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 – 19
Simple Random Sampling
Estimation of Population Mean

Hello, welcome to the course Essentials of Data Science with R Software-2, where we are trying to understand the basic concept of Sampling Theory and Linear Regression Analysis and in this module we are going to continue with the topics of Simple Random Sampling. So, up to now in the last couple of lectures, we have considered the basic fundamentals of the simple random sampling and we have learned that how we can draw the samples by SRSWOR and SRSWR using the R software.

Now, today I am going to discuss on another issue. My question is why do you do sampling? What is your objective? You may recall that in the beginning, I had explained you that our basic objective is this that we want to know something about the entire population, but that is very difficult; that is not possible.

So, we try to take the help of sampling theory and we try to draw the sample and then we conduct our entire analysis on the basis of sample, we try to apply all the statistical tool only on the sample and whatever are the outcomes they are supposed to be valid for the entire population. For example, if somebody wants to conduct a clinical trial and suppose, there is a new medicine which can control the blood pressure in a much better way than the earlier medicine, so how to conduct the clinical trial?

Some number of persons having the problem of blood pressure will be selected in the sample, they will be given the medicine and their recordings will be noted after some time and then the two values that what was the blood pressure before giving them medicine and after giving them medicine they will be compared.

But once the results are analyzed and it is established that the yes, the medicine is effective, then the medicine is valid for the entire population not only for the selected sample. So, that is our objective.

So, in simple random sampling also when we try to draw the sample then our objective is to estimate something for the population. When I say something for the population from the statistical point of view there can be several statistical quantities; they can be population mean, population variance, skewness, kurtosis and all sorts of parameters.

So, our basic objective is this; how to estimate a population value or a population parameter on the basis of sample which is drawn by simple random sampling technique. Well, now when I come to the aspect of sampling theory or in general, people are more interested in finding out the population mean and population variance; that means, they want to see that where the values are scattered or where the values are more concentrated, how the values are scattered, how the values are spreaded?

For example, as a student, you always try to inquire from your teacher, what is the class average. Based on the class average you always try to determine that whatever are the marks obtained on an average inside the class, based on that you try to decide whether what type of grades you are going to get. So, you are interested in the average value.

And similarly once you try to see the average value, you always try to see how is the spread of the values. For example, if some person has got 1 mark out of 100 and say another person has got say this say 100 marks out of 100. So, then the average is going to be close to 50, right; 100 plus 1 that is 50.5. And there is another possibility that all the students have got the marks between say 45 and 55. So, when you try to take the average of those marks, that will again be close to 50. So, in both the samples, the average value is coming out to be close to 50, then what is the difference?

The difference can be ascended, can be known, can be found using the concept of variability and you would always like that the observation should have as minimum as possible variability, whatever tools you are going to use they should give you a result which has got the minimum variability and that we also have discussed in the past right.

So, that is the precise reason that in sampling theory usually you will see people are more concentrated or more interested in estimating the population mean and population variability.

Now, the next question is this, you know from your statistics knowledge that there are different measures to find out the central tendency of the data; that can be arithmetic mean, that can be geometric mean, that can be harmonic mean, that can be median, that can be mode, and similarly for the variability; there is variance, there is absolute deviation, there are quartile deviation and etc. There are many-many things, but now the question is which of the measure you would like to choose?

Well, now at this moment I would try to address two things; first I will try to concentrate on the mathematical easiness means; once you are trying to do the algebra, then algebra should be decent and manageable and second thing is this it becomes very difficult. So, in the first case when we try to concentrate on algebra and statistical properties of the estimators then arithmetic mean turns out to be a reasonable choice.

Means, you take the observation and try to find out their arithmetic mean and use it to estimate the population mean, because from the statistical inference also you have seen that the sample mean has got nice statistical properties. And similarly, when you want to study the variability in the data one of the most convenient most popular and statistically efficient measure is variance.

So, usually you will see that people are trying to use the sample mean or arithmetic mean of the values to estimate the population mean and they are trying to use the sample variance; that means the variance of the sample to measure the population variance. So, that is the reason that why people are trying to consider these two aspects using these two measures, but I am not saying at all that these are the only possible measures.

There are other things also and we will see at a later stage that if you have some complicated types of thing then we have a bootstrap methodology which can help you in computing the different types of statistical function based on only computation.

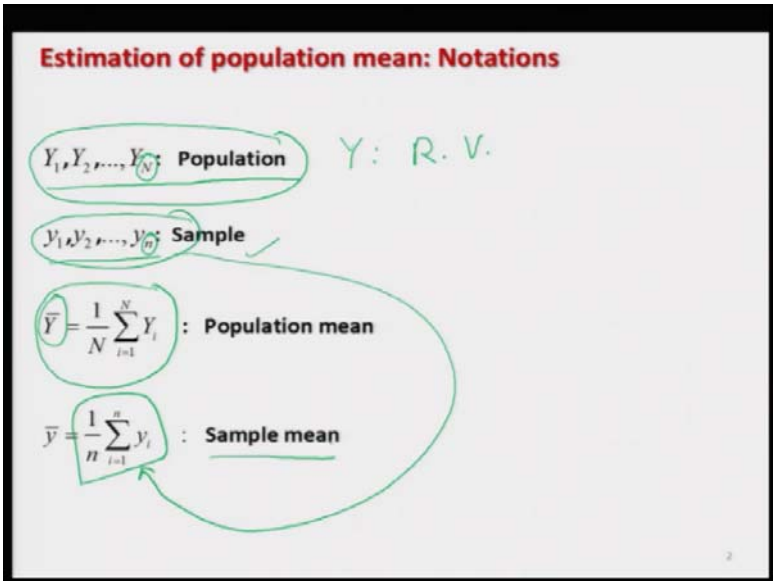
So, this is the basic idea and this is the basic philosophy behind the computation of sample mean and sample variance or treating sample mean and sample variance as an estimator of population mean and population variance, respectively.

Nobody is stopping you in using median, mode or say anything else to estimate the population mean, but these are the reasons in my opinion which motivate us to consider

the sample mean as a estimator for population mean and a variant of sample variance as a estimator for the population variance.

So, with this objective let us begin our this lecture. In this lecture I am going to consider the estimation of population mean and in the next lecture I will try to address the estimation of population variance.

(Refer Slide Time: 09:37)



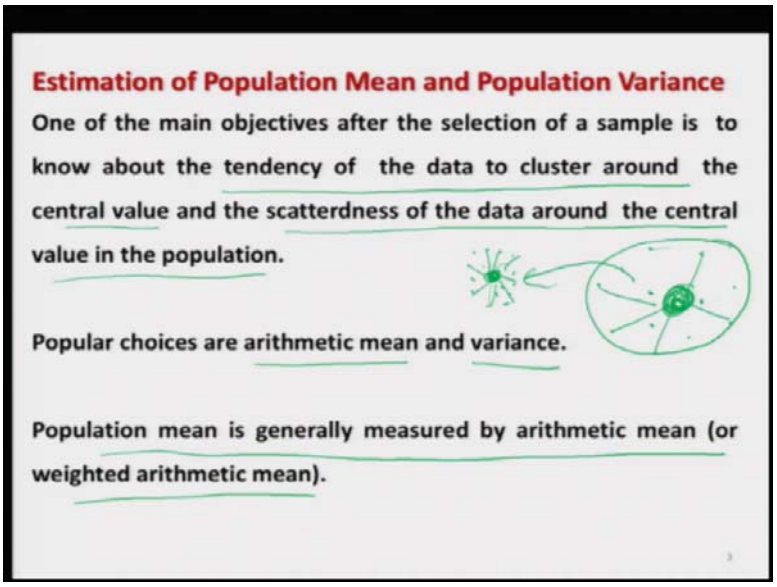
So, let us begin our lecture now ok. So, first let me briefly explain you our symbols and notation that we are going to use. So, as we had discussed earlier also, capital Y_1, capital Y_2,…, capital Y_N that will be denoting the sampling units in the population and; obviously, Y is my here random variable.

And then small y_1, small y_2, small y_2, they are going to indicate the sampling units in the sample and by looking at this capital N and this small n you can automatically know that what is my population size and what is the sample size.

Now, if I try to define the population mean one option is this; I can use the concept of arithmetic mean and using the concept of arithmetic mean the population mean of this Y_1, Y_2,…, Y_N in the population can be defined by here. 1 upon capital N, summation i goes from 1 to capital N Y i and this is going to be denoted by $\bar{Y}$, that will be our standard notation through the entire lecture on sampling theory that $\bar{Y}$ will be denoting the population mean.

And whatever sample you have obtained here, try to find out the arithmetic mean of these values and this is going to be denoted by here $\frac{1}{n}\sum_{i=1}^{n} y_i$ and this is my here, a $\bar{y}$ this is denoting the sample mean and the same notation will continue throughout the lectures on sampling theory, right ok.

(Refer Slide Time: 11:25)



So, as we have discussed that one of the main objective after the selection of the random sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value, but that we want to know in the population.

What does this mean? For example, if I say if my data is here like this then we can see here, this is the point here where most of the values are concentrated and this distance from the central value to this individual point that is giving us an idea about the spread or the scatterdness of the data.

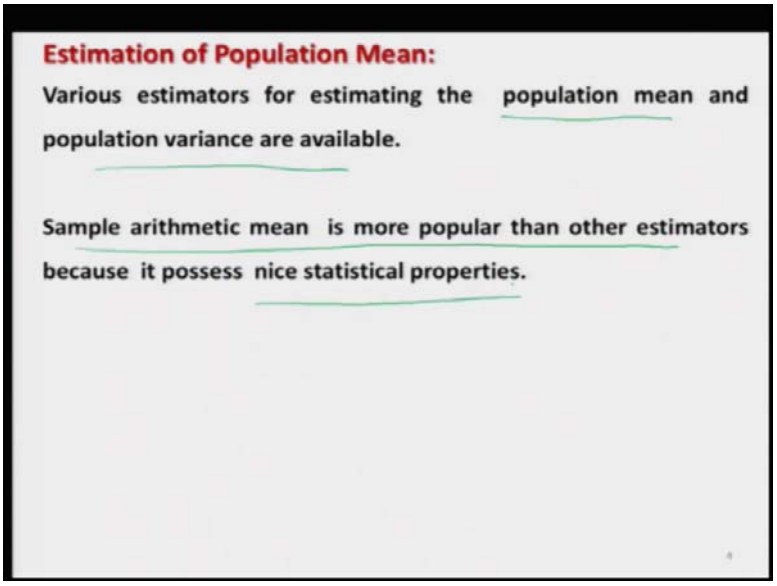
So, if I try to take here another data set here on the same scale so, you can see here this has central tendency somewhere here, the data is clustered somewhere here around this point, but this variation, the difference between the central value and all the individual point this is much larger in comparison to this data set. So, in this case the variation is

going to be higher. So, we want to know the point where the data is concentrated and what is the scatterdness of data around the central value.

So, the popular choice to know these things for central tendency; the popular choice is arithmetic mean and for the scatterdness, for the variability the popular choice is variance. So, that is what we try to denote in the population also. So, the population mean will be denoted by arithmetic mean and the population variance will be denoted by the concept of variance, right.

Yeah sometime, people also try to consider the weighted arithmetic mean depending on the choice. For example, when we do the stratified sampling there you will see that we will try to consider the weighted arithmetic mean instead of simple arithmetic mean.

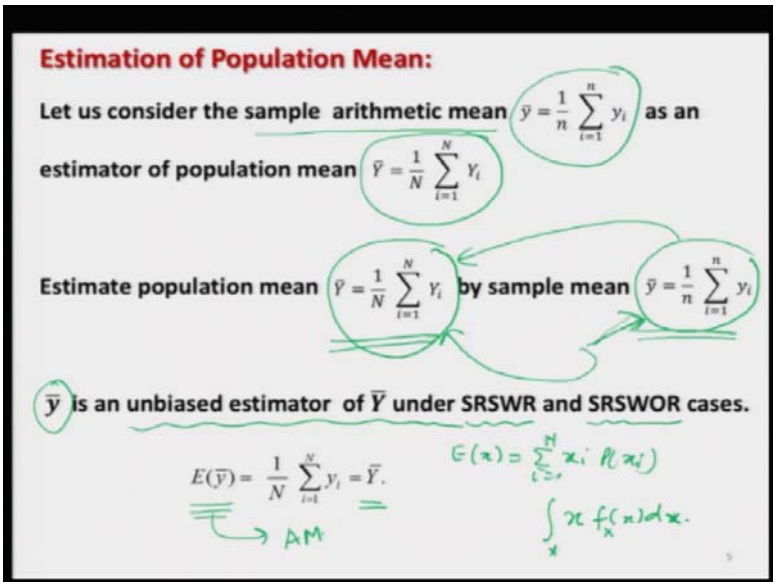
(Refer Slide Time: 13:34)



So, as I told you that various estimator for estimating the population mean or the population variance they are available. And can we try to consider the estimation of population mean?

Then sample arithmetic mean is one of the popular choice in comparison to other estimators and the reasons for this are manifold as I told you and one of the important reason you can see that whenever you are trying to define the sample arithmetic mean, it is easy to handle it mathematically and it will also have nice statistical properties in most of the cases, right ok.

(Refer Slide Time: 14:20)



So, now with this objective let us try to move further. So, now, we are going to consider the sample arithmetic mean which is here like this $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as an estimator for the population mean which is defined here as a $\bar{Y}$ as we defined earlier also.

So, now, we will estimate the population mean by sample mean. This means what? We want to know the value of $\bar{Y}$, but we do not know, because this is in the population. Due to various reason it is not possible for us to consider these things.

So, we are trying to find a sample and then we are trying to find the mean of the values in the sample and then we are estimating $\bar{y}$, we are using a $\bar{y}$ to estimate $\bar{Y}$.

In simple words, we are saying whatever is the value of the sample mean based on the values of sample of observation, this is the same value is going to be the mean of the entire population. Well, there can be many question that how it is possible and whether it will give us a good value or bad value. So, these those aspect we will now try to address one by one.

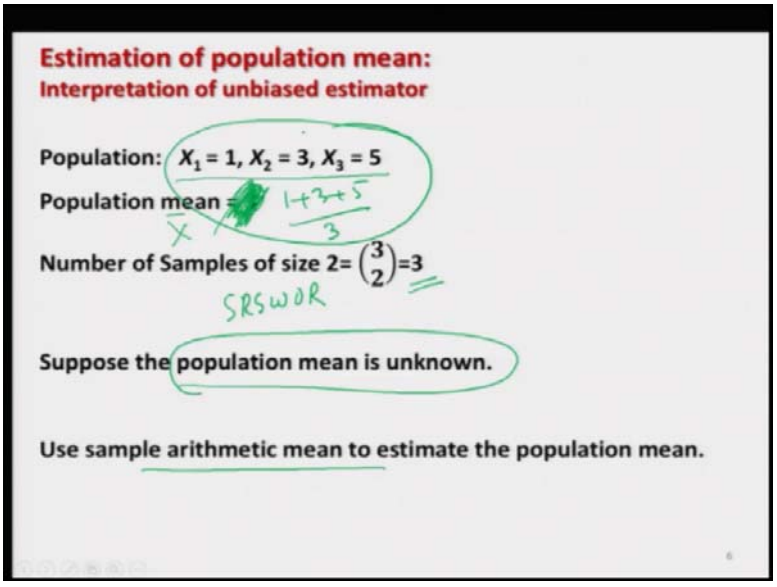
One important property of this sample mean is that, that sample mean is an unbiased estimator of the population mean under both type of sampling scheme that is SRSWR as

well as SRSWOR. What does this mean? From the statistics point of view in the language of statistics we try to denote this property by saying that $E(\bar{y}) = \bar{Y}$, right.

What is this expectation? Expectation those who are from statistics background they know it, for those who are not, I will request them to please go through with some statistic books and try to see, but you can see here these are defined as say some possible values of xi into probability of say here x_i; i goes from 1 to here N if the population is discrete and capital P is the probability mass function and if population is continuous, then it is a sort of $\int f(x)xdx$ right and over the values of here x you have to integrate at integrate it.

So, essentially this expected value of a function, this indicates the arithmetic mean. Why? Because in this case what we are going to say that we have different sample and the probability of drawing any of the sample is the same.

(Refer Slide Time: 17:28)

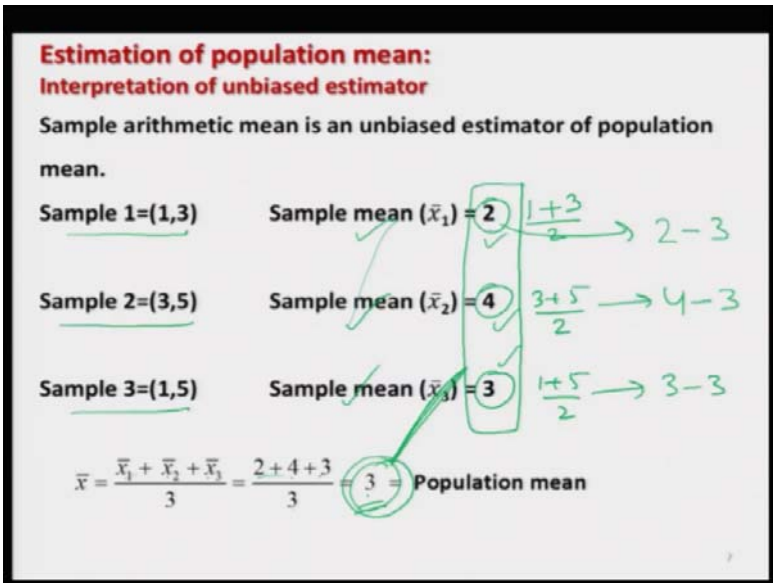


So, now, I do try to do one thing first I try to take a very simple example to explain you that what is the concept of unbiasedness and how it is useful in sampling theory or why we want that the estimator has to be an unbiased estimator of the population value. Suppose, I take a population of size 3 and I have a three values X_1, X_2, X_3 which takes value 1, 3 and 5.

So, the population mean in this case will become 1 plus 3 plus 5 divided by 3 which is equal to here 3. And number of possible samples which can be drawn by SRSWOR here are $\binom{3}{2}$ which is equal to here 3 and for a while, you assume that the this population mean is unknown to us. So, we do not know here this here 3, we do not know, right.

Now, what are we trying to do? We do not know this population mean $\overline{X}$, but we want to know its value. So, what I do? I try to take here as sample and try to compute the arithmetic mean of the sample values, ok.

(Refer Slide Time: 18:36)



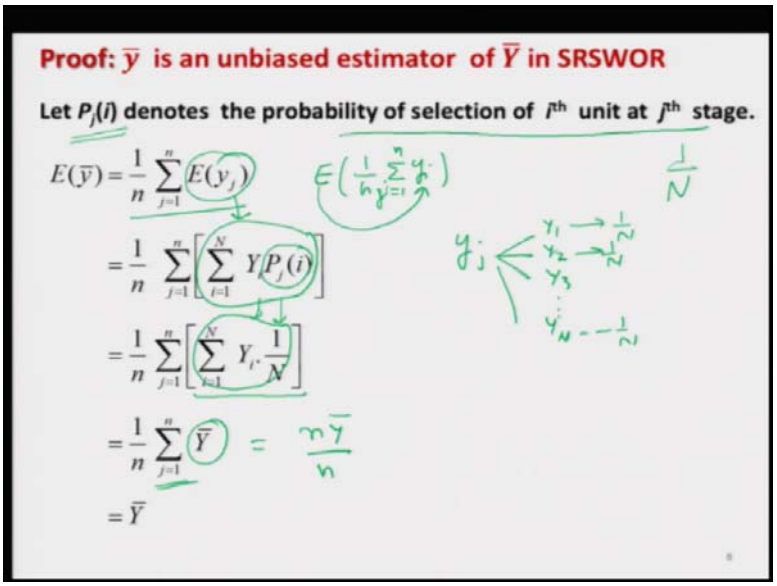
So, here in this case I have three possibilities to draw sample; sample number 1 can be the 1 and 3 unit, sample number 2 can be 3 and 5 the unit and sample number 3 can be 1 and 5 unit right. So, if you try to take the value of here x_1, x_2, x_3. So, you can see here X_1 is 1, X_2 is 3 and X_3 is 5.

So, now; if I try to consider the sample mean of the first sample that is 1 plus 3 divided by 2, this is going to be here 2, simply the sample mean of second sample this is 3 plus 5 divided by 2 which is equal to here 4 and third sample mean is 1 plus 5 divided by 2 which is equal to here 3.

So, now, you can see here, I have got here three sample means 2, 4 and 3 and now, if I try to take the sample mean of these three means. So, you can see here this, I am taking $\frac{\overline{X}_1+\overline{X}_2+\overline{X}_3}{3}$ which is equal to (2+4+3)/3 which is here 3 and now you can see here if you observe here the mean, the population mean of 1, 3 and 5 which was here capital $\overline{X}$, which is here same as here 3.

So, this is what we mean by the unbiasedness of an estimator. We say that sample mean is a is an unbiased estimator of population mean; that means, if you try to take all possible sample, try to find all possible sample mean and then try to find out the arithmetic mean of all the sample mean then it will turn out to be the same as the population mean, right.

(Refer Slide Time: 20:24)



So, what is the utility of this thing? Because I will try to well, I can show you from here also. And this utility is actually more useful from the computation point of view particularly, when you are trying to deal with bigger data sets and you know that when you are working in a real life, you will have only one sample, you will not have sample number 1, sample number 2 or sample number 3. You will have only one sample and you have to work with only one sample.

So, in case if you are trying to see, trying to take only one sample mean and you are trying to infer about the entire population, you can see here that the sample means are 2, 4 and 3 and the population mean here is 3. So; that means, the difference between this sample mean and population mean this is not actually much.

In one case this is 2 minus 3, in this case this is sample mean 2 minus population mean, this is 3, in the second case the sample mean is 4 minus the population mean is 3 and in the third case the sample mean is 3 and the population mean is 3. Actually, this is the concept of the bias in the estimator.

So, you can see here if the estimator is unbiased then the chances that the individual values will vary from the unknown population mean large is very-very less. So, usually we expect that the sample mean will be quite close to the population mean. So, before going into the further detail, let me try to give you here a brief sketch of the proof that how we try to prove it.

So, as we have discussed earlier in the lecture, we had computed different types of probability, among them one of them was $P_j(i)$. $P_j(i)$ was denoting the probability of selection of the ith unit at the jth stage and if you remember we had proved in case of SRSWR and WOR that this was 1 upon capital N.

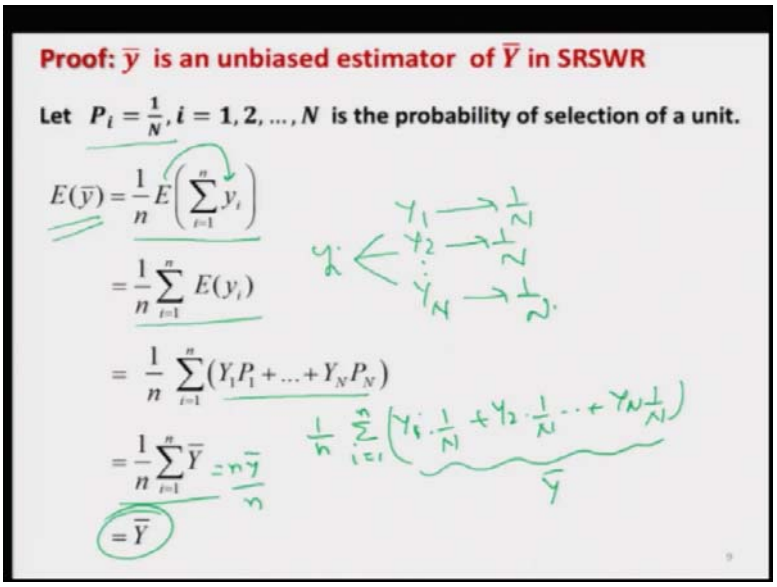
So, now, once I try to find out here $E(\bar{y})$. So, $E(\bar{y})$ can be written as $\frac{1}{n}\sum_{i=1}^{n} E(y_i)$ and then means I can take this expected expectation operator inside and I can write down here $\frac{1}{n}\sum_{j=1}^{n} E(y_j)$, right.

Now, the expected value of any random variable is defined as the sum of the values of random variable multiplied by its corresponding probability. So, now, you can see here y_j is my sample value. This is one of the value which is obtained from the population. So, it can take, it can be 1st value, it can be 2nd value, it can be 3rd value or it can be last Nth value also.

And every unit has got a probability 1/ N, 1/ N and so on that is what we have what we already have discussed. So, now, this E(y_j) will become here $\sum_{i=1}^{N} Y_i$, because this is on the population $P_j(i)$.. So, Y_i. So, in the next step this $P_j(i)$. will be replaced by here 1/N and so, we have here the quantity inside the bracket as $\sum_{i=1}^{N} Y_i \frac{1}{N}$ and you can see here this quantity is nothing, but your population mean.

So, this becomes here $\frac{1}{N}\sum_{i=1}^{N} \bar{Y}$. So, this becomes here n times y bar upon n which is here same as here $\bar{Y}$. So, this proves that in case of SRSWOR the sample mean y bar is an unbiased estimator of the population mean $\bar{Y}$. right.

(Refer Slide Time: 24:35)



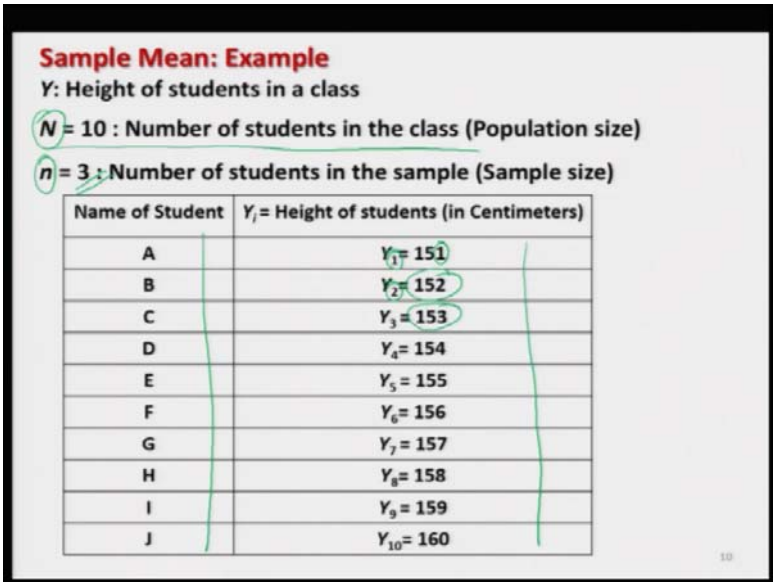
Similarly, if you try to take the case of simple random sampling with the replacement, in that case also the similar thing happen the probability of selection of any unit is 1/N. So, if you try to consider here respected value of here $\bar{y}$. So, that will become expected value of $\frac{1}{n}\sum_{i=1}^{n} y_i$ and so, this can be written here like this and this expectation operator can be taken inside the summation sign.

So, this becomes here $\frac{1}{n}\sum_{i=1}^{n} E(y_i)$. Now, once again this small y_i can take what possible value? This can be capital Y_1 capital Y_2 or this can be capital Y_N and all these units can be selected with the probability, equal probability 1/N. So, this I can write down expected value of Y i here. I say $Y_1 P_1 + Y_2 P_2 + \ldots + Y_N P_N$.

So, now this P_i are equal to 1 upon N. So, this becomes here 1 upon n summation i goes from 1 to here n Y i or say Y_1 X 1/N + Y_2X 1/N +…+ Y_N X1/N and this quantity is nothing but your $\bar{Y}$.

So, this quantity becomes here 1 upon n summation i goes from 1 to n capital Y bar and which is again here n $\bar{y}$ /n which is equal to here $\bar{Y}$. So, once again we can see here that the sample mean is an unbiased estimator of population mean in case of simple random sampling with replacement also.

(Refer Slide Time: 26:29)

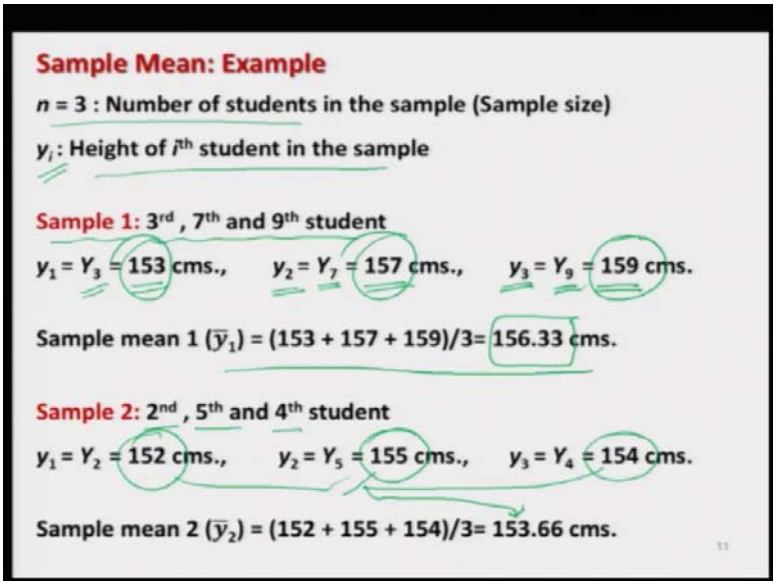


Now, let me try to take the same example which I took in the earlier lectures also and try to give you some more insight that how the things are going to happen in the situation of real data when you have a large number of data under the topic of data science. So, suppose I am just trying to take a small setup here. So, I have here 10 observation on the heights of the students in a class. So, this is my population size.

So, this N now you can imagine that this can be the total number of data set which are available in the framework of our data science and from there you want to draw a sample of size small n here, I am trying to take a small n to be 3.

So, the name of the students are here A to J and then their heights are denoted by Y_1, Y_2, Y_{10} and as I explained you earlier and these values have been taken keeping in mind the convenience. So, this Y_1 is same as here 1 that is 151, second value is 152, third value is 153 and so on, right ok.

(Refer Slide Time: 27:40)



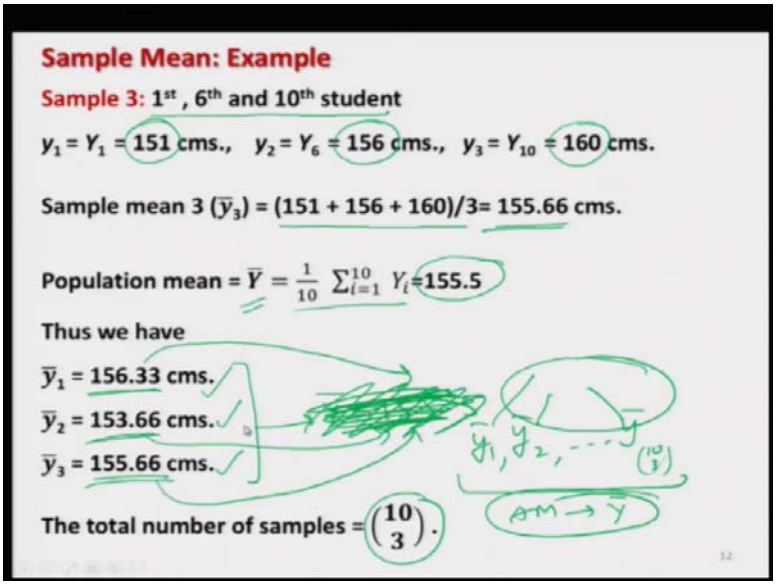
Now, from this suppose I want to draw a sample of size 3 and what is your here the sample value small y_i, this is the height of the ith student in the samples. So, suppose I try to draw here three possible samples.

So, sample 1 suppose, it comes out to be 3rd, 7th and 9th students. So, a small y_1 is the height of the third student which is 153, small y_2 second value in the sample is the seventh value in the population which is 157 centimetres and third value in the sample is the 9th value in the population 159 centimetre and if you try to find out the sample mean of 153, 157 and 159 this comes out to be 156.33 centimetres.

And similarly, if you try to take here another example in which I get 2^{nd}, 5^{th} or the 4^{th} student in my sample. So, the sample values are 152 centimetres, 155 centimetre, 154

centimetre and if you try to find out their arithmetic mean this comes out to be say 153.66 centimetre.

(Refer Slide Time: 28:53)



And similarly, if you try to take one more sample here in which we select 1st, 6th and 10th student, so there values are 151 centimetre, 156 centimetre, 160 centimetre and you find out its sample mean. So, this comes out to be 155.66.

Just for the sake of illustration, although it is not practically possible means I can here compute my population mean. This population mean is denoted by here $\bar{Y}$ which is based on all the 10 values of capital Y_i's and if you try to compute it this comes out to be 155.5.

So, now you can see here you have drawn here three possible sample and the sample mean comes out to be 156.33, 153.66, 155.66 and if you try to compare these values with the $\bar{Y}$ is equal to 150.0 155.5 centimetres, then you can see that there is not much difference between the sample values and this population value. And in case if you try to obtain here all possible samples, the total number of samples that you can obtain here is $\binom{10}{3}$.

So, that is going to be a big number, but if you, but as you have computed here $\overline{y}_1, \overline{y}_2,$ up to here you can compute here $\overline{y}\binom{10}{3}$ and then if you try to take out the arithmetic mean of all such sample mean this will come out to be $\overline{Y}$, but in practice this $\overline{Y}$ is unknown to us and you have only one of the values among these sample means. So, you have to work only with the one sample mean.

So, here I have shown you that if you try to choose an unbiased estimator, then the difference between the estimated value of the sample mean and the true value true unknown value of the population mean the difference is not much.

So, if you try to hide this information that if you do not know then you can expect that whatever values you are estimating here on the basis of random sample they are not far away from the true population value which is completely unknown to us and this is the basic concept of unbiasedness from the computational point of view.

And this is how also we try to test for the for an estimator to be unbiased or not in the simulation also that we try to compute the means of the values of the estimator and then we try to see whether it is close enough to the population mean or not if it is coming out to be close enough, because in the in the setup of a simulation means everything is known to you. And so, this also gives us an idea that why do we choose sample mean.

Well, if you try to choose median or mode, harmonic mean or geometric mean, they will not come out to be an unbiased estimator of the population mean. And one thing what you have to notice here that is very important.

Usually, when you try to do any statistical inference you want to do estimation or you want to do something usually, in statistics we assume a probability distribution that the sample is coming from a normal population from a binomial population or something else, but here in this case we have not assumed any distribution up to now at least.

We have not assumed that my sample y_1, y_2, y_n it is coming from a normal population or from a binomial population or a Poisson population. So, as far as the estimation is concerned, estimation of parameter is concerned we are not going to make any assumption about the distribution. Later on when you come to the test of hypothesis and

confidence interval estimation then we will require a probability distribution, but not now.

So, today I have explained you how to estimate the population mean. So, the rule is very simple one good estimator is this, take a sample and then try to find out its sample mean, but now here you can imagine if your sample is bad, then it will give you a bad value of the population mean and the value of the population mean is not known to you that is completely unknown, that is known only to the God.

So, even you have no choice, you have no option to judge whether your estimated value is good or bad, but the only option that is, which is in your hand is that you try to select a representative sample, a good sample, a sample which reflects all the characteristic which are hidden inside the population, then you expect that once you try to find out their sample mean those properties of the population will be reflected from the sample mean also. I hope I have made it clear. Think about it, try to look into books, try to read more and I will see you in the next lecture.

Till then take care and good bye.