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 – 20
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 learn the topics of Sampling Theory and Linear Regression Analysis. In this module we are going to continue with the topic of Simple Random Sampling.

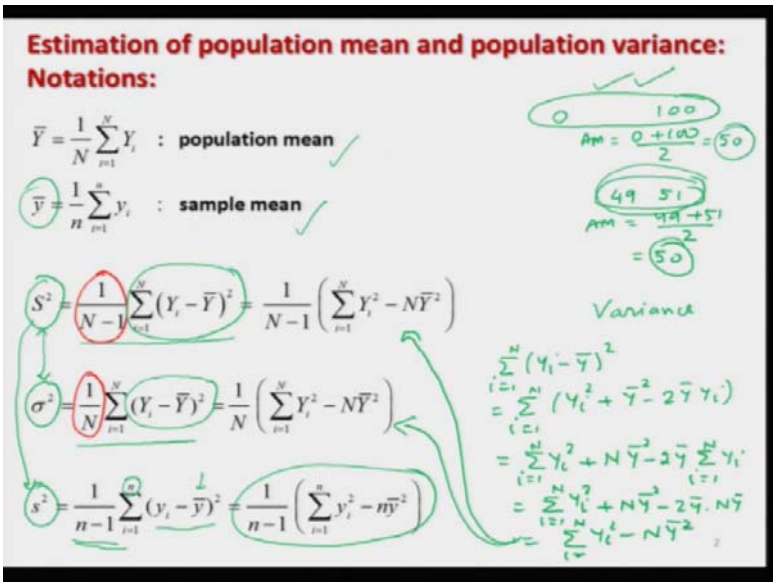
You may recall that in the last lecture we considered the estimation of population mean. And during that lecture, I had communicated with you that there are different types of parameters which are the values inside the population and usually they are unknown to us and our objective in statistics is to estimate them. And in order to estimate them what we do? The first step is that we have to draw a representative sample, how to draw a representative sample? That we already have discussed. So, I believe that since now we are considering the simple random sampling.

So, you have judicially used the simple random sampling to get a representative sample. Now the sample is with you, you already have estimated the central tendency of the data by finding out the sample mean. The value of the sample mean will reflect that what is the possible value in the population, but this is not really sufficient.

We also need to know what is the variability mean and variability the data's is scatteredness as well as central tendency both together will give you some relevant information, how? Let me first try to explain you this concept that why it is important to estimate the variability of the population based on the sample and as discussed earlier we are going to use the variance to measure the variability.

So, in the population we will indicate or we will measure the variability by the tool variance, variance you know. And similarly we will use the sample version of the variance that is, whatever sample you have drawn try to find out its variance. So, using this estimator we will try to estimate the population variability.

(Refer Slide Time: 02:46)



So, the first thing is comes over here why should I do it. So, if I say here that, suppose I get here a sample, suppose I just take here 2 values here say here 0 and 100. So, these are the 2 marks which are obtained by the 2 students in the class they are in my sample. So, if you try to find out the arithmetic mean, arithmetic mean will come out to be 0 plus 100 divided by 2 which is 50 and there is another sample in which 2 students have been drawn and their marks out of 100 are 49 and 51.

So, if I try to compute its arithmetic mean this will come out to be 49 plus 51 divided by 2. And which is again here 50, but do you think that this data set consisting of 2 values 0 and 100 and this data set consisting of value 49 51, are they the same? Because they are giving the same sample mean 50 and 50. So, the central tendency of the both data set is indicating that the population mean is close to 50, but is that really true?

There is a lot of variation in this data set whereas, in this case the values are closely concentrated around 50. So, definitely these 2 data sets are not the same and this gives us a clear indication that estimating only the mean is not sufficient to characterize a population. We also need to estimate other type of parameter.

I am not saying at all that only mean and variance are important all other parameters are also important, but here now we move to next step and now we try to estimate the

variance of the population or the variability in the population, which we are going to measure by the variance quantity, right.

So, these are our symbols and notations as earlier just for a quick review $\overline{Y}$ will denote the population mean based on N value a $\overline{y}$ will indicate the sample mean which is based on a n number of sampling units. And now I am introducing here two more quantities first one is here S^2 and one here is σ^2 and these quantities are defined as $\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2$.

And σ^2 is defined here as say $\frac{1}{N}\sum_{i=1}^{N}(Y_i-\overline{Y})^2$. So, you can see here these two quantities S^2 and σ^2 they are differing only with respect to one factor. The divisor in S^2 is N - 1 and in σ^2 the divisor is N, right. So, I am trying to introduce here these two quantities and you will see that these quantities have different types of roles at a later stage.

And you know that if you try to expand this quantity over here I can show you that $\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2$ this can be written as $\sum_{i=1}^{N}Y_i^2+\overline{Y}^2-2\overline{Y}\sum_{i=1}^{N}Y_i$; and if you try to take the summation sign inside the bracket.

So, this becomes here $\sum_{i=1}^{N}Y_i^2+N\overline{Y}^2-2\overline{Y}\sum_{i=1}^{N}Y_i$ and this can be written as $\sum_{i=1}^{N}Y_i^2+\overline{Y}^2-2N\overline{YY}$. So, this quantity is nothing but your $\sum_{i=1}^{N}Y_i^2-N\overline{Y}^2$ and which is written over here as well as here

So, these are the new symbolic notations and yeah sometime people get confused that or they ask that why there are these two quantities which are differing only with respect to the divisor. So, I can give you some idea for example, if you are considering normal distribution normal probability density function.

So, there we have two parameters μ and σ^2, μ is the population mean and σ^2 is the variance. So, when you try to estimate σ^2 then there are different estimation methods, method of moments, method of least square, maximum likelihood estimation and so on.

So, when you try to use the principle of least square then the σ^2 is estimated by the version of this σ^2 and S^2 by S^2 with the divisor there is 1 upon the divisor there is N - 1.

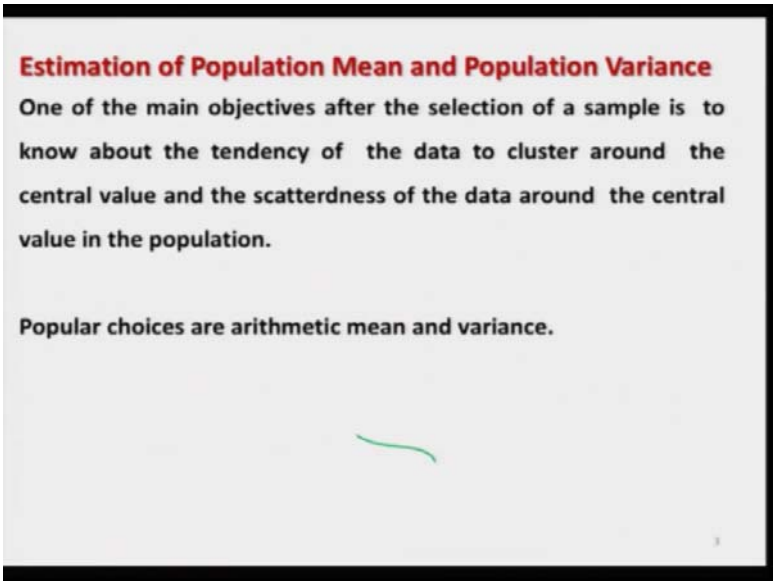
So, the least square estimate of σ^2 will be $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. On the other hand if you try to use the maximum likelihood estimation to estimate the σ^2 then the estimate of σ^2 will come out to be $\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$, right.

So, that is why we have two forms of these estimator and in general I can tell you that the if you try to take the sample versions of the two, then the quantity with divisor N - 1 will turn out to be an unbiased estimator of the population variance; whereas, the maximum likelihood estimator that is the sample variance with divisor N will not be an unbiased estimator, but the variability of the maximum likelihood estimator that is with divisor N will be er than the variability of the least square estimator which has got divisor N - 1.

So, this is a sort of strange situation one estimator which is unbiased, but it has got a higher variability and the estimator which is not unbiased that has got a er variability, but they have their own utility and they are used accordingly, right. So, now I introduce here one more notation that is the sample version of this S^2.

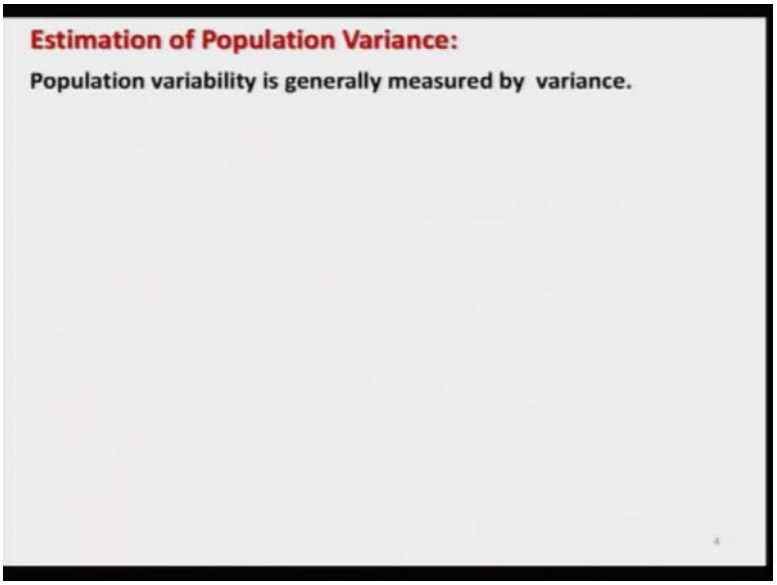
So, you can see here I am considering $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $\bar{y}$ here is the sample mean and yeah this can be expanded exactly in the same way like this. So, these are the notations which I am going to use in the further chapters while estimating the variance and the confidence interval estimation, ok.

(Refer Slide Time: 10:20)

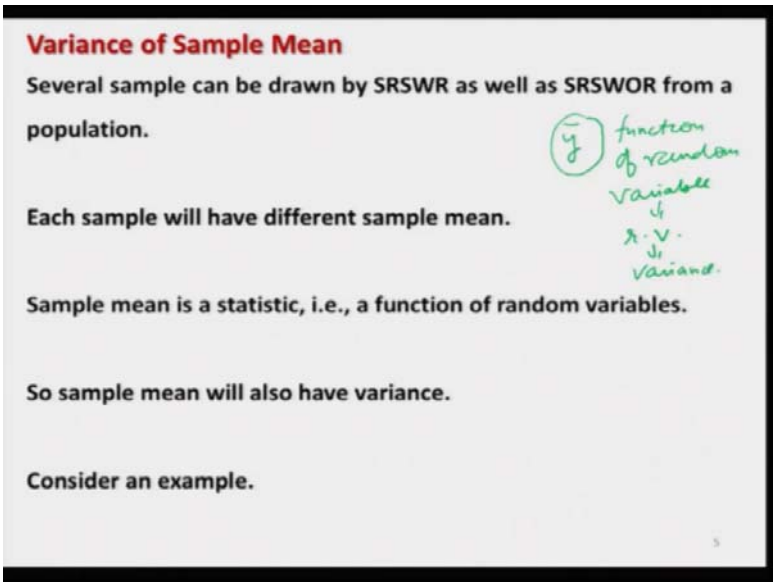


So, now as we have discussed earlier that once we have got the sample one of the important objective is that we would like to measure the central tendency of the data as well as the variability or the scatteredness of the data. And popular choice to find out the population mean is arithmetic mean for the variability this is variance, right.

(Refer Slide Time: 10:41)



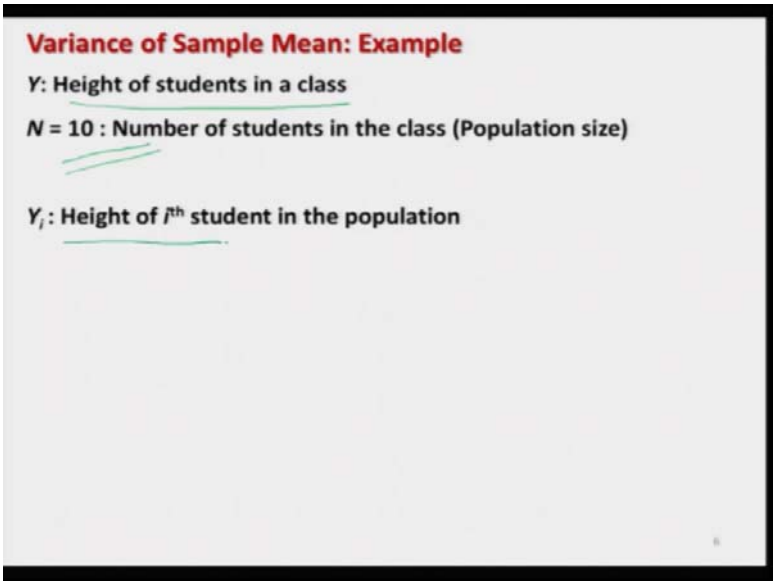
(Refer Slide Time: 10:48)



So, and in general this population variability is measured by variance, now the question is this whose variance we want to consider. So, now, if you see we have considered $\overline{y}$ which is the sample mean as an estimator of population mean. So, this is a function of random variables, what does this mean I will try to explain you soon.

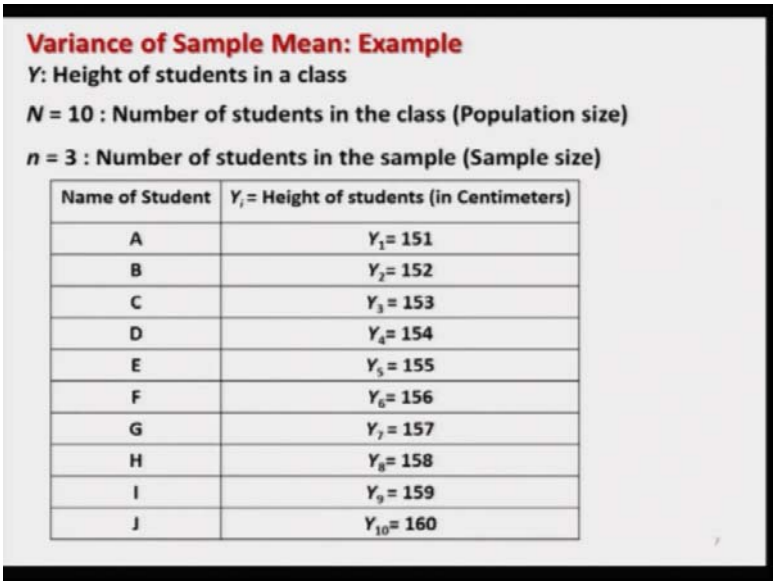
So, now, since this is a function of random variable, so this itself is a random variable and since this is a random variable, so it will have some variance. So, we would like to find out the variance of $\overline{y}$, right ok, but what is this variance and how it is coming into picture?

(Refer Slide Time: 11:32)



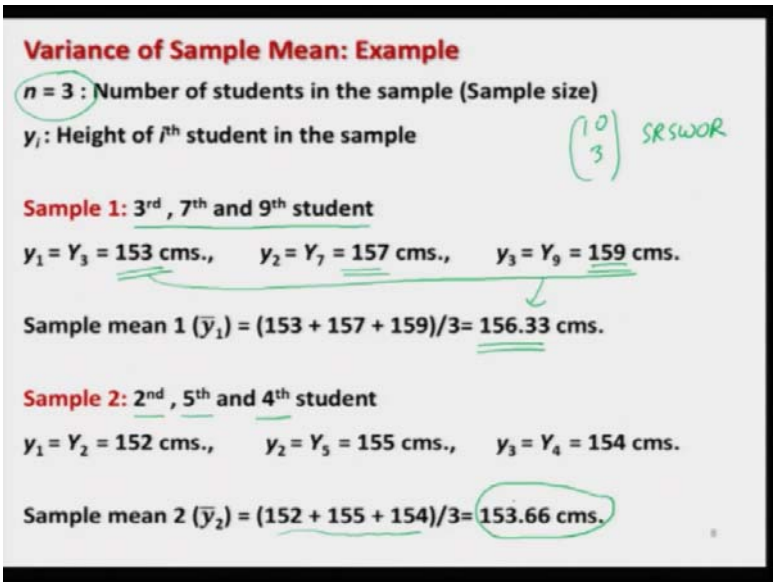
Let me try to take a simple example that we have considered earlier to explain you. So, suppose I take the same example where I am denoting Y as the height of the students and I have 10 number of students in the class and Y_i is the height of the ith student in the population.

(Refer Slide Time: 11:49)



And this is the data what we have obtained we have here 10 students and their heights are given like this. So, I will be quick here because I already have considered explained this example couple of time.

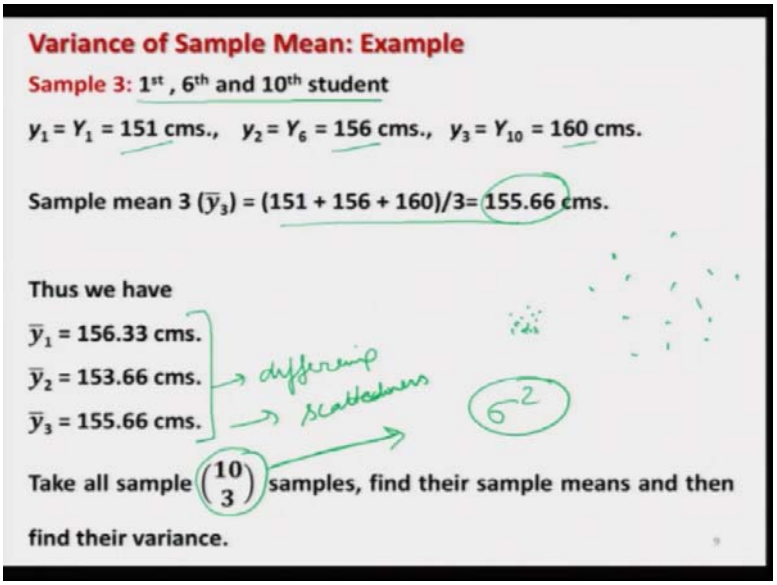
(Refer Slide Time: 11:59)



Now, you see this is the same slide which I use in the case of estimation of population mean also. So, now, you can see here I want to draw samples of size 3. So, you know that there will be $\binom{10}{3}$ number of samples in the case of SRSWOR right. So, now, let me choose only 3 samples here and we try to find out their sample means. So, Sample 1 gives me 3rd, 7th and 9th student and the heights of the 3rd, 7th and 9th strings are 153, 157 and 159 centimetres respectively.

And when I try to find out the arithmetic mean of these 3 values this comes out to be 156.33. Similarly I try to take the second sample where I have 2nd, 5th and 4th students I find their height and then I try to compute their the sample mean of their heights. So, this comes out to be here 152 plus 155 plus 154 divided by 3 which is 153.66 centimetres.

(Refer Slide Time: 13:12)



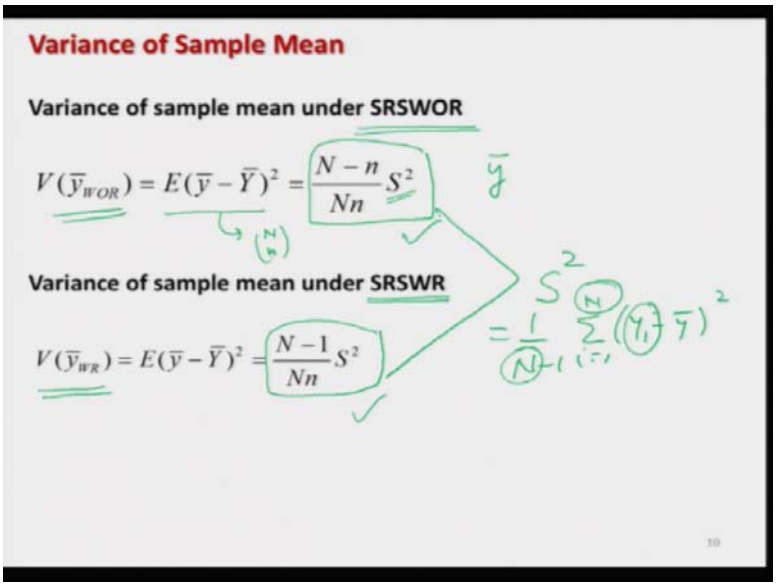
And now I try to take the 1st, 6th and 10th student in my sample and their heights are 151, 156 and 160 centimetre and their sample mean comes out to be here 155.66 centimetre. So, now you can see here that I have got here 3 sample means and you can see here that these 3 sample means are not the same.

Although they are coming from the same population, but they are differing; so, this is exactly what I want to measure, that how much are they differing from each other? Or in simple word, what is their scatteredness? For example, I can have here $\binom{10}{3}$ possible samples large number of sample and if I try to plot them here are they lying close enough or are they lying quite far enough.

This is what we want to explore, this is another feature of the data which we would like to explore and my objective is this whatever is this variability we are trying to consider the variability of the entire population which is whose variance will be something say σ^2, but it is unknown from that population I can draw $\binom{10}{3}$ samples, but I am choosing here only 1 sample.

So, we are working only with 1 sample. So, that is why looking at the value of the variance of 1 sample we have to judge for the variability of the entire population that is my trouble.

(Refer Slide Time: 14:59)



So, first I try to give you the expression that what will be the variance of sample mean under WR and WOR cases and after that I will try to give its derivation so that you get convinced that these expressions are correct. So, when we are considering the simple random sampling without replacement.

Under this situation when we consider the sample mean $\bar{y}$ then the variance of $\bar{y}$ under WOR is essentially $E(\bar{y} - \bar{Y})^2$. And yeah, the meaning of this thing I had already explained you that there will be $\binom{N}{n}$ samples you try to find out all the samples try to find out there all the variances and then try to take their arithmetic mean and this mean will be the population value.

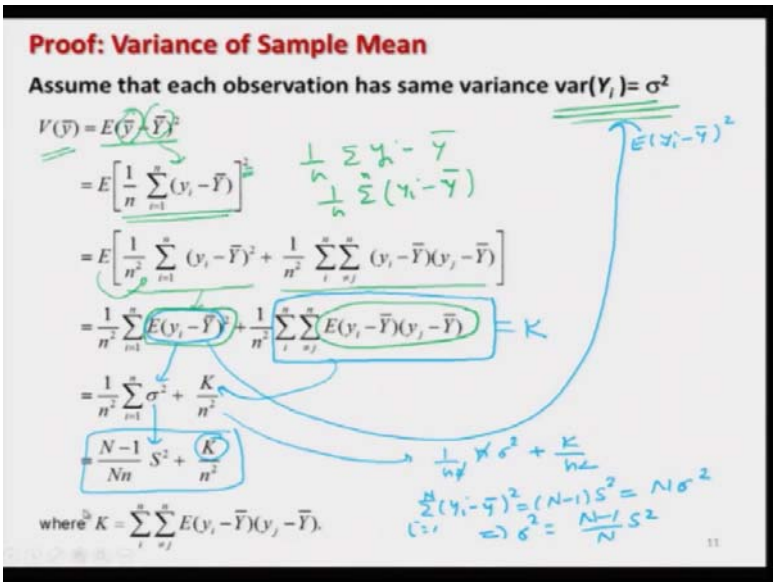
For this variance turns out to be like this (N – n)/ Nn that is N - n divided by N into n times S^2 and S^2 was like a population variance with a divisor N - 1. And when we are using simple random sampling with replacement in that case the variance of the sample mean turns out to be (N – 1)/Nn S^2.

So, these are the two expressions, but remember one thing in these two expressions we have a quantity here S^2 which is $\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. So, this is depending on the

population values and these values are not known to us. All the population values Y_1, Y_2,…, Y_N are not known to us.

So, although I am finding out here the variances to be like this, but they cannot be used on the basis of the sample they cannot estimate. So, first I am going to explain you what is the variance of $\bar{y}$ and then later on I will show you that how to estimate these two variances under SRSWOR and SRSWR conditions, right ok.

(Refer Slide Time: 17:26)



So, let me first give you the derivation of finding out the variance of the sample mean. So, in the beginning I will take a general expression and then I will divert to SRSWOR and WR cases. So, we assume that in the population the variance of every unit is the same that is σ^2, right.

So, now, you can see here this relates to one of the observation which I made couple of lectures back that when you have to choose a sampling scheme, then how are you going to decide that which of the sampling scheme is usable under what type of condition. At that moment I explained you that simple random sampling is more useful where the variability is uniform across the population.

So, this is what I am trying to say here that the variance of every unit in the population is σ^2. So, it should not happen that in some parts of the population the variability is very high and in some part of the population the variability is extremely low.

So, now if I try to find out the variance of here $\bar{y}$; so, this is nothing but $E(\bar{y} - \bar{Y})^2$. So, I try to substitute here the values of a $\bar{y}$ and $\bar{Y}$. here. So, if you try to see here just by substituting the value of a y bar, I can write down this quantity here like this right.

This is the same thing here $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})$ and if I try to take out this outside the bracket $\frac{1}{n-1}\sum_{i=1}^{n}$ this becomes here $y_i - \bar{Y}$, right. So, now, this is here like this and if I try to this is here a square quantity.

So, if I try to open the bracket this will become the square quantity plus its cross product. So, this becomes here $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and plus $\frac{1}{n^2}\sum_{i}^{n}\sum_{\neq j}^{n}(y_i - \bar{Y})(y_j - \bar{Y})$. So, that is a very simple rule and now I try to bring this expectation operator inside the bracket.

So, you can see here this quantity can be written here as a like this now this $(y_i - \bar{Y})^2$ will become expected value of $(y_i - \bar{Y})^2$ this is what you have to keep in mind. And the second term when the expectation sign when the expectation operator comes inside the bracket this becomes your $E(y_i - \bar{Y})(y_j - \bar{Y})$.

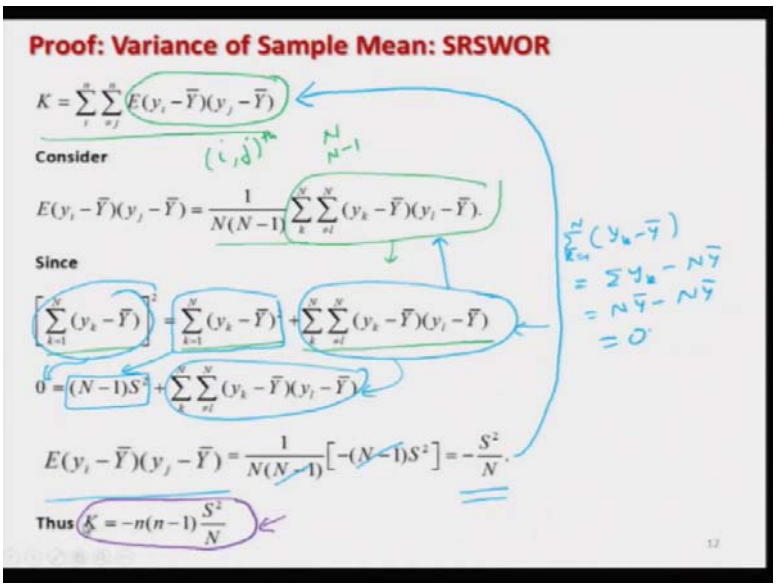
So, let us just assume that whole this quantity this is suppose equal to here K, and what is this quantity here? $E(y_i - \bar{Y})$ this is the same thing which you have assumed here variance of Y_i is nothing, but $E(y_i - \bar{Y})$, right. So, this the same quantity and that you have assumed that this is σ^2.

So, I can replace here $E(y_i - \bar{Y})^2$ by σ^2 and this quantity here by K. So, now, this quantity becomes here $\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 + \frac{K}{n^2}$. So, this is here σ^2 / n and you can see here that

there is a relationship between σ^2 and S^2 because $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ this is same as (N – 1)/(N – 1) X S^2 and this is same as N X σ^2.

So, σ^2 is the same as (N – 1)/N X S^2. So, I try to replace here this σ^2 by this here S^2. So, now, I have this expression here like this right. Now, what I will do? I will try to consider two cases for SRSWR and SRSWOR and will try to find out the value of K and then I will substitute it here to find out the exact expression for the variance of $\bar{y}$ under SRSWR and WOR.

(Refer Slide Time: 22:00)



So, first I consider the case SRSWOR. So, K is here like this. So, now, first I try to compute this quantity $E(y_i - \bar{Y})(y_j - \bar{Y})$. If you try to see what is this quantity indicating this is trying to indicate that two units i and j they are drawn one by one by SRSWOR.

So, when ith unit is drawn there are N units in the population and when the jth unit is drawn in the second draw ith unit is drawn in the first draw and jth unit is drawn in the second draw, then in the second state there are N - 1 units available in the population. So, the probability of collecting of y_i and y_j will be $\frac{1}{N(N-1)}\sum_{k}^{N}\sum_{\neq l}^{N}(y_k - \bar{Y})(y_l - \bar{Y})$.

Now, I just need to simplify this particular expression. So, you know that is a very simple formula that possibly you have learnt in class 10 or class 12 that $\left[\sum_{k=1}^{N}(y_k-\bar{Y})\right]^2$ can be written as $\sum_{k=1}^{N}(y_k-\bar{Y})^2+\sum_{k}^{N}\sum_{\neq l}^{N}(y_k-\bar{Y})(y_l-\bar{Y})$, right.

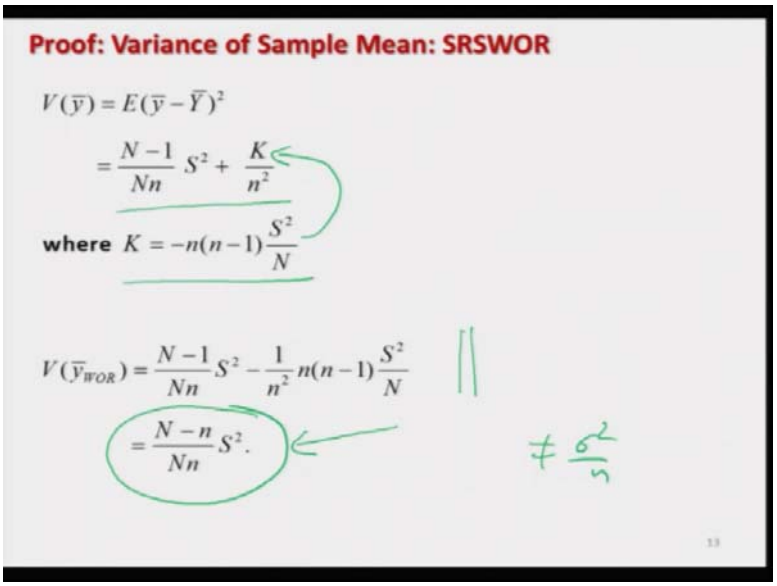
So, now, you can see here that this expression what I have written here this is the same expression like here this thing. So, I will try to find out the value of this expression and then I will replace it there, but before that you also know that if I try to take here $\sum_{k=1}^{N}(y_k-\bar{Y})$ this is $\sum_{k=1}^{N}y_k-N\bar{Y}$ and summation $N\bar{Y}-N\bar{Y}$ which is equal to 0.

So, that is why this quantity inside the bracket becomes here 0. So, 0 square is 0 and this quantity here $\sum_{k=1}^{N}(y_k-\bar{Y})^2$. This I can write as N Xσ^2 or (N – 1)XS^2. So, I am preferring here to write this quantity as (N – 1) S^2 and yeah this quantity as such.

So, now from there I can find out this quantity comes out to be here $\frac{1}{N(N-1)}\left[-(N-1)S^2\right]$. So, this N - 1 on and this N - 1 in the numerator and denominator they get cancel out and I have the quantity here - S^2/N.

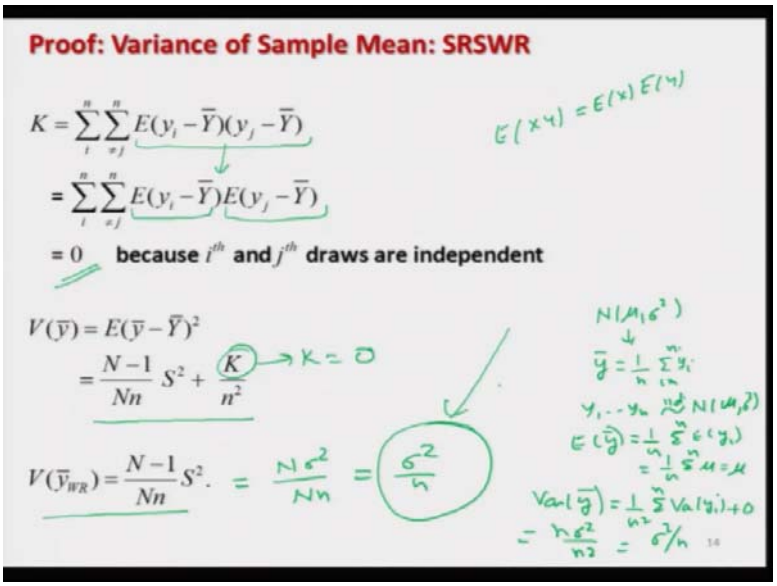
So, this quantity can be brought back here and then if you try to solve it this K turns out to be here like this; I will try to use a different color so that you can identify this here like this right. And now I can substitute this quantity what I have obtained here as K under the case of SRSWOR in the expression of variance of $\bar{y}$.

(Refer Slide Time: 25:31)



So, the variance of $\bar{y}$ which I had obtained in general was $\frac{N-1}{Nn} S^2 + \frac{K}{n^2}$. Now in the case of SRSWOR, the value of K is like this $-n(n-1)\frac{S^2}{N}$. So, if I try to substitute this quantity to quantity over here and here I just make a calculation and the variance of $\bar{y}$ under SRSWOR comes out to be like this $\frac{N-n}{Nn} S^2$, right. So, this is the variance of $\bar{y}$ under SRSWOR.

(Refer Slide Time: 26:12)



Now, I try to find out the value of K under SRSWR. So, under simple random sampling with replacement we know that all the units are independently drawn in the sense that they are replaced back in the sample. So, this K contains the quantity $E(y_i - \bar{Y})(y_j - \bar{Y})$.

So, we know from the rules of statistic that if two random variables are independent say if two random variables are x and y then they can be written as E(x) and expected; the E(x y) can be written as E(x) X E(y). So, I try to use this result over here.

So, this quantity can be written as $E(y_i - \bar{Y})E(y_j - \bar{Y})$ and this $(y_i - \bar{Y})(y_j - \bar{Y})$, I already have shown you here that this is 0. So, this quantity K becomes here 0 in case of SRSWR.

So, the variance of $\bar{y}$ which we had obtained earlier like this here I have to simply substitute K is equal to 0 and then this term turns out to be like this $\frac{N-1}{Nn}$ S^2, right. And if you want to write it in some books you will find it because (N – 1) S^2 can be written as N σ^2 /N n.

So, this can be written as σ^2 /N and now I will ask you do you remember? Can you identify? Or do you remember that that how this term is resembling? Right.

If you remember when you try to consider a probability density function like normal μ σ^2 under that case you try to find out the sample mean, sample mean is the arithmetic mean of say all the sample values i goes from 1 to N. And you assume that all Y_1, Y_2,…, Y_n whatever is your sample they are i i d from N(μ , σ^2).

Now if you try to find out here expected value of $\bar{y}$ this will come out to be here $\frac{1}{n}\sum_{i=1}^{n} E(y_i)$ which is equal to here $\frac{1}{n}\sum_{i=1}^{n} \mu$ which is equal to here μ and variance of $\bar{y}$ this comes out to be say $\frac{1}{n^2}\sum_{i=1}^{n} \text{var}(y_i)$ and plus covariance terms becomes 0.

So, this comes out to be nothing, but your $n\sigma^2$ /n^2 which is σ^2/n. So, now, do you remember that earlier when you studied the statistical inference in that situation in those cases you have been taught that the variance of $\bar{y}$ is σ^2/n, but here in the case of SRSWR you can see that this is matching, but in, but if you come to SRSWOR this is not matching, this is not the same as σ^2/n. So, why this is happening? Right.

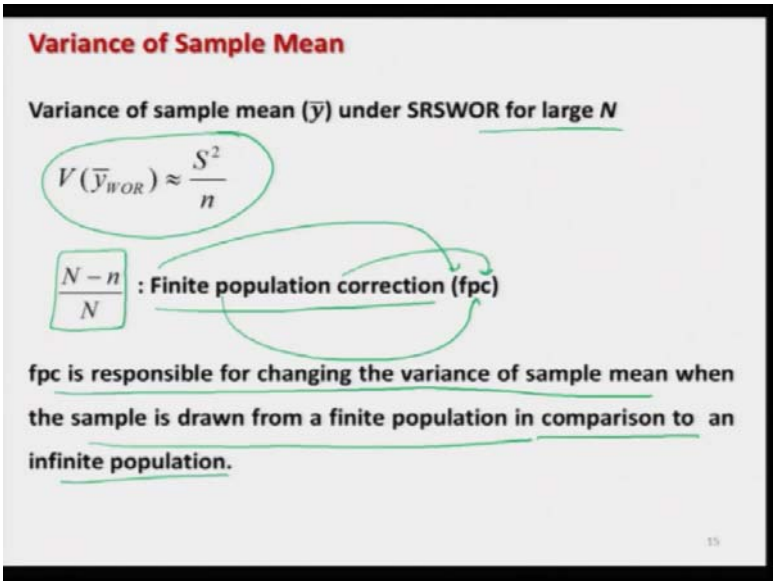
So, in this case actually my explanation goes like this when you are trying to consider the situation of N(μ, σ^2), where the observations are i i d from N(μ, σ^2), you are trying to draw a sample from an infinite population; whereas, when you are trying to deal with simple random sampling or the sampling theory usually our first attempt is that our population is finite.

So, that is why this difference is coming. So, when you are trying to take a sample from a finite population and when you are trying to take a sample from an infinite population they are going to be different and this difference is coming here. The variance of $\bar{y}$ that you studied during the statistical inference that was actually a sample of size n is drawn from an infinite population.

But in simple random sampling you are trying to draw the same size of random sample, but from a finite population and that is the reason that we need a correction factor here and this correction factor can be logically called as finite population correction. As the name suggest that when we are trying to consider a finite population then I have to make

this correction in the variance. So, this is what happened here now I try to address this concept over here, right.

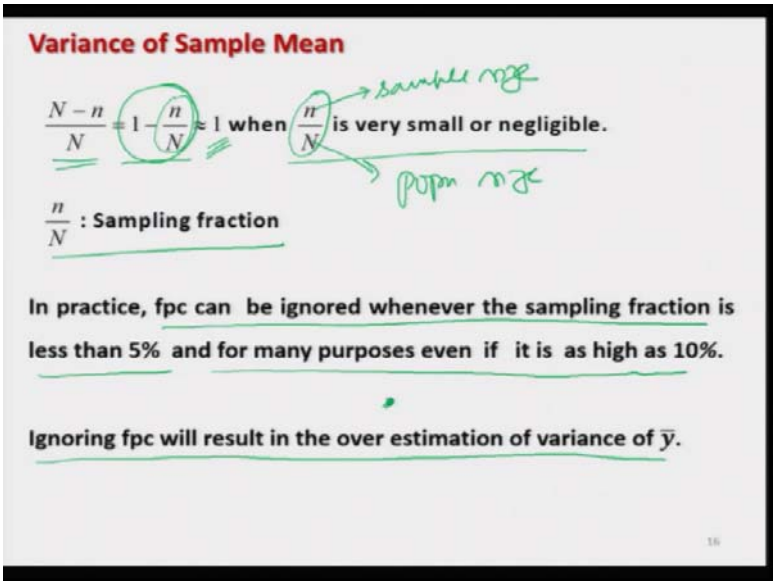
(Refer Slide Time: 31:10)



So, whenever N is large in that case the variance of $\bar{y}$ under WOR can be approximated by this quantity S^2/n. And this factor here (N – n)/ N this is called as finite population correction and briefly that is called as fpc, f stands for finite p stands for population and c stands for correction.

So, this fpc is actually responsible for changing the variance of sample mean when the sample is drawn from a finite population in comparison to an infinite population, this is exactly what I explained you, right.

(Refer Slide Time: 31:55)



And if you try to look at the structure of this finite population correction, this is this can be written as 1 – n/N. Now if you try to look at this structure suppose if this factor n/N is very then it is close to 1. So, I can say that this fpc is close to 1 when n/N is very or say negligible.

And if you try to look at the structure of n/N, what is this? n is the sample size and N is the population size. So, if you try to see this is the ratio of the sample size and population size and this is called as Sampling fraction. So, this factor actually helps us in a deciding in real situation that whether the finite population correction has to be ignored or not whether a population can be considered as a finite population or a or an infinite population this factor can help us.

Well, there is no strict condition, there is no rule of bible that well this must happen, but in practice based on the experience this has been suggested that, this fpc can be ignored whenever the sampling fraction is less than 5 percent and for many purposes even if it is as high as 10 percent even then it can be ignored.

And definitely ignoring fpc will result in the over estimation of the variance of y, please ignore this $\bar{y}$, right, but the thing is this that you have to simply see how much is the difference. And if the difference by considering or not considering fpc is not very much possibly the population can be considered as sufficiently large right, now I stop here with the estimation of variance.

So, now I have told you I have explained you that how can you estimate the population mean, how can you estimate the population variance from the theory. Now the thing is this, how are you going to estimate them? Estimation means whatever you have obtained here variance of $\bar{y}$ they contain the term S^2 which is based on the population.

So, this from this expression you cannot find out what is the variability in the population based on the sample. So, on the next end I will try to consider the estimation of variance and after that I will show you that how these quantities can be estimated on the basis of R software. So, you try to revise it, try to understand it, try to think about it and I will see you in the next lecture, till then good bye.