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 – 14
Simple Random Sampling
SRSWOR, SRSWR and Selection of Unit

Hello, welcome to the course Essentials of Data Science with R Software part 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 the topics of Simple Random Sampling with R Software.

So, you may recall that in the earlier lecture I started on the topic of simple random sampling and I have shown you that how the simple random sampling procedure can be executed and I also have given you a demo, my idea of demo was to just to show you some possible mistakes which one can make in real life.

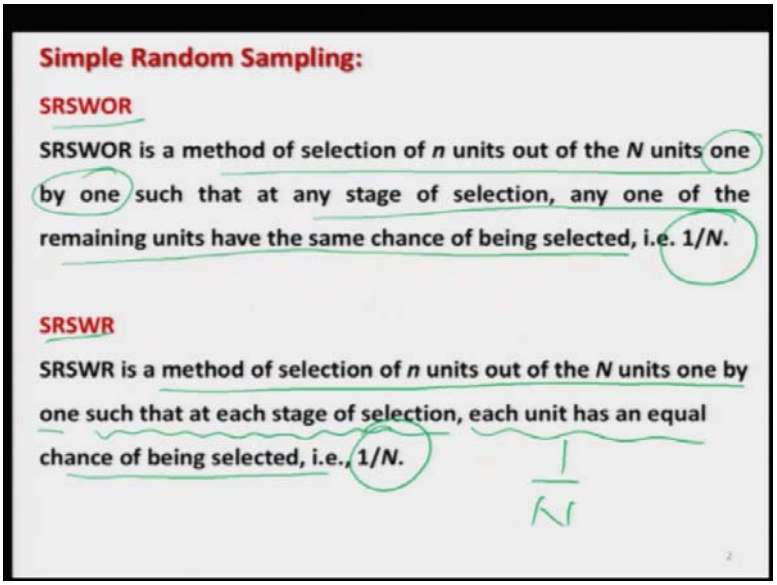
So, the moral of the story that we learnt from the earlier lecture was that we have to use the simple random sampling only when you do not expect that the variability among the sampling unit with respect to the characteristic under study is going to be large. And you will see that when we go to the mathematical say steps statistical analysis, we will assume that the variance of all the observation remains the same.

So, now if you try to translate this idea to data science, data science, remember, the data science is very large, you cannot view that data with your eyes. So, you need to do some checks to see whether the variance of different samples across the population they are nearly the same or not and there are some other checks which you need to do and you need to verify.

That if you are going to select a sample which sampling scheme has to be used. Please do not close your eyes and use only the simple random sampling as an ultimate solution. If the variability among the units is high and if you use the simple random sampling, the results are not going to be efficient, well what to do in those cases that we will see in the

forthcoming topics. Well now let me start this lecture and I will try to give you some more ideas, I will try to continue on the same lines that I did in the earlier lecture.

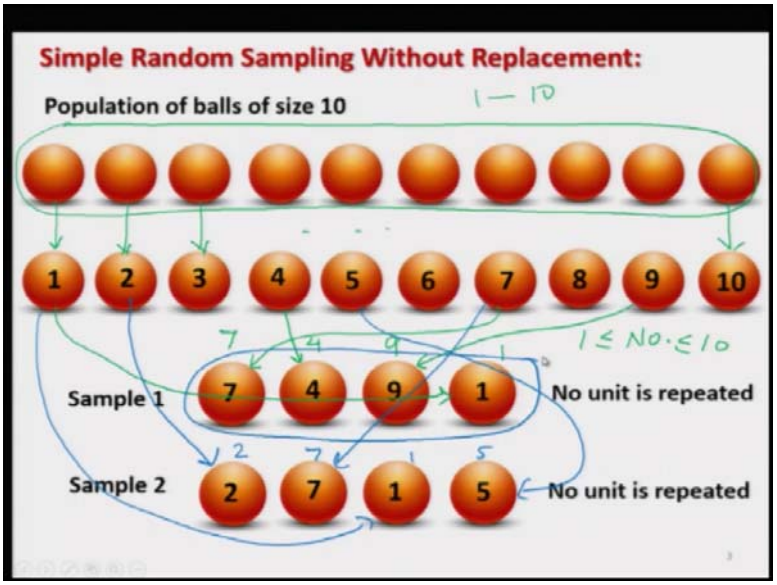
(Refer Slide Time: 02:59)



So, you may recall that we started the definition of simple random sampling and we had discussed 2 things one was SRSWOR and SRSWR and just for the sake of quick revision. Simple random sampling units out of a population of N number of units and the sampling units have to be drawn one by one.

Such that at any stage of selection any one of the remaining units have the same chance of being selected. And in case of simple random sampling with replacement, this is also a method of selection of n number of units out of N number of units, one by one, but in this case the units are drawn such that at each stage of selection each unit has got an equal chance of being selected and in both the cases, the probability of selection of a unit at any stage or at every stage, this will always remain the same as 1/N. In this lecture I will try to show you that in case of SRSWOR, how it is possible and how it comes out to be 1/ N. Whereas, in the case of SRSWR this is straightforward this is obvious.

(Refer Slide Time: 04:08)



Now, let me take here one more example and I try to clear some more concept, right. So, you can see here I am considering here a population of size 10 in which I have considered these 10 number of balls, they all are in the same color red and now my objective is this I would like to draw a sample of size 4 using SRSWOR. So, the first step is this I have to identify all the sampling units in the population and I have to create the sampling frame.

So, I try to choose here the numbers from 1 to 10 because my population size here is 10 and I number all the balls. So, which I have done here for example, this ball has been numbered as 1, second ball as number 2, third ball as number 3 and so on the last ball is now numbered as 10. Now, now I try to draw a sample.

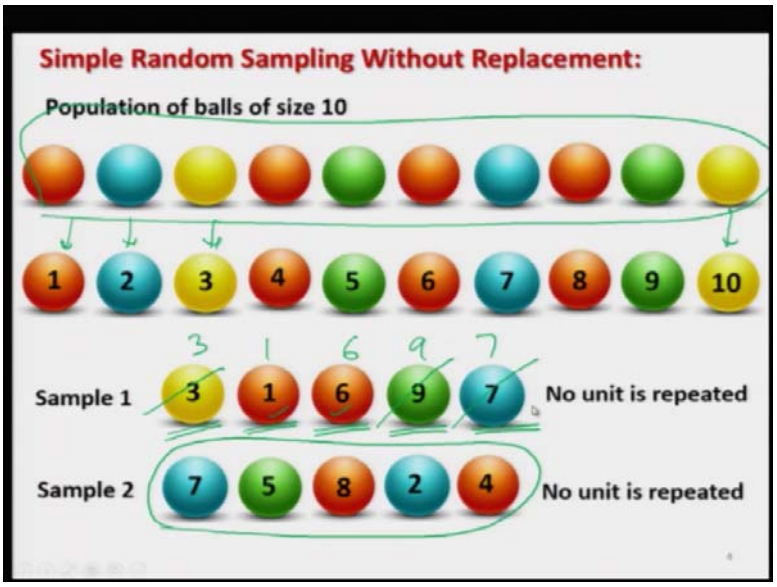
So, what I try to do? At random, I try to choose a number which is say, any number which is lying between 1 and 10 and suppose this number comes out to be 7, then what I do? I try to choose this 7^{th} number ball in my sample, then I choose another number I try to for example, I can use the random number table and I draw second random number and suppose this random number comes out to be 4.

So, now I try to choose here 4 number ball and this is now a part of my sample. Then I try to choose the third random number and suppose this comes out to be 9. Now I try to choose the 9 number ball and now 9^{th} ball is a part of my sample, then I draw the fourth random number suppose this comes out to be 1 and then I try to choose the ball number 1 in my sample.

So, this is how you can see here that now this is my here sample consisting of 4 ball and you can see here that none of the ball is repeated similarly if I try to take another sample from the same population. And so, I try to choose 4 number numbers from the random number table suppose those number comes to be 2, 7, 1 and 5 and then I try to draw the corresponding number balls in my sample. right you can see here.

And in this case also you can see here that none of the ball number is repeated. One thing what you have to keep in mind that when you are trying to draw different samples. Usually the probability of having the two exactly same sample is very very less and this I will try to show you little later on when I try to compute the probability of selection of a sample. So, this probability is going to be extremely less. So, it is; so, it is very highly unlikely that two samples are going to be exactly the same, ok.

(Refer Slide Time: 07:41)



And now I try to extend this example with different colors of ball so that it becomes more clear. Suppose once again I try to take here these 10 balls, but this time the balls are of different colors, right. So, you can assume that the variability is here high.

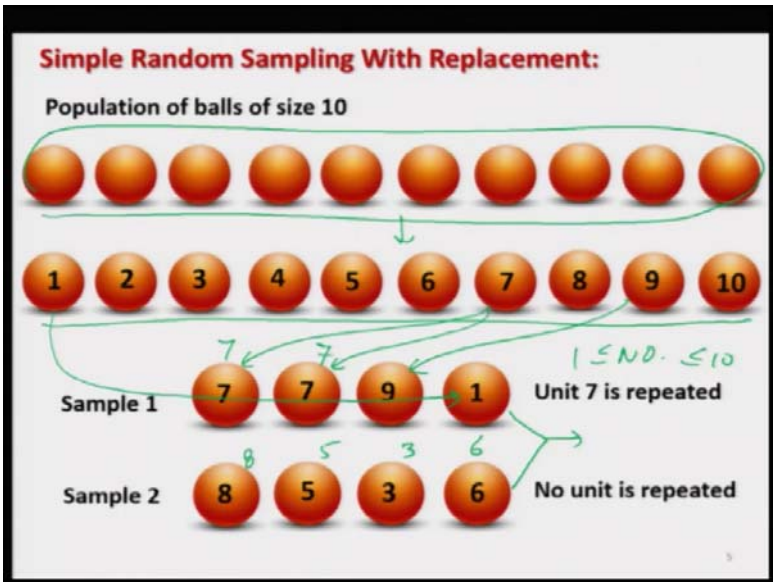
So, now you can see here I try to give the balls a number, ball number first as 1, second as 2, third as 3 and so on here 10^{th} is yellow ball is given as number 10. Now I have to draw a sample of suppose here size 5. So, I try to draw here 5 numbers by the using the random number table suppose those number comes to be 3, 1, 6, 9, 7.

So, you can see here that to these 2 ball 1 and 6 which are in color red, there are 2 red balls, 1 yellow ball, 1 green ball, 1 blue ball. So, you can see here none of the sampling unit is repeated because this is simple random sampling without replacement once you have taken out one ball in the sample, you cannot get it from the population again.

Similarly, if you try to choose your second sample you can see by the colors of the ball that the second sample may be different and remember one thing that the ordering in which the balls are drawn, they will also be different with a very high probability, ok.

For example, you can see here in the 1st draw you try to draw ball number 3, in the 2nd draw you are drawing ball number 1, in the 3rd draw you are drawing ball number 6, in the 4th draw you are trying to draw 9th number ball and in the 5th draw you are drawing 7 number ball in sample number 1, ok.

(Refer Slide Time: 09:26)



Now, I come to a simple random sampling with replacement. So, I try to repeat the same example under this new setup, you can see here I have taken here 10 red balls. Now I try to give them number and I try to create my sampling frame and then once I have given these balls a number from 1 to 10, then I have to choose any number between 1 and here 10, but now here the difference with respect to the earlier example is that in case if a number is repeated, I will not ignore it, but I will consider it.

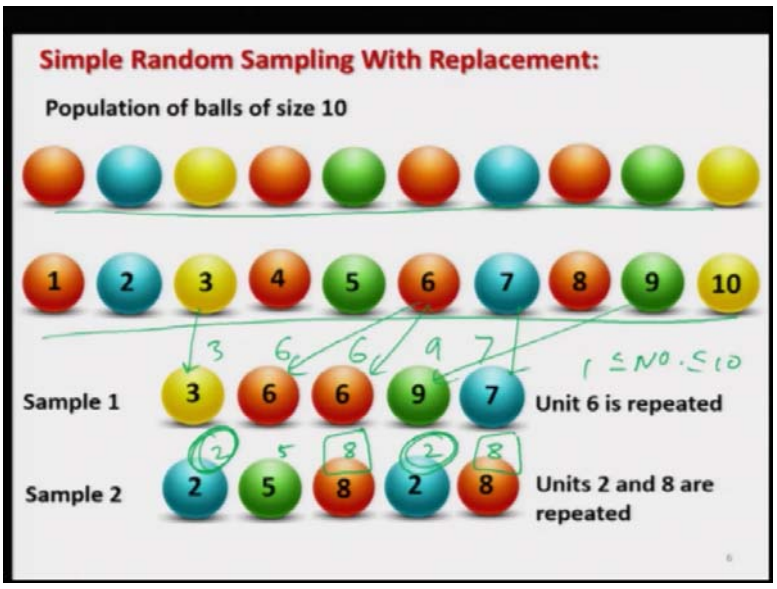
So, suppose I try to draw here a number suppose that number comes out to be here 7. So, I try to draw here the seventh ball in my sample, now the second number also comes out to be 7. So, I try to take the same ball once again because after drawing the seven number ball I have replaced it back in the population.

And similarly, the third number comes out to be 9 and then I try to choose the 9^{th} ball in my sample and suppose the fourth number comes out to be here 1. So, I try to choose the ball number 1 in my sample.

And similarly if I try to take one more sample and suppose my random numbers are obtained as 8, 5, 3 and 6. So, ball number 8, 5, 3 and 6 will now be included in my sample, but the difference between the sample number 1 and sample number 2 is that, you can see that in the sample number 2 no sampling unit is repeated.

So, the sample may look like as if this is the simple random sample without replacement, but it is actually not. So, by looking at the sample you really cannot judge whether the sample has been drawn by SRSWR or SRSWOR.

(Refer Slide Time: 11:11)



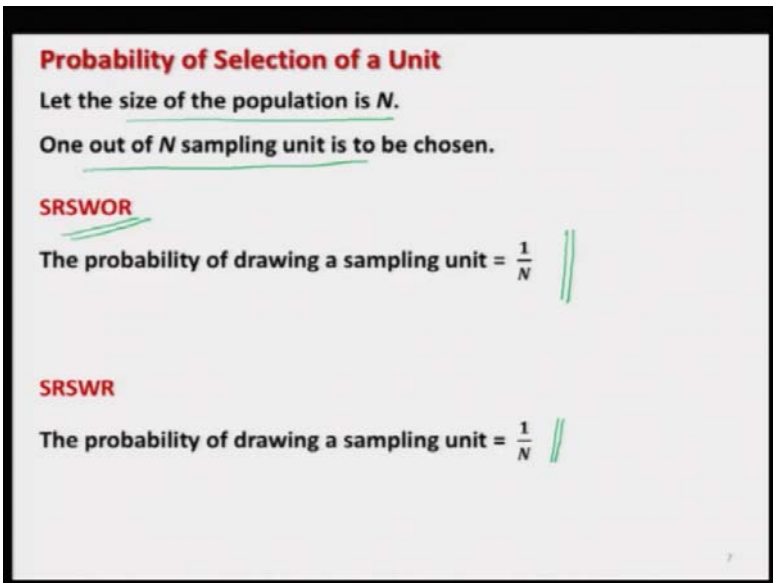
And similarly. if I try to take one more example of colored ball. So, again I have here 10 balls which I am trying to number from 1 to 10 and after that I try to find out here a number which is lying between 1 and 10 and then I have to draw my sample. So, suppose if I draw here 5 numbers and those number comes out to be suppose 3, 6, 6, 9 and 7. So,

I try to draw the corresponding number balls in my sample ball number 3, ball number 6, 2 times ball number 9 and ball number 7.

Similarly, if I try to take one more sample. So, in this case you can see here that the random number drawn are 2, 5, 8, 2 and here 8. So, you can see here that number 2 is repeated 2 times and number 8 is also repeated 8 times.

So, it is possible that 1 number can be repeated more than 2 and in an extreme case the number 2 can be repeated 5 times. So, the entire sample consists of only blue balls. Well that is a different thing whether this is representative or not, but this is the way samples are going to be obtained using the simple random sampling with replacement, ok.

(Refer Slide Time: 12:27)



So, now I come to computing different types of probabilities. So, first of all I try to compute the probability of selection of a unit, but before going into the computation of probability of selection of a unit. let me clarify one thing. I am going to compute two types of probabilities; one is the probability of selection of a unit and another is probability of selection of a sample.

So, remember one thing probability of drawing a sampling unit means I am trying to draw only one unit. Whereas, when I am saying that I am going to draw a sample; that means, I am choosing a n number of units, sometime people get confused with these two probabilities, but this is my simpler simple clarification.

When you are trying to choose the probability of drawing a sampling unit; that means, one sampling unit from a population of size N and once you are imposing a condition that you are following the simple random sampling methodology. So, obviously, you have to draw the sampling units in such a way such that all the units have got the same probability.

So, the most simple answer to this question that what is the probability of selection of a unit is 1 upon total number of possible cases that is N because any unit can be chosen by N number of ways. For example, if I try to choose the first draw, in the first draw there can be 1^{st} unit or 2^{nd} unit or 3^{rd} unit or the Nth unit. So, the total number of possible cases are N. So, the probability of selection of any unit will be 1/ N. So, that is what I am going to explain here, ok.

So, suppose that will be our standard symbol that the size of the population will be assumed to be N and now my objective is this one out of N sampling unit is to be chosen. So, in the case of simple random sampling without replacement, the probability of drawing a sampling unit will be simply 1 / N, 1 / N as I said and in case of SRSWR also the probability of drawing sampling unit is here 1/ N.

Now, at this stage there is a confusion the confusion is when you are trying to draw a sample by SRSWOR, then we are saying that the probability of selection of unit is equal to 1/ N, where N is the population size.

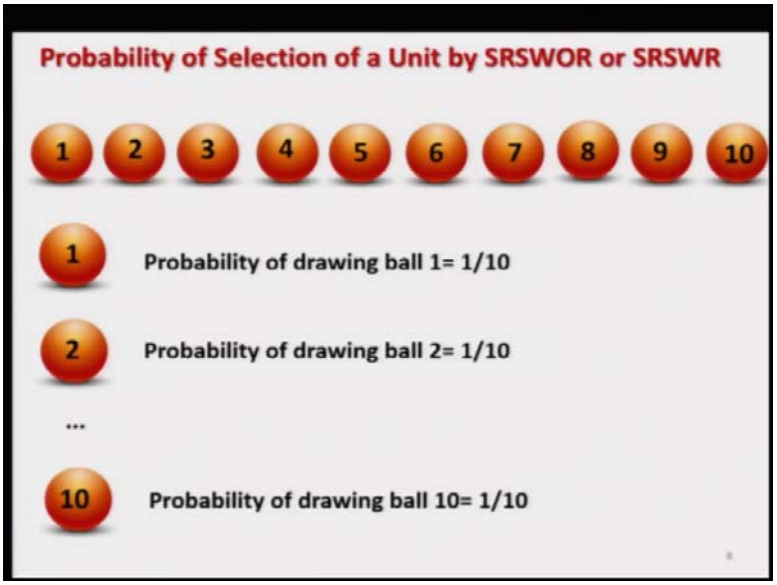
So, when you have drawn the 1^{st} unit then the total number of units remaining in the population is only N - one, not N. So, how it is possible that once you have drawn the 1^{st} unit out of the population the remaining units still have the same probability 1/N how it is possible.

This is the simple very that is simple question that sometime create a big confusion in the minds of the students, but as I said earlier these things are not coming by thought or by logic they are coming by pure statistical and mathematical ways. So, I will try to show you with the mathematical proof that how this probability is going to be 1/ N.

Well in case of SRSWR with replacement, there is no such confusion. Because at every stage you try to suppose you try to draw one ball then you are putting it back. So, at

every stage the total number of balls in the balls will remain the same, the total number of sampling units will remain the same as N, right.

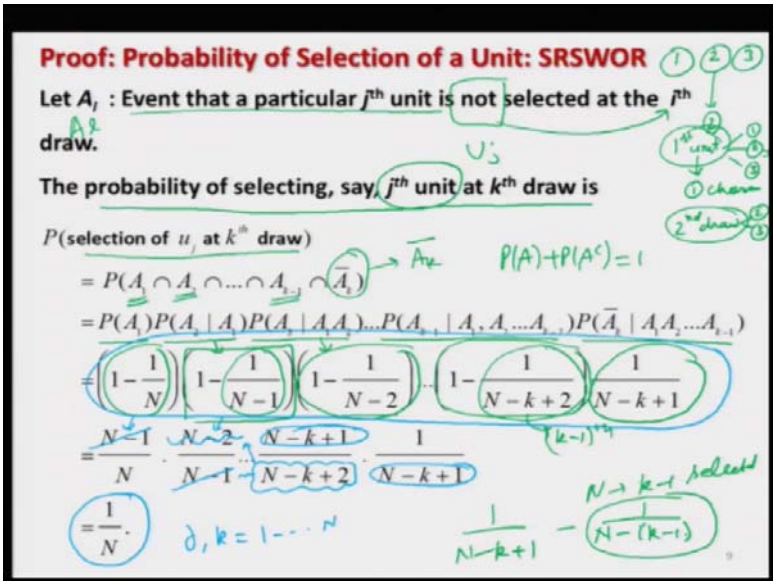
(Refer Slide Time: 16:31)



So, now let me try to just explain it here little bit and then I will try to show you the proof of this probability. So, now, in the same example where I have taken 10 balls and I have numbered them from 1 to here 10 like here like this.

So, once I said the probability of drawing the sampling unit is 1/N; that means, if I want to draw the ball number 1 then the probability of selection of ball number 1 is 1/10, similarly the probability of selection of ball number 2 is 1/10 and similarly the probability of selection of ball number 10 is also 1/10, right ok.

(Refer Slide Time: 17:04)



Now, I give you here a proof of this probability. So, now, you can see I have explained you and you also have understood that once you give a mathematical proof then the things become more clear more convincing. So, now, I am going to use little bit probability theory at an elementary level and, that is what I assume that you have a fair idea of the basic probability and basic statistics.

So, I try to define here an event this event here is A l and this is an event that a particular jth unit is not selected at the ith draw because you can see here suppose if I try to take a simple example, that suppose I have got here three balls- ball number 1, ball number 2 and ball number 3. So, suppose I consider the ball number 2 and suppose I want to have a sample of size 2.

So, now when I try to draw the first unit, then there are 3 possibilities that ball number 1 may come, ball number 2 may come, ball number 3 may come. And suppose in the 1st draw suppose the ball number 1 is chosen. Now once then I come to the 2nd draw then in the 2nd draw I have a possibility that ball number 2 and ball number 3 are there the. So, these balls are going to be chosen.

So, now if you try to see here, ball number 2 can be chosen in two different ways either it is selected in the 1st draw here or in the 2nd draw here. So, this is what I mean when I say that the jth unit is not selected at the ith draw for example, the 2nd unit is not selected at the 1st draw or at the 2nd draw something like this, ok.

Now, I try to find out the probability of selecting the jth unit at the kth draw. So, now, suppose I try to indicate this jth unit by here u_j. So, u_j is in nothing, but your sampling unit, right, whatever is the whatever are the units in the population one of them is going to be jth unit. So, now, if you try to see just using the basic theorem of multiplication of probabilities. means I can write down here. What is the probability of selection of the jth unit at the kth draw? This is going to be that the kth unit is not selected in the 1^{st} draw.

So, this is going to be the A_1, A_1 is the even that the kth unit is not selected at the 1^{st} draw. Then if kth unit is not selected in the 1^{st} draw, not selected in the 2^{nd} draw, not selected in the 3^{rd} draw and up to the kth unit is not selected in k - 1^{th} draw. And finally, in the kth draw the unit is selected the kth unit is selected. So, the probability will become A_k complement, right, you know that probability of A plus probability of an event a complement is equal to 1.

So, this is the complement probability. So, now, using the multiplicative theorem of probability I can write this expression it is something like $P(A \cap B \cap C)$ and so on. This I can write down here probability of A_1 into $P(A_2|A_1)$ conditional probability, $P(A_3| A_2A_1)$ and so on $P(A_{k-1}| A_1A_2A_{k-2})$ and at the end probability of $P(A_k^c| A_1A_2A_{k-1})$.

Now, look let us try to write down these probabilities and now you have to be very careful you will try to understand what you are thinking, is actually what? So, here I will give the answer. Now I try to write down probability of A_1, probability of A_1 is that the kth unit is not selected in the 1^{st} draw. So, the probability of selection of kth unit at any draw is 1/ N that we have understood. So, the probability of not selection will become 1 - 1 / N. I think that is straightforward.

Now, you come to 2^{nd} draw; in the 2^{nd} draw since one of the unit has already been selected in the 1^{st} draw. So, the total number of possible units available in the 2^{nd} draw will be N - 1 and the probability of choosing any unit in the 2^{nd} draw will be 1/N - 1, but you are interested in the event that the particular kth unit is not selected in the 2^{nd} draw.

So, this will become 1 - 1/N - 1 like as here. Now we come to draw number 3 and the 3^{rd} draw since out of N unit 2 unit have already been drawn they are not the kth unit, but they are some other unit. So, the total number of unit which are available for the selections are only N - 2.

So, the probability of selection of any sampling unit in the 3rd draw will become 1/N - 2 and the probability of not selecting the kth unit in the 3rd draw will become 1 - 1/N - 2 and so on I can continue up to k - 1th draw.

So, at the k - 1th stage the probability of selection of the kth unit in the k - 1th draw will become 1 - 1/N - k plus 2 and the probability of not selection will become 1 - 1/N - k plus 2 and finally, out of N units k - 1 units have been already selected. So, now the total number of remaining units are N - k - 1.

So, the probability of selection of any unit in the kth draw will be 1/N – (k – 1). So, and this is same as suppose here 1 / N - k + 1. So, the probability of selection of the kth unit at this draw will become 1/N - k + 1. Now this is your here entire probability from here to here, you have the slides also so you can clearly look at this expression I have to write over this.

So, now this first term becomes here N - 1 / N second term becomes here N - 2/N - 1 and so on. I try to simplify and now if you try to see I try to simplify it this N - 1, N - 2 get cancel out this N - 2, N - 2 will get cancel out with something in the denominator of the third term.

And similarly this N - k - 1 and N - k - 1 also get cancel out and this term will also get cancel out with the numerator of the term receding to it. And finally, you have here only 1/ N.

So, you can see here the probability of selection of jth unit u j at the kth draw is simply 1/ N. Now I have not given any particular number to j or k. So, j and k they can take any number between 1 to N, right. So, this means that in case if you try to draw any sampling unit by simple random sampling without replacement the probability of selection of a unit will still remain 1/ N, good.

Now, let me stop in this lecture, but this lecture has given you one very clear idea that whatever you are thinking that may or may not be correct unless and until it is verified by the statistical tool by the mathematical tools.

Now once you come to data sciences as I explained in the first lecture sometime the algebra becomes very complicated, the mathematical tools give us something which is

very difficult for us to understand or to get a very clear idea whether things are increasing, decreasing or what is happening, in that case these computations help us.

But still we have there is a long way to go and these things will get cleared in a stepwise way as we go further into the course, but now this is the time where you have to start thinking that, what is the role of statistics in a data science? And what is the role of mathematics in a data science? The concepts are coming from statistics that is playing a very important role.

And now mathematics and computers they will help us in handling the large data sets big data set doing computation using optimization techniques and helping us in solving mathematical complexities. So, you think you practice and I will see you in the next lecture till then, good bye, take care.