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 – 21
Simple Random Sampling
Estimation of Population Variance

Hello, 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 will consider once again the topics of a Simple Random Sampling. You may recall that in the last two lectures, we considered the estimation of two population parameters; population mean and population variance.

And what we had done? That we considered the sample mean as an estimator of population mean; that means, you have got a big population and you want to study about the central tendency of the population. That means, what are the point where most of the data is scattered around.

So, you propose that we will try to estimate the unknown population mean by sample mean; that means, you take a small sample, try to find out the arithmetic mean of the observation in that sample and that will act as if that is the value of the arithmetic mean in the entire population.

And then, what you did? You obtained the variance of the sample mean. So, that will give you an idea that when you are trying to use the simple random sampling with replacement and without replacement, then what will be the amount of variability that will be involved, when you are trying to estimate the population mean by sample mean.

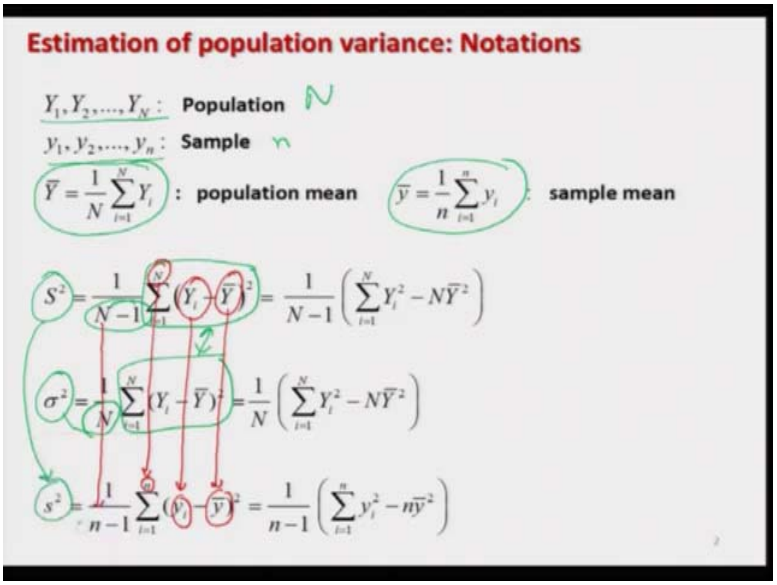
So, we found the variance of the sample mean and I have shown you with some example that when that when you are trying to take different samples, their sample mean will vary among themselves and they will also be deviate from the population mean. That is the difference between the sample mean and the population mean will not always be 0.

So, now, we want to then we wanted to see how close they are; so, that is why we estimated the variance of $\bar{y}$; but variance of $\bar{y}$ had one issue that it was involving the quantity S^2, which is related to the population variance. So, unless and until you know the value of S^2, you cannot estimate the variance of $\bar{y}$, variance of sample mean on the basis of given sample of data. That is my trouble now.

So, now in this lecture, I am going to take up this issue that on the basis of given sample of data, how are you going to estimate the population variance or the variance of the sample mean, ok.

So, once you know how to get it done, then on the basis of sample, by using the sample mean, you can find out the value of the population mean and by finding out the estimate of the variance, you can find out the variability in the population. So, and both of them are going to help you in taking a wise decision right. So, let us begin our lecture.

(Refer Slide Time: 03:42)

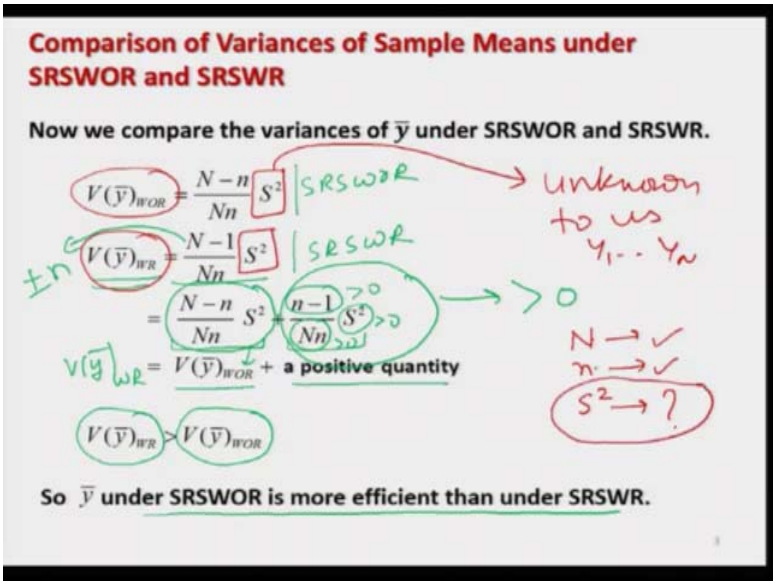


So, just for the quick review, these are our symbols. We had discussed it in the earlier lecture also that by the notation capital Y_1, capital Y_2, …, capital Y_N, we are going to denote the population of size capital N and with small y_1, small y_2,… small y_{nn}, we are going to represent the sample values and the sample is of size small n. Then, capital $\bar{Y}$ is the population mean and small $\bar{y}$ is the sample mean, arithmetic mean.

And we had defined the quantity the S^2 and σ^2 and S^2 has a divisor capital N - 1 and σ^2 has a divisor capital N and these two quantities in S^2 and σ^2, they are the same and after that, we had defined one more quantity a s^2. So, this was a sort of sample part of S^2.

You can see here I it looks like as if I have simply replaced the population values by small values. You can see here capital N is replaced by here small n; this capital N, this is replaced here by small n; this capital Y, this is replaced by here the sample value and this $\bar{Y}$ is replaced by here sample mean right. So, this is our quantity and now we will try to use this s^2 in this lecture.

(Refer Slide Time: 05:17)



So, just for your quick review, we had found the variance of sample mean under SRSWOR that was given by this quantity $\frac{N-n}{Nn} S^2$ and the variance of $\bar{y}$ under SRSWR was given by this quantity. So, now, question arises suppose you want to take a sample from a population, the question is whether you should use simple random sampling with replacement or without replacement?

So, obviously, you would like to choose the sampling scheme in which the population mean is estimated with smaller variability. So, the population mean is being estimated by sample mean. So, now, you have obtained the sample mean under with replacement and without replacement.

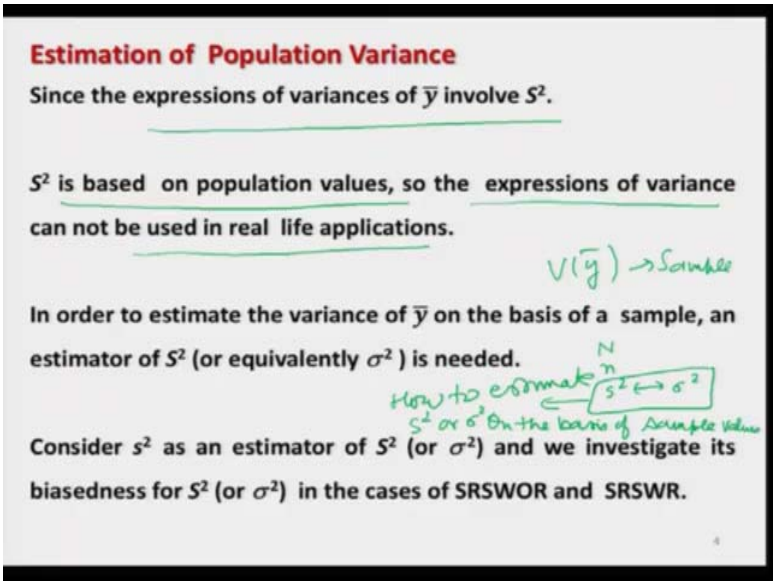
So, whichever variance of $\bar{y}$ either under SRSWOR or under SRSWR, whichever is minimum that should be your preferred choice. So, now what I am going to do? I am simply going to compare the variances of $\bar{y}$ under simple random sampling with and without replacement. So, now, if you try to look into the expressions of this variance of $\bar{y}$, you can see here variance of $\bar{y}$ under with replacement is given by $\frac{N-1}{Nn} S^2$.

Now, what I try to do here that in the numerator, I try to add and subtract here small n. So, I can write down this expression as $\frac{N-n}{Nn} S^2 + \frac{N-1}{Nn} S^2$. And now, you can see here this small n - 1 will always be greater than 0; obviously, mean sample size cannot be 1 and capital N and small n, they are greater than 0 and S^2 is also greater than 0.

So, this entire term is always positive. So, obviously, now what will happen? That this first term $\frac{N-n}{Nn} S^2$, this is nothing but your variance of $\bar{y}$ under WOR. So, now, this variance of $\bar{y}$ under WR is expressed as sum of the variance of $\bar{y}$ under without replacement + a positive quantity. So, now, obviously, this variance of $\bar{y}$ under WR will always be greater than variance of $\bar{y}$ under WOR.

So, this indicates that the simple random sampling without replacement will yield more efficient results than the simple random sampling with replacement, when sample mean is chosen to estimate the population mean. So, that is a very important result. So, in practice as far as possible, we always try to choose simple random sampling without replacement to draw our sample. Why? Because that is going to give us a more efficient result, ok.

(Refer Slide Time: 08:41)



Now, you can also see here that in these expressions variance of $\bar{y}$ under WOR and under WR, we have here the quantity S^2 and this S^2 is unknown to us. Why this is unknown to us? Because this is based on the population value $Y_1, Y_2, \ldots, Y_N$ and we do not know the entire population because if I know the entire population, why should I go for sampling, right.

So, now, one option is this that if you want to use this expression in a real life, this variance of $\bar{y}$ have several quantities. One here is capital N, small n and S^2. Capital N is known to us because this is the population size, sample size is known to us because that is what we have to decide and S^2 is unknown to us.

So, in case, if I want to estimate variance of $\bar{y}$, then one option is this, one possible solution is this that if somehow, I can replace S^2 by some sample-based quantity. So, now, this is what I am going to do now here. So, since the expressions of variances of $\bar{y}$ involve S^2 and S^2 is based on the population value.

So, this expression cannot be used on the basis of a given sample values because and so, in any real-life application, they are not possible to be used. So, what we try to do? My objective is this, I want to estimate the variance of $\bar{y}$. That means, I want to know the variance of $\bar{y}$ on the basis of a sample. I just told you that in the variance of $\bar{y}$, there are three quantities; small n, capital N and S^2.

But you can see here that in case of SRSWR or WOR, you can also express S^2 or say σ^2 by any notation, they are equivalent, right. So, either I say S^2 is known or σ^2 is known, that is the similar thing they are the equivalent statement.

So, now what I need to do? I need to find out an estimator of S^2 or σ^2. So, the question is how to estimate this S^2 or σ^2 on the basis of sample values? At this moment, we try to take the help of statistical inference. In inference whenever we have an unknown parameter and we and if we want to estimate it, we need an estimator.

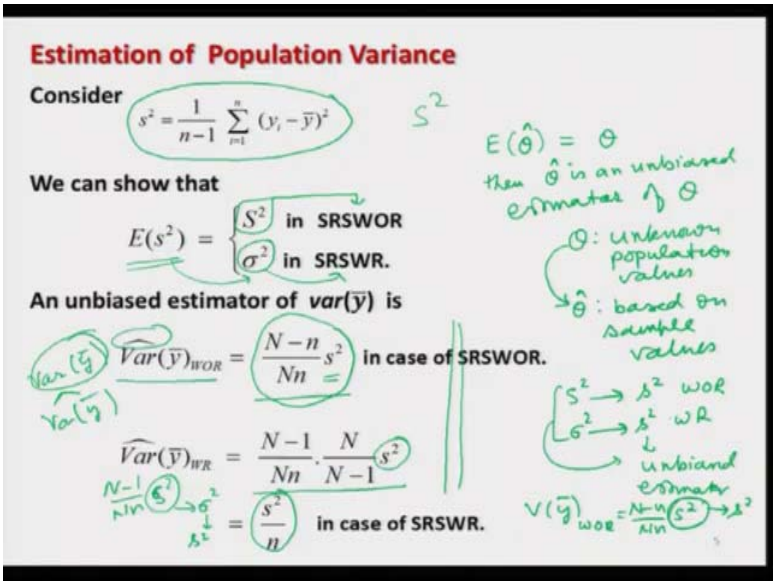
In order to find out an estimator of the population parameter, one approach is that we try to find an estimator or I will say we try to find a possible estimator of the population parameter by guess; some intuition, some guess and then, we try to manipulate it. So, now you can see here that you want to estimate population variance.

So, our natural guess comes out to be- why not I can use the sample variance. Sample variance means whatever the variance has been defined for the population value, the same quantity can be translated to sample values.

So, that is our approach now. So, now we have here two quantities; S^2 and σ^2. So, what we are going to do? We are going to consider here S^2 and the sample version of S^2 is a s^2.

So, what we will try to do? We would try to find out an unbiased estimator of S^2 and to begin, we are going to start with a s^2. So, we will try to consider a s^2, we will try to find out the spectral value of s^2 and based on that, we will try to see how we can obtain an unbiased estimator of S^2. So, this is our approach and this is what I am now going to do through algebra ok.

(Refer Slide Time: 13:23)



So, now we consider a s^2 that I already explained you earlier and if you see this look like as if this is a sample variant of S^2. Now, first I give you the final result and then, I will show you its mathematical proof also. If I try to find out the expected value of s^2, then this will come out to be the same as S^2 in case of SRSWOR and a spectral value of s^2 will come out to be σ^2 in case of SRSWR.

So, what we do in statistics is that if I have an estimator $\hat{\theta}$ to estimate the population parameter θ, then we say that expected value of $\hat{\theta}$ is equal to θ, then we say that $\hat{\theta}$ is an unbiased estimator of θ.

So, this means what? Now, θ is unknown to us and this is based on population values; whereas, $\hat{\theta}$ this is based on sample values. So, we can estimate it, we can compute it on the basis of given sample of data. So, now, I can say if θ is unknown to us, this can be estimated by $\hat{\theta}$.

So, now what I am going to do here? I am now saying that S^2 is unknown to us. So, this can be estimated by a s^2 and in case of SRSWOR and σ^2 is unknown to us. So, this can also be estimated by s^2 in WR case and in both the cases this s^2 is going to be an unbiased estimator, right. So, what I can do now? That we had an expression of variance of $\bar{y}$ for example under WOR as $\frac{N-n}{Nn}S^2$.

So, now, this S^2 can be replaced by s^2 and similarly, in case of variance of $\bar{y}$ under WR, σ^2 can be replaced by a s^2. So, this is what I am going to do here. So, you can see here, here in case of variance of $\bar{y}$ under WOR, we had an expression $\frac{N-n}{Nn} S^2$, now this is replaced by a s^2.

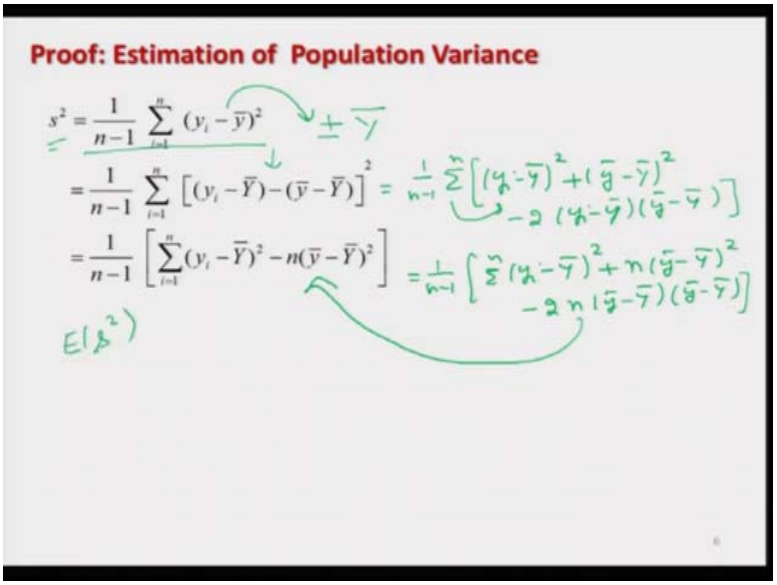
So, now this quantity can be estimated can be computed on the basis of given sample of data in case of SRSWOR. So, now, this becomes an estimator of variance of $\bar{y}$ under WOR and in order to indicate the estimator, we use here a hat. So, I have here two quantities say variance of $\bar{y}$ and variance of $\bar{y}$.

So, now, suppose this is population value, that is unknown; but this is based on sample value, so I will put it here hat to identify that. This is the value which can be estimated on the basis of given sample of data. And similarly, in case of WR, what we can do over here? This expression was $\frac{N-1}{Nn} S^2$.

So, now, in this case if I want to replace here S^2 by s^2, then in the first step, I will try to replace it by σ^2 and then, I will try to replace it by s^2. So, that is what I am trying to do here. So, if you try to see here this is the expression, I simply have replaced σ^2 by s^2 and this comes out to be s^2/n; s^2/n.

So, this is an estimator of variance of $\bar{y}$ under WR. So, this is a quantity which you can estimate on the basis of given sample of data. So, these two expressions which I have written here, they are actually the unbiased estimator of variance of $\bar{y}$ under the two cases; SRSWOR and SRSWR.

(Refer Slide Time: 18:11)



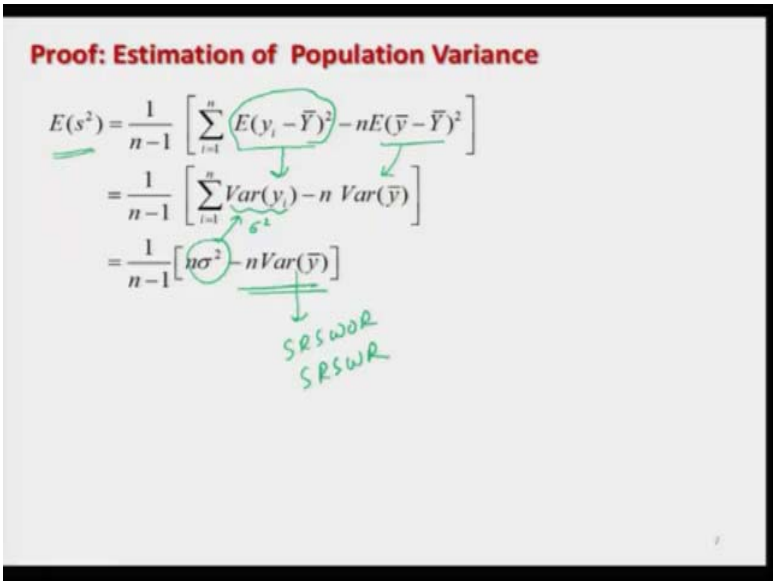
So, now let us try to consider its proof so that you get convinced. So, now, let us consider the quantity S^2 which is here defined by like this. So, what I do there inside this parenthesis in this quantity, I try to add and subtract $\bar{Y}$.

So, this quantity can be written here like this and in case if I try to expand it, so this will become here $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ + N($\bar{y}$ - $\bar{Y}$)² and then, 2(y_i - $\bar{Y}$) and ($\bar{y}$ - $\bar{Y}$).

And in case if you try to take this summation sign inside, so this will become here 1 over n - 1 summation n y i - Y bar whole square + this will become n ($\bar{y}$ - $\bar{Y}$)².

And now, once I try to bring it here, bring here this quantity, so this will become n ($\bar{y}$ - $\bar{Y}$) x ($\bar{y}$ - $\bar{Y}$) and this once you try to solve it, this will give you this expression $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ - n x ($\bar{y}$ - $\bar{Y}$)².. Now, what I do? I try to find out the E(s^2) under two cases and that is WOR and WR, right.

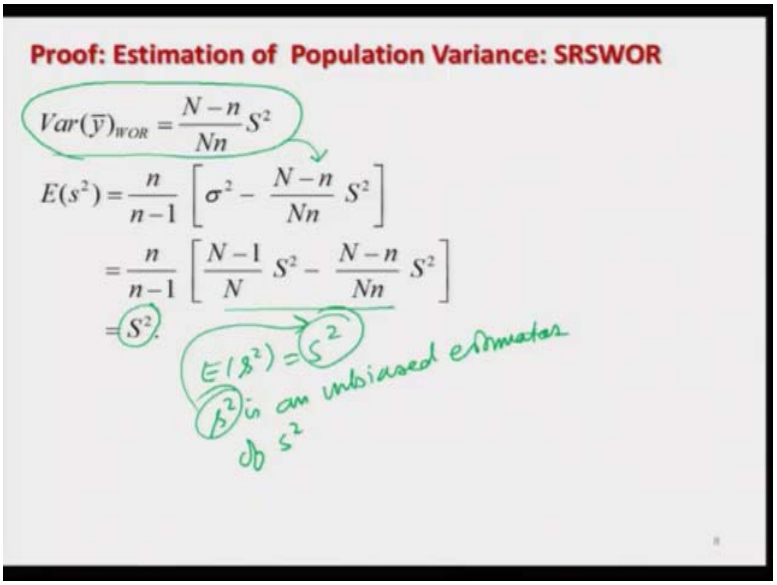
(Refer Slide Time: 19:49)



And so, first we try to find out the expectation of s square. So, this will become $E(y_i - \bar{Y})^2$ here and $nE(\bar{y} - \bar{Y})^2$. So, you can see here this quantity, this is nothing but the variance of y_i and this quantity you have assumed that this is σ^2. So, this becomes here n times σ^2.

Now, for the second term, you can see here E ($\bar{y}$ - $\bar{Y}$). This is variance of $\bar{y}$; a small $\bar{y}$ right. So, this is nothing but the variance of $\bar{y}$, right. So, now, what I try to do? I try to substitute this variance of $\bar{y}$ under two situations, under the case of SRSWOR and under the case of SRSWR and we try to simplify it.

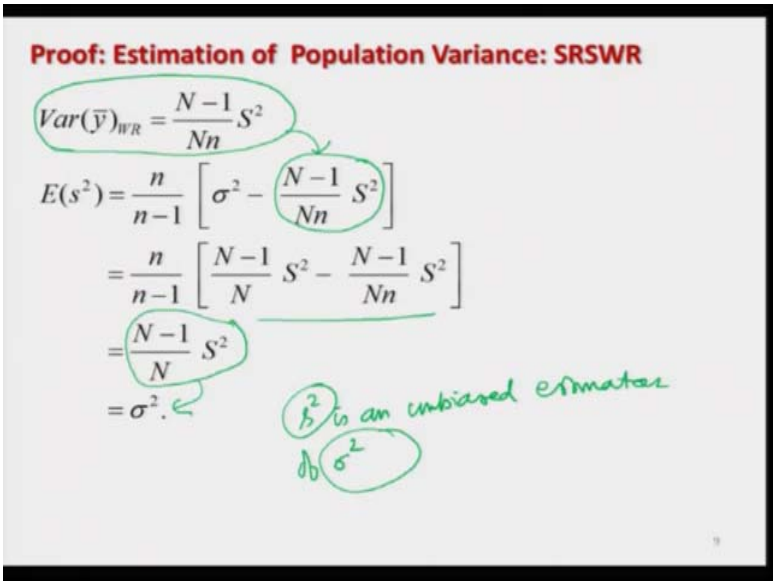
(Refer Slide Time: 20:45)



So, first case variance of $\bar{y}$ under WOR was given by this expression. So, now, what I try to do. So, I try to substitute the variance of $\bar{y}$ here by this quantity and if you simply try to solve it, you can you easily get here S^2. So, I can say here expected value of s^2 is equal to S^2; that means, s^2 is an unbiased estimator of S^2.

That means, well, I do not know here the value of S^2, S^2 because this is based on population; but I have got a sample here Y_1, Y_2,…, Y_N. Based on that, I will try to compute the value of s^2 and this is going to indicate the value of S^2 in the population and the advantage is that it is an unbiased estimator. So, what is the advantage of unbiased estimator particularly in the data sciences, when you are trying to concentrate on the computation part that I had shown you earlier ok.

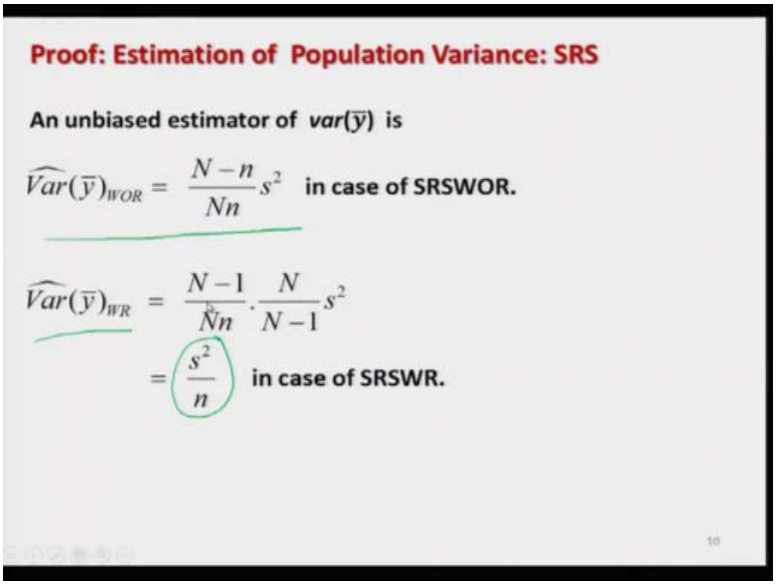
(Refer Slide Time: 21:58)



Now, similarly in case of with the replacement the variance of $\bar{y}$ under SRSWR is given by this quantity. Capital $\frac{N-1}{Nn} S^2$. And now, I try to substitute this expression in the expression of expected value of s^2. So, this variance of $\bar{y}$ becomes here like this and if you simplify this quantity toward here, this comes out to be $\frac{N-1}{Nn} S^2$. So, this is nothing but your σ^2 .

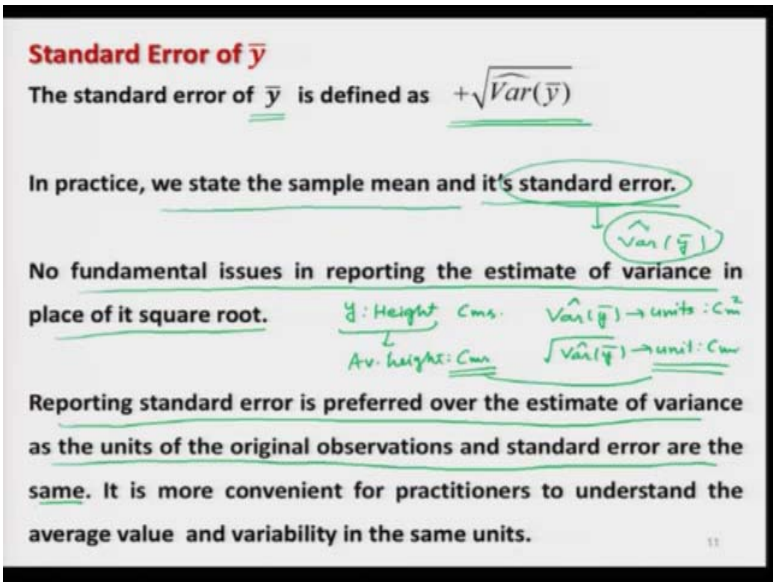
So, now you can see here that in case of SRSWR, this s^2 is an unbiased estimator of σ^2 . So, this means, I do not know the value of σ^2 in the population; but now, I am not worried. I can compute the quantity s^2 from the given sample of data and that will reflect the value of σ^2 in the population right.

(Refer Slide Time: 23:10)



So, you can see here the variance of $\bar{y}$ under WOR can be estimated by this quantity and the variance of $\bar{y}$ under WR can be can be estimated by this quantity right. I would try to address one issue more, but first I try to tell you the result and then, I try to address this.

(Refer Slide Time: 23:36)



Now, the standard error of sample mean $\bar{y}$ is defined as the positive square root of the estimated variance of $\bar{y}$, right and this is called actually as a standard error and the advantage of this standard error is that this is more convenient for the practitioner to use it because in practice, whenever we are trying to conduct say any analysis using this sample mean; then, we always state the value of the sample mean as well as its standard

error right. Well, you also have an option that you that instead of this standard error, you can also report the estimate of the variance of $\bar{y}$, right.

And I would like to make it a 100 percent clear here that there is no fundamental issue in reporting the estimate of the variance in place of its positive square root, right. The only difference is for example, if I say suppose y is pure height some observation on height and they are measured in centimeters.

Now, in case if you try to report the estimate of variance of $\bar{y}$, then its units are going to be in terms of centimeter square. But in case if you try to report the positive square root of $\widehat{Var}(\bar{y})$, then the unit of this quantity, they are going to be in centimeter.

So, now, when you are trying to report the average height right, then this average height will also have units in centimeters. So, now you can see that details of the height in terms of its average value, they are in centimeters as well as the variability is also in terms of centimeters.

So, that is why this reporting of standard error is preferred over the estimate of variance as the units of the original observations and the standard error are the same right and actually, the only advantage is that the practitioners feel more convenient, if they have to report the variability and mean in the same unit right.

For example, suppose if I ask you a simple question that suppose the variability of the variable height is measured in centimeter square. That means, the observations are originally measured in say centimeters. So, now, the unit of the variance will be centimeter square. So, now, if I ask you that the variability of the sample mean is 100 centimeter square and if I ask you that the variability of the sample mean is 10 centimeters. What do you think mean that which is more easy to understand?

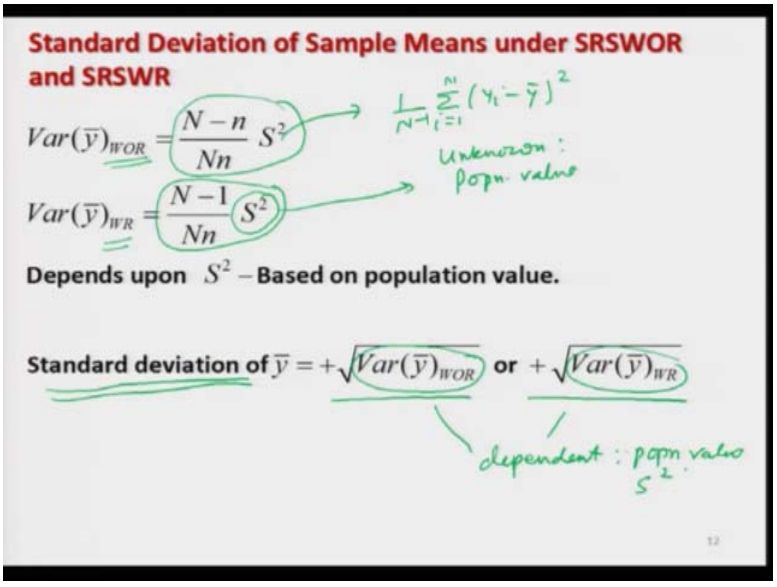
Means obviously, when I say the variability is 10 centimeters, then it is more easy to understand that is the only thing and on the other hand, if I give you two data sets, there are two samples on the heights of the students and suppose I try to measure the standard error and their estimated variance of the sample mean. Suppose, now I give you two statements. The estimated variance of sample mean in the two data sets are say 100 and 196 centimeter square respectively right.

So, you can see the difference looks to be something like 196 - 100 which is 96 centimeters square. Now, I give you second sentence that the standard error of the sample mean in the two data set is 10 and 14 respectively. Now, you can see the difference is only 14 - 10 that is 4 units.

So, as a practitioner for a user, who is not so much familiar with the statistical tool, it becomes easier for the user to understand 10 and 14 instead of 100 and 196. That is the only reason; means, otherwise in practice, whenever you are trying to report your result. So, in case if you are trying to report the average height in terms of centimeter, you can choose whatever you want to report the standard error or the estimated variance, right.

There is no problem at all, but there are some issues and where you have to be more careful, right. So, now, I will try to explain you that what will be the difference between the two and what are the basic fundamental differences from the statistics point of view, ok.

(Refer Slide Time: 28:59)

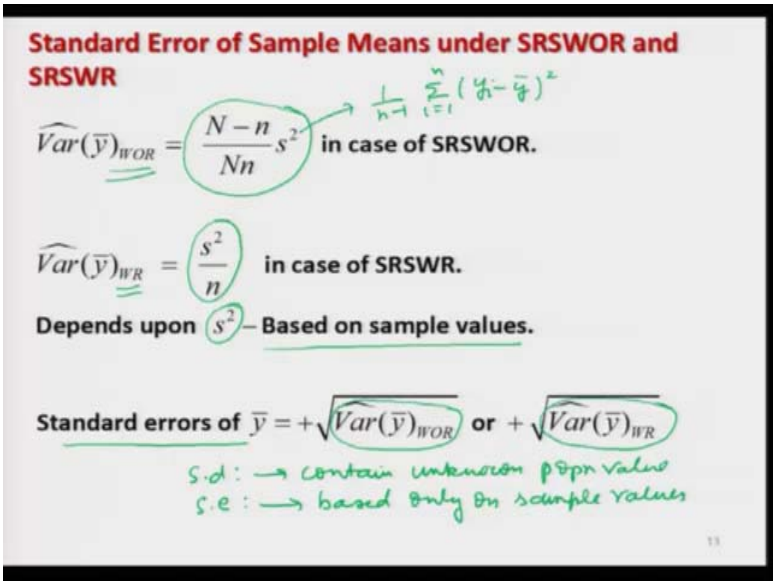


So, now let us see. So, in case if you try to see the variance of $\bar{y}$ under the case of WOR was defined as capital $\frac{N-n}{Nn} S^2$. So, S^2 is going to be $\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$ and now, you know that this is unknown to us that is a population value unknown because this is a population value.

And similar is the case in case if you are trying to use the simple random sampling with replacement, then your variance is given by capital $\frac{N-1}{Nn}S^2$, right. So, S^2 is again going to be dependent on the population and this value is unknown to us, right.

Now, before I move forward, let me try to address one issue more. If you remember that usually people are talking of standard deviation right. So, the first question is what is the difference between the standard deviation and standard errors. Now, in case if you try to take the positive square root of these two variances- variance of $\bar{y}$ under WOR or say variance of $\bar{y}$ under WR, then this is called as standard deviation; but now, you can see that they are dependent upon say population value, S^2, right.

(Refer Slide Time: 30:46)



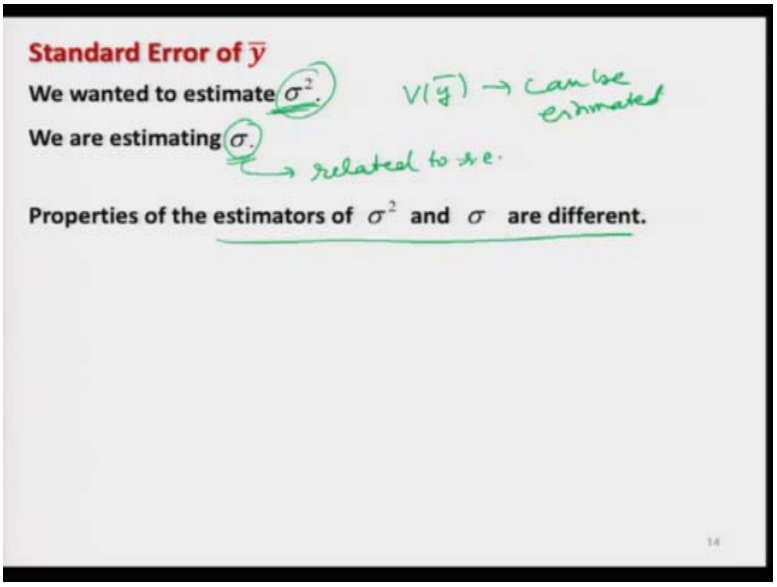
So, now in order to solve this thing what we had done? That we had estimated these variances and an estimate of the variance of $\bar{y}$ under WOR, this was obtained here like this capital $\frac{N-n}{Nn}s^2$; s^2 was $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

So, now, you know now you can estimate this estimated variance on the basis of the given sample of data and similarly, in case of this estimate of variance is given by s^2/n. So, now you can see that these two estimate of variances, they are dependent only on s^2 and this entire quantity is based on sample values and you can estimate it.

Now, in case if you try to take the positive square root of these estimates of the variances, either under WOR or say WR, whatever you want, then this positive square root of the estimate of the variance that is called as standard errors. So, that is the basic difference between the standard deviation and standard error. Usually, the standard deviation, they will contain unknown population value, right and what about the standard error?

Standard error will be based only on sample values, right. And you have to be careful here that in practice many times you will find that people are using these two terminologies, standard deviation and standard error in the same context; but that is fundamentally wrong. But, those people who are not from a statistic background, it is difficult for them to understand; but now you know it, so you must use the correct terminology, ok.

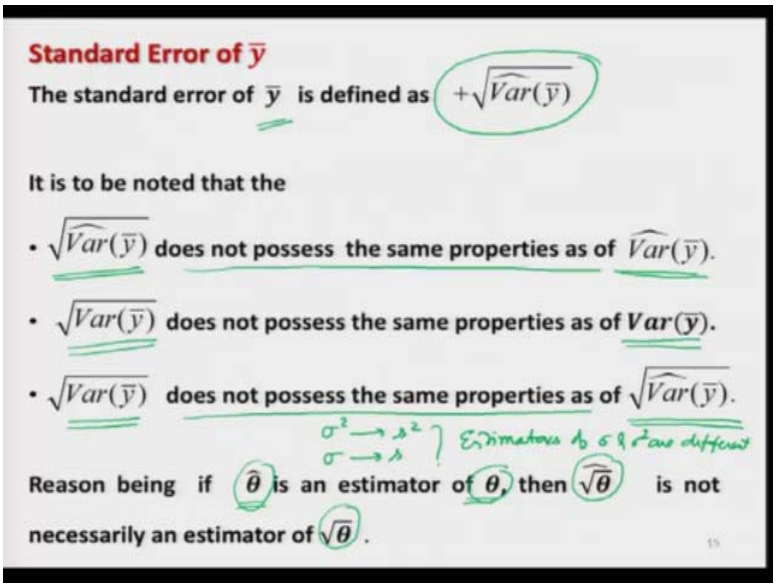
(Refer Slide Time: 32:46)



Now, in case if you try to understand what are we really doing. You wanted to estimate σ^2, right. Means, σ^2 in the sense because once you estimate this σ^2, then the variance of $\bar{y}$ can be estimated right. But when you are trying to deal with the standard error, then you are trying to essentially estimate σ, right. So, this is related to standard error.

Now, if you try to see the estimate of σ^2 and estimate of σ, they are going to be different. They cannot be the same that is the basic difference because of which the difference in the properties of the variance of $\bar{y}$, its standard deviation or estimate or variance of $\bar{y}$ or standard error that comes into picture.

(Refer Slide Time: 33:53)



And there, you have to be little bit careful and you always have to keep in mind that the standard error of $\bar{y}$ will be defined as the positive square root of the estimated value of the variance of $\bar{y}$, right. And you have to note that this value, this standard error does not possess the same properties as of the estimated value of the variance, right and similarly, the square root of the variance of $\bar{y}$ and the variance of $\bar{y}$, they are also two different parameters for us.

So, they will also have different types of properties and when you are trying to estimate the variance of $\bar{y}$ by the standard deviation, so the standard deviation which will be positive square root of the variance of $\bar{y}$ and you are trying to estimate it by the positive square root of its estimator, that is the estimated value of the variance of $\bar{y}$, they will also have different properties.

And they and both of them does not possess the same properties that you have to be very clear in your understanding and there is a statistical fundamental behind these

conclusion. From the statistical estimation point of view, in case if we are trying to estimate a parameter θ by its estimator say $\hat{\theta}$. So, this $\hat{\theta}$ is an estimator of θ.

Now, in case if you want to estimate $\sqrt{\theta}$, then it is not necessarily always possible that the estimate of $\sqrt{\theta}$ is going to be $\sqrt{\hat{\theta}}$. This means in case if you are trying to estimate σ^2 by s^2, it does not necessarily mean that σ can be estimated by s.

I am not saying that σ cannot be estimated, but the estimator of σ and σ^2, they are going to be different. So, I can write down here very clearly that estimators of σ and σ^2 are different and so, they will have different statistical properties.

But what happens? That for a practitioner, who is not very well worse with this concept, they try to use them in the same sense ok; my suggestion is that you should know it that what type of mistakes you are trying to create; means, I agree that in practice we are always trying to take the positive square root of the estimate of the variance. But that is not really going to estimate the value of σ; that can be close to σ that is acceptable, right.

So, now, I have explained you how to estimate the variance of $\bar{y}$ from the on the basis of given sample of data. So, now, up to now I can comprehend that we had the parameter $\bar{Y}$, we wanted to estimate it that was our population mean. So, that we have estimated by sample mean. You also had population variance, so that now you also have estimated it.

So, whatever estimator you employed to estimate the population mean, now you have its value as well as its standard error; that means, that value will give you an idea how much is the variability involved in the estimated value and definitely, the value which has got a smaller variability that will be preferable. You can recall that earlier I had taken the example. In that example, I had taken two values. Suppose, one set of the data has two values 1 and 100. So, the arithmetic mean will be 49.5 which is close to say 50.

Now, I have another data set. There also we have two values; one value is 49 and another is 51. So, now if you try to find out the average that is again 50. So, the average values of both the samples are nearly the same 49.5 and 50, then which data set you would like to

use, that is the question and now, that is answered with the estimate of variance or the standard error of $\overline{y}$.

You will simply try to find out the standard error of the two data sets and whichever standard error is smaller that should be your choice, that is more preferable. And you can see it means you have two values 1 and 100 which are this much apart, there is a lot of variability between the two values; whereas, you have two values say 49 and 51 which are pretty close to each other.

So, the scatteredness is very very small. So, obviously, in statistics we have a rule that any data set or any statistics which is resulting in the outcome with the smaller variance, that is preferable. So, this standard error will help you in finding out that variability.

So, up to now, I have taken two aspects of estimation of parameter that is the point estimation and there is the point estimation of $\overline{Y}$. means how population mean at a point and my answer was sample mean. Sample mean is a single value, so we call it as a point estimate and then, we also estimated the population variance. So, we obtain the point estimate of population variance and based on that, we have obtained the standard error.

Now, next question is can we estimate these population mean and population variance in the form of an interval? So, that is done under the confidence interval estimation. So, I believe that you have a fair idea what is called confidence interval estimation.

But if not, then I would request you please try to take any statistics book and try to look what is called a confidence interval estimation, that is not difficult. That is my promise to you. If you just say spend half an hour, you will easily understand it and there are many books on statistics, almost every basic book on statistical inference is giving you this thing.

So, in the next class, in the next lecture, I will try to take up the Issue of confidence interval of the population mean. So, I am informing you beforehand. So, that if you do not know, you please come with a quick revision. So, you try to revise your course, try to study, try to practice and I will see you in the next lecture with confidence interval estimation. Till then, good bye.