

**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 - 24**  
**Simple Random Sampling**  
**Estimation of Mean, Variance and Confidence Interval in SRSWR using R**

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. And in this module we are going to continue with the Sampling Theory with R Software, and we are going to discuss the topic of Simple Random Sampling with R Software.

So, you may recall that in the earlier lecture that how are we going to compute the mean, standard error of sample mean as well as its variance and the confidence interval estimation in the R software. Now, the similar thing I will try to do here, but in the case of simple random sampling with replacement.

So, most of the basic fundamental concept behind the computation, I already have explained you. So, in this lecture I will try to be quick, because same concept, same rule, same notations everything is the same, they will be following here, ok. So, let us try to begin our lecture.

(Refer Slide Time: 01:23)

```
Using R software: Confidence interval of  $\bar{y}$  in SRSWR
First we define a population units containing the numbers 1 to 20.
This can be defined by a sequence as x.
> x <- 1:20
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Let us define ywr as a sample data vector using sample statement
in R.
(wr indicates that the sample is drawn with replacement)
> ywr <- sample(x, size=5, replace = TRUE)
> ywr
[1] 10 17 18 1 18

In ywr vector we can see that the value 18 is repeated two times.
```

So, now I will try to create here the same population, I will try to consider the same what we considered in the earlier lecture that is the numbers from 1 to 20. So, suppose there is a population of size 20 which is denoted by  $x$ . Now, I use the command here `sample` and I try to draw here a sample of size 5 from this data vector, which is population here  $x$ .

And here since I am using simple random sampling with replacement, so, I will use here `replace` equal to `TRUE`. So, now, this is my here sample which I have got, when I was preparing my slides. So, you can see here that these two values 18 and 18 they are repeated here. So, they are repeated two times ok.

(Refer Slide Time: 02:11)

**Using R software: Confidence interval of  $\bar{y}$  in SRSWR**  
**Case 1: When population variance ( $\sigma^2$ ) is known**

The estimates of the population mean is the sample mean. For a data vector  $y$ , it is found by the statement `mean(y)` in R.

The estimates of the variance of sample mean of the sample of size  $n$  data in  $y$  is found by the statement `var(y)/n` in R.  
 $\hat{var}(\bar{y})_{WR} = \frac{s^2}{n} \rightarrow var(\bar{y})$

Note that the command `var(y)` gives the value of  $s^2$  based on the values of  $y$ .

The standard error of mean of  $y$  is found by the statement `sqrt(var(y)/n)` in R.

Now, I will try to take here the first case when the population variance  $\sigma^2$  is known. Well, anyway as far as you are considering the estimation of population mean by the sample mean, that is sample arithmetic mean it will not make any difference, right. When you go for the computation of variance and standard error as well as test of hypothesis then this will make difference.

So, if you want to estimate the population mean by sample mean that means, we are going to find out only the arithmetic mean of the values. So, this can be computed by using the command `mean(y)`,  $y$  is my data vector that is that will indicate the sample values.

And in case of simple random sampling with replacement, the computation of variance of  $\bar{y}$  is pretty straight forward. Why? Because, if you remember the estimate of variance of  $\bar{y}$  under WR this was  $s^2/n$ . And in R this the command var this computes the quantity  $s^2$  ok.

So, in case if you want to find out the standard error, then you simply have to take the positive square root of this command. So, the variance of  $\bar{y}$  that is sample mean can be estimated using the command variance of y divided by n that is the sample size, and its standard error can be computed by taking its positive square root.

(Refer Slide Time: 03:52)

```
Using R software: Estimation of population mean and
population variance in SRSWR
Case 1: When population variance ( $\sigma^2$ ) is known

For
ywr
[1] 10 17 18 1 18

> mean(ywr)
[1] 12.8

> var(ywr)/n
[1] 10.94

> sqrt(var(ywr)/n)
[1] 3.307567

R Console
> n
[1] 5
> ywr
[1] 10 17 18 1 18
> mean(ywr)
[1] 12.8
> var(ywr)/n
[1] 10.94
> sqrt(var(ywr)/n)
[1] 3.307567
> |
```

So, now, we have this sample here and those sample values have been stored in a variable ywr. And, if I want to find out the mean I simply have to find out the sample mean that is mean of ywr and this value comes out to be here 12.8. And, if you want to find out the standard error or the variance of  $\bar{y}$  that is pretty straight forward. To find out the variance of  $\bar{y}$  under SRSWR you simply have to find out variance of ywr divided by n.

So, you can see here this value here is 10.94. And, if you want to find out standard error you simply have to take the square root of this value. So, this can be obtained by the command sqrt of the same command which is used here. So, this comes out to be 3.30 approximately and this is here the screenshot. So, as I said earlier when I try to repeat on

the R console here, then these results will change because I will be hopefully I will be getting a different sample.

(Refer Slide Time: 04:57)

**Using R software: Confidence interval of  $\bar{y}$  in SRSWR**  
**Case 1: When population variance ( $\sigma^2$ ) is known**

In R the value of the quantile of Normal distribution is found by the function `qnorm()` .  $Z_{\alpha/2}$

The quantile for significance level  $\alpha = 2.5\%$  is obtained by the function `qnorm(.975)` .

The quantile for significance level  $\alpha = 5\%$  is obtained by `qnorm(.95)` .

```
> qnorm(.975)
[1] 1.959964

> qnorm(.95)
[1] 1.644854
```

R Console

```
> qnorm(.975)
[1] 1.959964
> qnorm(.95)
[1] 1.644854
> |
```

And when you try to construct the confidence interval, as I discussed in the earlier lecture you will need here the value  $Z_{\alpha/2}$ . So, I will not discuss it here, because I already have discussed it earlier. So, in order to find out the value of  $Z_{\alpha/2}$ , we simply have to essentially find the percentile those percentiles can be obtained using the command `qnorm`. So, for  $\alpha$  is equal to 2.5% or 0.025 the `qnorm` will be given as  $(1 - \alpha)$ .

So, this will be `qnorm 0.975` and if you try to execute it, this will give you the value 1.96 which is approximate value, and the same command can be obtained from the R on the R console also. And if you want to put this  $\alpha$  to be 5%, then the value inside the parenthesis in the `qnorm` command will be  $(1 - \alpha)$ . So, that will be 1 minus 0.05 which is 0.95 and this value comes out to be close to 1.64.

(Refer Slide Time: 06:03)

**Using R software: Confidence interval of  $\bar{Y}$  in SRSWR**  
**Case 1: When population variance ( $\sigma^2$ ) is known**

The commands to find 95% confidence interval for data  $y$  in R are the following.

```
> errorzwr <- qnorm(0.975) * sqrt(var(y)/n)
> leftCIzwr <- mean(y) - errorzwr
> rightCIzwr <- mean(y) + errorzwr
```

The values of `leftCIzwr` and `rightCIzwr` are the left hand and right hand limits of the confidence interval respectively.

$$\left[ \bar{y} - Z_{\alpha/2} \sqrt{\text{Var}_{wr}(\bar{y})} \leq \bar{Y} \leq \bar{y} + Z_{\alpha/2} \sqrt{\text{Var}_{wr}(\bar{y})} \right]$$

*Handwritten notes on the slide:*  
-  $Z_{\alpha/2}$  is written above the `qnorm(0.975)` command.  
- `sqrt(var(y)/n)` is circled in green.  
- `leftCIzwr` and `rightCIzwr` are circled in green.  
- `mean(y)` is circled in green and labeled "mean(y)".  
- The formula below has  $\bar{y}$  circled in green and labeled "mean(y)", and the right side  $\bar{y} + Z_{\alpha/2} \sqrt{\text{Var}_{wr}(\bar{y})}$  circled in green and labeled "upper".

Now, I try to compute the confidence interval and as we had discussed, the we have to compute the lower limit of the confidence interval, please have a look in the bottom of the slide and the upper confidence limit. So, for this as I discussed it earlier, these things can be programmed very easily  $\bar{y}$  can be obtained by mean of  $y$ . And, this variance of  $\bar{y}$  is obtained already and then you have to use the square root.

So, this thing can be written over here. You can see here, I am writing here mean of  $y$  minus `errorzwr`, where `errorzwr` is  $Z_{\alpha/2}$  which is here obtained as obtained by `qnorm(0.975)` for  $\alpha$  is equal to 0.05. And the standard error is obtained by the command `sqrt` inside the parenthesis variance of  $y$  divided by  $n$ , right.

So, this will give you the second component of the lower and upper confident limit. And, if you try to just say mean of  $y$  minus this component `errorzwr`, this will give you the left confidence limit of the confidence interval. And, similarly if you try to write down mean of  $y$  plus `errorzwr` this will give you the upper confidence limit, which I have denoted by `rightCIzwr`; that means, right confidence interval of  $Z$ ; that means, under the normality under  $wr$ .

(Refer Slide Time: 07:50)

```
Using R software: Confidence interval of  $\bar{Y}$  in SRSWR
Case 2: When population variance ( $\sigma^2$ ) is known
Example:
> n <- 5      N=?
> errorzwr <- qnorm(0.975) * sqrt(var(y)/n)
> leftCIzwr <- mean(ywr) - errorzwr
> rightCIzwr <- mean(ywr) + errorzwr
> leftCIzwr
[1] 7.742465
> rightCIzwr
[1] 17.85754

95% CI (7.74, 17.86)

Confidence interval is [7.742465, 17.85754]
```

So, now we try to execute these things over the R console. So, as we have seen that  $n$  is equal to here 5 is  $n$  which is the sample size, and here practically you do not need here actually  $N$ . So, that is the reason that I am not giving it here. Although, if you wish you can give it you know it, but in order to compute the confidence interval you need only the variance and variance depends only on a  $n$  right.

So, you can see here I am writing here `qnorm` which is the  $Z$  value. And, the standard error of  $\bar{y}$  and then I try to compute the left and right, limits of the confidence interval using the same command and you can see here, this is my here outcome for the given data set.

So, my 95% confidence interval for this  $\bar{Y}$  comes out to be close to 7.74 to 17.86 ok. So, this is what I wanted to find and the interpretation exactly goes in the same way, as I discussed in the earlier lecture.

(Refer Slide Time: 09:03)

**Using R software: Confidence interval of  $\bar{Y}$  in SRSWR**  
**Case 2: When population variance ( $\sigma^2$ ) is unknown**

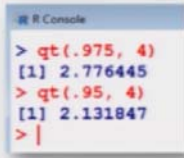
In R, the value of the quantile of  $t$  - distribution with  $n$  degrees of freedom is found by the function `qt( , df=n)`.

The quantile for significance level  $\alpha = 2.5\%$  and  $n = 4$  degrees of freedom is obtained by `qt(.975, 4)`.

The quantile for significance level  $\alpha = 5\%$  and  $n = 4$  degrees of freedom is obtained by `qt(.95, 4)`.

```
> qt(.975, 4)
[1] 2.776445

> qt(.95, 4)
[1] 2.131847
```



The screenshot shows an R console window with the following text: `> qt(.975, 4)` followed by `[1] 2.776445` on the next line, and `> qt(.95, 4)` followed by `[1] 2.131847` on the next line. A vertical bar is visible at the end of the second line of output.

Now, I try to take the second case where the variance is assumed to be unknown, so,  $\sigma^2$  is unknown. So, in this case we had used  $t$  distribution. So, for the  $t$  distribution I had discussed in the earlier lecture, that the percentile or the quantiles of a  $t$  distribution can be obtained using the command `qt`. And, inside the parenthesis you have to give the value of  $(1 - \alpha)$  and the degrees of freedom.

So, degrees of freedom remember that in the confidence interval, it is  $n$  minus 1 and here we have to give the exact degrees of freedom. So, either you write here in terms of sample size as  $n$  minus 1 or you give the directly the value of  $n$  minus 1. So, if you try to take  $\alpha$  is equal to 5%, then at then  $\alpha/2$  will be 2.5% and the 97 point 5<sup>th</sup> percentile of  $t$  distribution with 4 degrees of freedom can be obtained using this command.

And, similarly the at  $\alpha$  equal to 5%, this percentile the 95<sup>th</sup> percentile can be using can be obtained using this command. And these are the values, they are the same values which we have obtained in the earlier lecture.



(Refer Slide Time: 10:29)

**Using R software: Confidence interval of  $\bar{Y}$  in SRSWR**  
**Case 2: When population variance ( $\sigma^2$ ) is unknown**

The commands to find the 95% confidence interval for data  $y$  in R are the following.

```
> errortwr <- qt(.975, n-1) * sqrt(var(y)/n)
> leftCItwr <- mean(y) - errortwr
> rightCItwr <- mean(y) + errortwr
```

The values of `leftCItwr` and `rightCItwr` are the left hand and right hand limits of the confidence interval respectively.

$$\left[ \bar{y} - t_{\frac{\alpha}{2}} \sqrt{\frac{\text{Var}_{wr}(\bar{y})}{n}} \leq \bar{Y} \leq \bar{y} + t_{\frac{\alpha}{2}} \sqrt{\frac{\text{Var}_{wr}(\bar{y})}{n}} \right]$$

Now, exactly on the same lines, you can see here this is here the confidence interval. This is your here the lower limit, and this is here the upper limit and in order to compute it, so, this value can be obtained by the command `qt`, this value  $\bar{y}$  can be obtained using the mean of  $y$ . And this quantity square root of variance of  $\bar{y}$ , this can be obtained by  $\sqrt{\text{Var}(\bar{y})}/n$ .

So, these things, so, this part which is here like here this, this I am trying to compute here in this function `errortwr` which is here `qt` that is the value of  $t$ . And, then I am trying to find out the variance and then taking its square root, and based on that then I try to compute the left confidence interval like this.

Here in this command, which is here mean of  $y$  minus this `errortwr`, and here in the second one here I am trying to compute the upper limit of the confidence interval, that is pretty straight forward it is not difficult.



(Refer Slide Time: 11:54)

```
Using R software: Confidence interval of  $\bar{Y}$  in SRSWR
Case 2: When population variance ( $\sigma^2$ ) is unknown
Example:
Population:
> x <- 1:20
> x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
18 19 20
Sample:
> ywr <- sample(x, size=5, replace = TRUE)
> ywr
[1] 10 17 18 1 18
The degrees of freedom  $n = 5-1=4$ 
```

And, then if I try to put it on the R software. So, you may recall that our population was 1 to 20 and we had obtained and we are trying to obtain here a sample using SRSWR. So, for that the command is sample population x size 5 and then now replace will become here TRUE because, we want it by with replacement.

So, the sample is obtained here. So, now, the sample is here like this and for this sample I am denoting here n equal to 5 and then I am trying to compute it. So, I simplify it to give it here the value here ywr in place of y and I try to compute the error twr part and which comes out to be 9.18.

(Refer Slide Time: 12:41)

```
Using R software: Confidence interval of  $\bar{Y}$  in SRSWR
Case 2: When population variance ( $\sigma^2$ ) is unknown
Example:
> n <- 5
> errortwr <- qt(.975, n-1)* sqrt(var(ywr)/n)
> errortwr
[1] 9.183278
> leftCItwr <- mean(ywr) - errortwr
> leftCItwr
[1] 3.616722
> rightCItwr <- mean(ywr) + errortwr
> rightCItwr
[1] 21.98328
The 95% confidence interval is [3.616722, 21.98328]
```

95% CI (3.61, 21.98)

Then, I try to use the lower limit of the confidence interval, and upper limit of the confidence interval, by the usual commands. For this sample this value comes out to be 3.67 and the right limit of the confidence interval comes out to be 21.98 approximately. So, I can see here that the 95% confidence interval, when  $\sigma^2$  is unknown to us comes out to be say close to 3.61 to 21.98, right.

(Refer Slide Time: 13:22)

```
Using R software: Confidence interval of  $\bar{Y}$  in SRSWR
Case 2: When population variance ( $\sigma^2$ ) is unknown
R R Console
> n
[1] 5
> errortwr <- qt(.975, n-1)* sqrt(var(ywr)/n)
> errortwr
[1] 9.183278
> leftCItwr <- mean(ywr) - errortwr
> leftCItwr
[1] 3.616722
> rightCItwr <- mean(ywr) + errortwr
> rightCItwr
[1] 21.98328
> |
```

So, this is my confidence interval. In this case the interpretation goes exactly on the same way as we did earlier when I was trying to prepare this lecture. Now, you will be getting a different sample.

(Refer Slide Time: 13:31)

**Estimation of Population Total**

Sometimes, it is also of interest to estimate the population total, e.g. total household income, total expenditures etc.

Let  $Y_T$  denotes the population total defined as

$$Y_T = \sum_{i=1}^N Y_i = N\bar{Y}$$

$Y_T$  can be estimated by

$$\hat{Y}_T = N\hat{\bar{Y}} = N\bar{y}$$

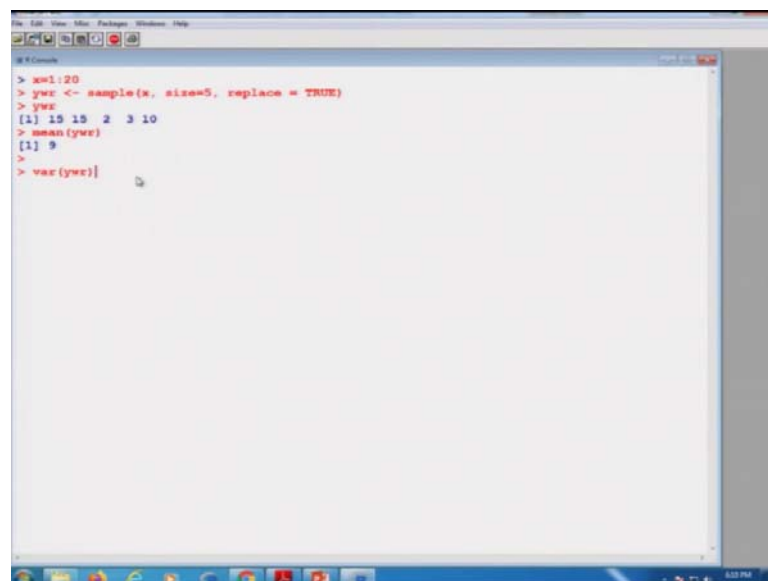
*Handwritten notes:*

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$
$$\Rightarrow \sum_{i=1}^N Y_i = N\bar{Y}$$
$$\Rightarrow Y_T = N\bar{Y}$$
$$\hat{Y}_T = N\bar{y}$$

13

So, these values may not match, but the expressions and formula will match ok. Before I go further let me try to show you these things on the R console itself. So, if you try to see here first I try to create here a sample here.

(Refer Slide Time: 13:47)



```
> set.seed(20)
> ywe <- sample(w, size=5, replace = TRUE)
> ywe
[1] 15 15 2 3 10
> mean(ywe)
[1] 9
> var(ywe)
```

On the R console I try to define my population 1 to 20. And, then I try to take a sample over here with replacement and you can see here, this is my sample. So, you can see here this time 15 and 15 are repeated two times. Now, if you try to find here its mean, so, will be mean of ywr so, this is straight forward this comes out to be 9. And, if you want to find out the standard error of the sample mean, which is variance of ywr divided by here n, right.

(Refer Slide Time: 14:22)

```

> x=1:20
> ywr <- sample(x, size=5, replace = TRUE)
> ywr
[1] 15 15 2 3 10
> mean(ywr)
[1] 9
> n=5
> var(ywr)/n
[1] 7.9
> sqrt(var(ywr)/n)
[1] 2.810694
> errorzwr <- qnorm(0.975) * sqrt(var(y)/n)
> errorzwr
[1] NA
> errorzwr <- qnorm(0.975) * sqrt(var(ywr)/n)
> errorzwr
[1] 5.508859
> lci=mean(ywr)-errorzwr
> lci
[1] 3.491141
> uci=mean(ywr)+errorzwr
> uci
[1] 14.50886
> errortwr <- qt(.975, n-1)* sqrt(var(ywr)/n)
> errortwr
[1] 7.803737
> lci=mean(ywr)-errortwr
> lci
[1] 1.196263
> uci=mean(ywr)+errortwr
> uci
[1] 16.80374
>

```

So, let me try to give it here first n, n is equal to here 5. So, I try to find out the variance of ywr divided by n, which will give me the variance of the sample mean under SRSWR which is here 7.9. And, if I want to find out the standard error then I simply have to take, the square root of this quantity. This is positive so, it will be always be positive. So, the standard error is 2.81.

And, now if you assume that the  $\sigma^2$  is known, then in that case if you want to compute the confidence interval, then the error part is computed here by this thing. So, you can see here error this comes out to be here NA, why this is here NA? Because, you have not replaced here y by ywr, so, you can see here now this comes out to be the correct value.

And, now if you try to find out here lower confidence limit, let me call it lci that will be here, mean of ywr minus errorzwr right. So, this is here lci, this is the lower confidence limit when  $\sigma^2$  is known. And, if you try to find out the upper confidence limit, let me

denote it by here  $\bar{y}$ . So, that will be mean of the observation plus error  $\bar{y}$ . So, this comes out to be  $\bar{y}$  which is 14.50. So, the 95% confidence interval turns out to be 3.49 to 14.50.

And, now similarly if you want to find out the confidence interval using the t distribution here when  $\sigma^2$  is unknown, so the error  $\bar{y}$  is obtained by this command. So, you can see its value is like this and if you want to find out the lower limit. So, let us call it  $\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$  lower confidence interval with t distribution, this is mean of  $\bar{y}$  minus error  $\bar{y}$ .

So, this comes out to be here  $\bar{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$  and if you try to find out the upper confidence limit then, let me call it  $\bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ . So, this  $\bar{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$  comes out to be here 16.80. So, once again the same conclusion that we discussed in the earlier lecture holds true here, that a lower confidence interval when  $\sigma^2$  is known is from 3.49 to 14.5 and when  $\sigma^2$  is unknown and it is being estimated from the sample.

So, it is 1.19 to 16.8 or so. So, that is obvious that for the same sample, when you are trying to replace a known value by an estimated value then, definitely the sampling variation will increase in this in these quantities right. So, the confidence interval which are constructed when  $\sigma^2$  is unknown, usually they will be they will have a their confidence width the width of the confidence interval will be higher than the width of the confidence interval when  $\sigma^2$  is known to us.

So, now, after this I try to come back on my slides and I try to give you the last topic of this, simple random sampling which is pretty straight forward ok. So, now, in this small topic I am simply going to give you an idea of estimation of population total that is pretty straight forward actually. Sometimes in practice the people are not interested in the average value, they are not interested in estimating the population mean, but they want to find out the total.

For example, I want to know what is the total of the population.

For example, if I say  $Y_i$  is my the  $Y_i$  is the height of the  $i^{\text{th}}$  person, then I have I simply want to find out the sum of  $Y_i$ 's over the entire population, yeah in some practical application this is of interest. So, here I would like to show you that there is a one to one

correspondence between the estimation of population mean and estimation of the population total.

So, in order to estimate the population total you do not have to do anything new. But, using the result what you have obtained in case of population total, they can be directly used here with a very minor modification, how? Let me try to show you in this slide ok. So, sometime we are interested in estimating the population total.

For example, sometime we want to estimate the total household income, total expenditure etcetera. But, if you try to see what is the population total, the population total is defined as the sum of all the sampling units in the population that means, sum of all the units in the sampling frame corresponding to the variable under steady.

So, this we are going to denote by  $Y_T = \sum_{i=1}^N Y_i = N\bar{Y}$ , right.

So, this implies that  $Y_T$  is nothing but your  $N$  times  $\bar{Y}$ . So, using the rules of statistics there is a one to one correspondence between the two parameters which is  $Y_T$  and say  $\bar{Y}$ . So, if I know the estimator of  $\bar{Y}$ , then I can obtain the estimator of  $Y_T$  directly using the one to one transformation.

So, this becomes here because we know that we had estimated  $\bar{Y}$  by  $\bar{y}$ . So, the estimator of total will become simply  $N\bar{y}$ . So, this is what I am trying to show you here, that is why total can be estimated by  $\hat{Y}_T$  which is equal to  $N$  times  $\bar{y}$ ; that means, sample mean into the population size.

(Refer Slide Time: 21:33)

**Estimation of Population Total**

Then  $E(\hat{Y}_T) = NE(\bar{y}) = N\bar{Y} = Y_T$ . *SRSWOR SRSWR  $Y_T$  remains an unbiased estimator of  $Y_T$*

Variance of  $\hat{Y}_T$  is

*$\hat{Y}_T = N\bar{y}$   
 $Var(\hat{Y}_T) = N^2 Var(\bar{y})$*

$$Var(\hat{Y}_T) = N^2 Var(\bar{y})$$

$$= \begin{cases} N^2 \left( \frac{N-n}{Nn} \right) S^2 = \frac{N(N-n)}{n} S^2 & \text{for SRSWOR} \\ N^2 \left( \frac{N-1}{Nn} \right) S^2 = \frac{N(N-1)}{n} S^2 & \text{for SRSWR.} \end{cases}$$

Estimate of variance of  $\hat{Y}_T$  is

$$\widehat{Var}(\hat{Y}_T) = N^2 \widehat{Var}(\bar{y})$$

$$= \begin{cases} \frac{N(N-n)}{n} s^2 & \text{for SRSWOR} \\ \frac{N^2}{n} s^2 & \text{for SRSWR.} \end{cases}$$

And once you do this thing then there is no problem in establishing the unbiasedness, because that will remain the same. Because, if you try to see expected value of  $\hat{Y}_T$  will become simply N times expected value of  $\bar{y}$ , which is N times  $\bar{Y}$  which is same as  $Y_T$ .

So, there is no issue even under SRSWR and SRSWOR, this  $\hat{Y}_T$  remains an unbiased estimator of  $Y_T$  or sorry  $\bar{Y}_T$  that is the population total. Now, in case if you want to find out the variance or the estimate of variance that is very simple, because you already have established that  $\hat{Y}_T$  is equal to N times  $\bar{y}$ . So, whatever is your variance of here  $\bar{y}$  that is same as  $N^2$  times variance of  $\bar{y}$ .

So, this relationship I am going to use here and so, what you have to do? You simply have to multiply the variance of  $\bar{y}$  under SRSWOR and under SRSWR by the quantity  $N^2$ . And you will get here these two terms here, which are the variances of  $\hat{Y}_T$  under the SRSWOR as well as under SRSWR that is straightforward. And, simply if you want to estimate these variances because, you can see here we have here  $S^2$  that is unknown in real data set in practice.

So, I can find out an unbiased estimator of  $S^2$  as we have done earlier and actually we already have found the estimates of the variance. So, the estimate of the  $Y_T$  mean



the variance of the  $\hat{Y}_T$  will be same as the variance of  $\hat{\bar{y}}$  under SRSWR and WOR, the only difference will be you have to multiply it by here the quantity  $N^2$  right. So, if you try to and see here that these are the expressions what we have obtained. So, that is pretty straight forward right.

So, now and similarly if you want to compute the confidence interval means, you can do it very easily and that is straight forward. So, now, I stop here and I will also try to complete the chapter of simple random sampling, but this was simple random sampling for quantitative variable, I have not told you.

But, we always assumed that whatever  $y_1, y_2, \dots, y_n$  we have taken they are some quantitative variables, they are not like qualitative variables like yes or no, low medium or high, but they are taking some numerical values. So, this is all about the simple random sampling for quantitative variables.

What to do with qualitative variables that will be our next topic that is not difficult. Because, once you have understood the simple random sampling technique, I will try to establish a one to one correspondence between the quantitative variable concepts and qualitative variable concepts and that will make your life very simple.

So, now this is your turn, I have completed this part, you try to take some example from the books from the assignments, and try to practice them. Whatever I have taught that is directly applicable in any bigger data set, which we try to consider under the purview of data sciences. There cannot be a situation in data sciences, where you would not like to estimate the population mean.

There will not be a situation where you would like to find out the standard errors, there will not be a situation, where you would like to find out the confidence interval for the population mean. So, be confident that these things are going to play a very important role, if you want to become a data scientist. So, you practice, take your ambitions forward and I will see you in the next lecture, till then good bye.