the characteristic understudy is qualitative. Now let me try to give you here the simple proof of the properties of sample proportion. Since sample mean and sample proportion they are interconnected. So, since we already have proved that sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$ so, this result can be extended for the sample proportion also.

So, in both the cases when the sample have been drawn by without replacement or with replacement; in both the cases sample mean is an unbiased estimator of the population mean, hence I can write that expected value of $\bar{y}$ will be same as expected value of p which is equal to here $\bar{Y}$. and that is the same as P in our situation. And hence I can say that sample proportion is p is an unbiased estimator of population proportion in case of SRSWOR as well as SRSWR, right.

(Refer Slide Time: 13:41)



Similarly, if I want to give you a quick proof of the variance of sample proportion so, you remember that variance of sample mean in the case of without replacement, we had found to be like this $\frac{N-n}{Nn} S^2$. So, this variance of $\bar{y}$ will be same as variance of sample proportion p under without replacement.

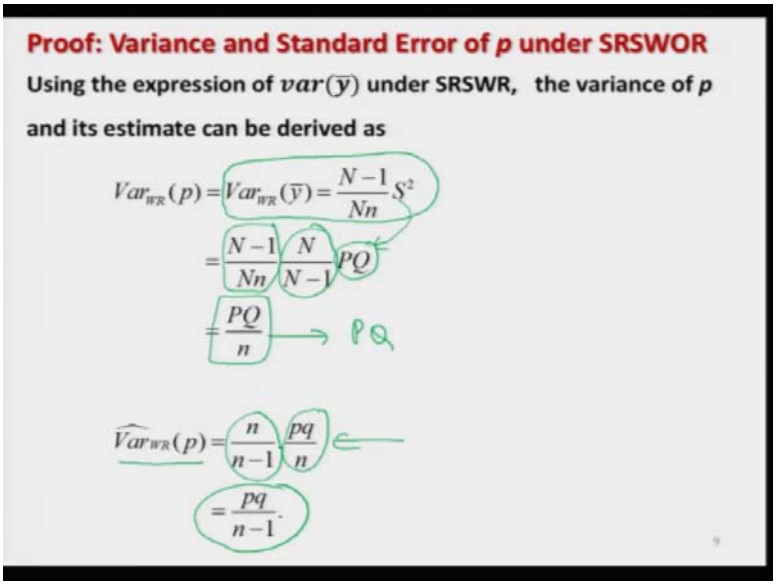
So, you can see here that in this quantity this and - N/n N n, this continues here and this S^2 is replaced here by N (n – 1) PQ and if you try to simplify it, you get here the result which I stated in the earlier slide like here this one. Similarly if you try to find out the estimate of this variance because you can see here that the variance of p involves here the quantities PQ which are unknown to us.

So, we would like to estimate this variance on the basis of given sample of data. So, we had earlier found in case of SRSWOR that the variant of WOR, the estimate of the

variance of $\bar{y}$ under WOR was obtained here like this; $\frac{N-n}{Nn}S^2$. And you already have established earlier that a s^2 is the same here as a $s^2 = \frac{n}{n-1}pq$ right.

So, what I do here that this quantity N - n/Nn, this continues here and for a s^2, I try to replace here this quantity. And if you try to just simplify it here, you get here an estimate of the sample proportion when the sample have been drawn by SRSWOR like this $\frac{N-n}{N(n-1)}pq$.

(Refer Slide Time: 16:06)



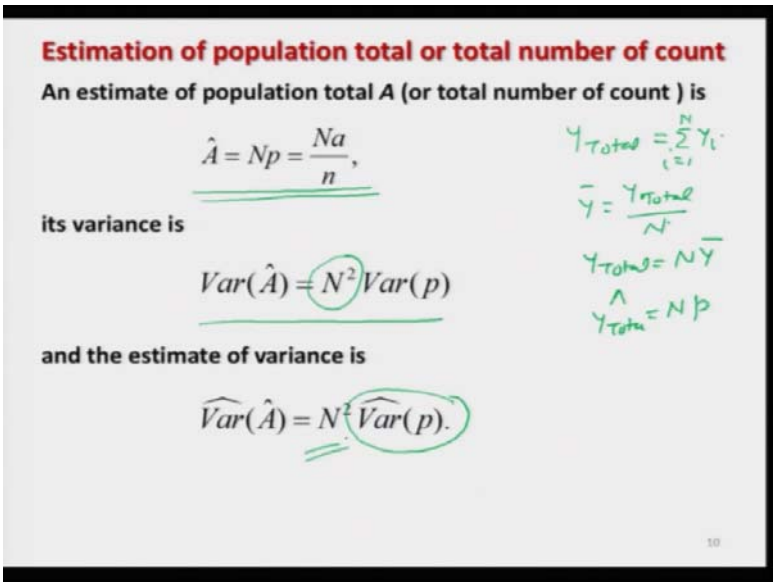
So, this is how you can estimate the variance on the basis of given sample of data and the same stories same story continues for the simple random sampling with replacement also. You can see here that we had obtained the variance of sample mean under without with replacement as $\frac{N-1}{Nn}S^2$.

And if you simply try to replace here the S^2 in terms of PQ, you get here $\frac{N}{N-1}PQ$ and multiplied by (N – 1)/Nn. And if you try to simplify it, you get the variance of sample proportion under SRSWRS PQ/n. And since here this quantity depends upon PQ which are the population values which are unknown to us.

So, we cannot find the variance of sample mean from the sample using this expression. So, we try to find out the estimator of variance and the variance estimator of sample proportion when the samples have been drawn by with replacement as obtained by $\frac{n}{n-1}.\frac{pq}{n}$, right.

Why? Because if you try to see you had earlier found the variance as here PQ/n, you can see here right. So, if you simply try to replace the quantities σ^2, you can obtain here this expression and from there I can say that the variance of sample proportion under WR can be estimated by this quantity $\frac{pq}{n-1}$, right.

(Refer Slide Time: 17:48)



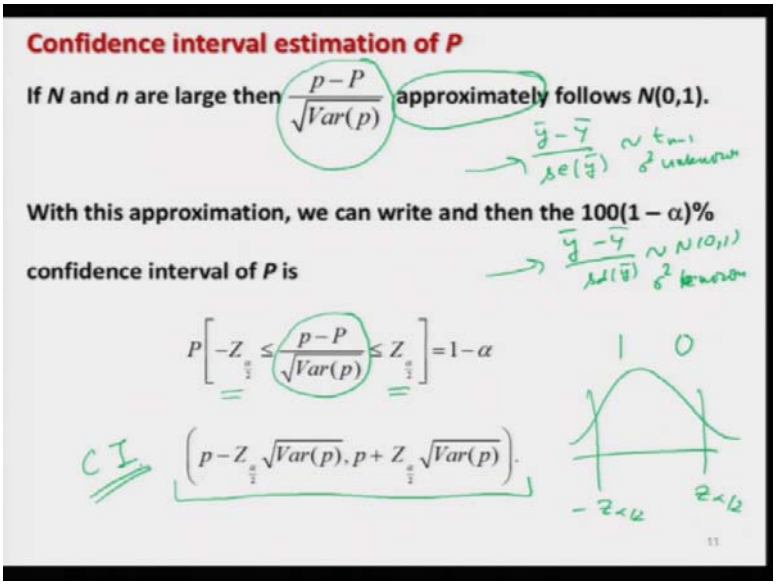
So, this was all about the sample mean or the sample proportion. You can you may also recall that we had discussed the estimation of population total. So, the population two total, we had defined earlier as a Y_{total} which was $\sum_{i=1}^{N} Y_i$.

So, corresponding to this the $\bar{Y}$ is defined here as a Y_{total} /N. So, Y total can be defined by here N times here $\bar{Y}$ and if you want to estimate here, Y_{total} then N will remain as such and $\bar{Y}$ can be replaced by here sample proportion Np.

So, this is what I am trying to write down here. So, if you want to estimate the population total that can be estimated by N a/n. And its variance we had to discuss that is straight forward that you simply try to multiply the variance of sample mean or variance of sample proportion by N^2.

So, the variance of this $\hat{A}$ can be obtained here directly by substituting the expression what we have obtained earlier. And the estimate of this variance of $\hat{A}$ can be just obtained by estimating the variance of the sample proportion. So, simply try to replace the variance of p by the estimator of variance of a p and multiplied by N^2. So, this is how you can obtain all these results straight forward.

(Refer Slide Time: 19:34)



Now, you may recall that when we had discussed the confidence interval estimation for the population mean in case of SRS, then we had two cases that σ^2 is known or $\sigma^2$is unknown. At that moment, I had discussed the result that ($\bar{y}$ - $\bar{Y}$)/standard error($\bar{y}$), right.

This follows a t distribution with n - 1 degrees of freedom and if and if σ is unknown and the other result what I discussed that ($\bar{y}$ - $\bar{Y}$)/standard deviation($\bar{y}$) which follows a normal distribution with N(0, 1) with when σ^2 is known.

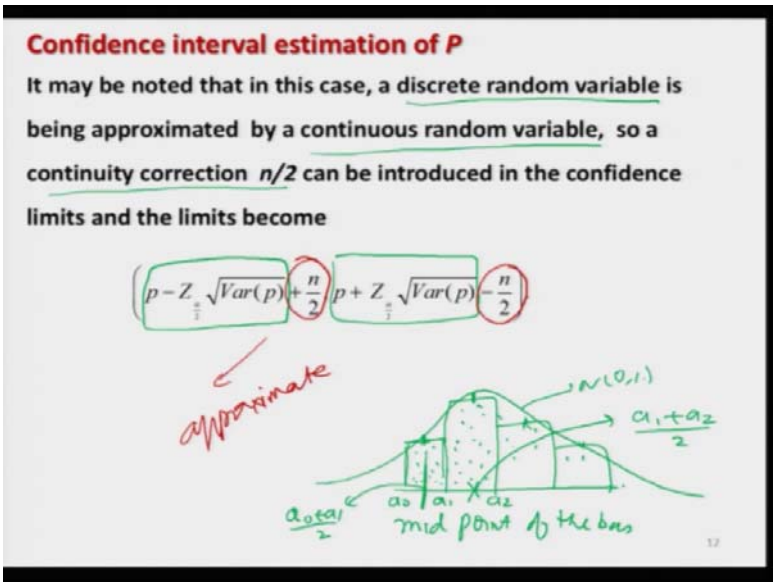
But in the case of qualitative characteristic, we have a problem. These two results they are valid where all also when y i's are you quantitative variable, but in this case your y_i is a qualitative variable it takes only two possible values; 1 and 0. So, we tried to use here a result that sample proportion - population proportion divided by square root of variance of p that is the standard deviation of p.

This approximately follows a N(0, 1) right exactly it does not follow. And now using the this result exactly on the same line as we discussed in the case of simple random sampling, we can construct the 100 (1 - α)% confidence interval for P.

So, if you remember we had written that the probability that this quantity (p - P) divided by square root of variance of p will lie between - $Z_{\alpha/2}$ and + $Z_{\alpha/2}$ like this one, you can recall. So, this was here $Z_{\alpha/2}$, this is - $Z_{\alpha/2}$. So, and then we had solved it and we had obtained the expression earlier in terms of $\bar{y}$.

Now, that $\bar{y}$ can be replaced by sample proportion p and the 100 (1 - α)% confidence interval for sample proportion can be obtained here by this expression p - $Z_{\alpha/2}$ square root of variance of p to p + $Z_{\alpha/2}$ square root of variance of sample proportion. So, this is your here confidence interval.

(Refer Slide Time: 22:19)



Well, in this confidence interval you have to so, observe one thing actually this is a topic that is taught in the statistical inference also when we are trying to construct the confidence interval for a binomially binomial distributed random variable.

What happens? That you are trying to assume that all the observations are here scattered here like this. And we try to create here a sort of histogram and once you try to join the midpoints of the histogram and if you try to join them by a smooth curve, this will look like leaf.

So, we assume that the distribution of p - P divided by square root of variance of p will follow a normal distribution. So, you can see here inside this bar the observation are scattered at different places within the bar, but we assume here that all the observations are concentrated at the midpoint of the bar.

So, all the observations which are lying in the interval a_0 to a_1, they are essentially lying at a $0 + a_1/2$; that is our basic assumption similarly if you try to take here another bar from a_1 to a_2. So, all those observations are here, but we try to assume that all the observations are concentrated at the midpoint of the interval that is $a_1 + a_2/2$.

So, in order to do this thing in the case of a discrete random variable, we try to make a correction. Correction is for what? Because we are trying to approximate the confidence interval of our discrete variable by a continuous random variable; so, we apply here a continuity correction right. So, this continuity correction can also be introduced in the case of the confidence interval when we are trying to estimate the population proportion by sample proportion.

So, in the same confidence interval that we have just obtained, we have obtained this thing this as the lower limit and this as the upper limit. But we try to add here a quantity $+n/2$ in the lower confidence limit and we try to subtract the quantity $n/2$ in the upper confidence limit.

So, this becomes your confidence interval in case of sample proportion, but remember this is a sort of approximate one whereas, in the case of the quantitative variable, you had obtained the exact values of the lower and upper limits of the confidence interval, right. So, in this lecture now you see we had obtained all the results very easily very quickly.

So, this should give you a confidence that why we had spent so much of time in the case of simple random sampling and now you will see that later on when we try to consider another sampling or any other sampling which I am not considering here suppose, you want to study it themselves, suppose you want to study yourself; then the same concept will help you there and you will very quickly understand any type of sampling.

Now, the question is how this type of sampling scheme is going to be useful in the case of data sciences. In data sciences we have got a very huge data sets and you see in practice, most of the surveys are concerned with qualitative variables. People simply call your phone. Suppose you have given your car or scooter for servicing after that you will get a call from the customer service that can you please give us a rank whether your whether the quality of servicing was excellent, average or bad.

Sometime they will ask you a question are you satisfied with the servicing and then they will ask you many questions are you satisfied with the cleaning, are you satisfied with the engine tuning etcetera etc. And they are always interested in finding out their estimates that is how they try to make different statement for the advertisement of their firm. They will say last year we had a consumer survey in which 99% people were satisfied.

So, how do you get all these things or how do you estimate the proportion of the people who are happy with the service quality of the vehicle? Similarly, whenever you do any shopping on the online website after the material is delivered, you get a call or you get an email that please try to classify your satisfaction level or they will ask you are you satisfied with that delivery boy or are you happy with the packaging of the material, are you happy with the quality of the variable, are you happy with the website that the website was clearly indicating the characteristic of the product.

And these numbers are huge they can be in millions or billions. So, you will need a procedure or a direct automated procedure to compute the satisfaction level or the number of persons who are agreeing, number of persons who are not agreeing and so on. So, this sampling for proportion and population, this is a major part of any sample survey in data science and that is how I can make a link between the classical statistics with the data science.

Now, the next question is this whatever we have done here, how can you use them in the software? So, in order to use them in the software, we will try to use the similar results what we have done in the case of simple random sampling for quantitative variable. But now you have to modify them so, that you can handle the situation of qualitative variable.

So, this topic I will try to cover in the next lecture, but my request will be that if you want to really understand those things, then you need to revise whatever we have done in the case of sampling for proportions and percentages as well as what we have done in the case of simple random sampling particularly the lecture where I had explained the implementation on the R software. So, you try to have a look, get yourself prepared for the next lecture and I will see you in you and I will see you in the next lecture. Till then, good bye.