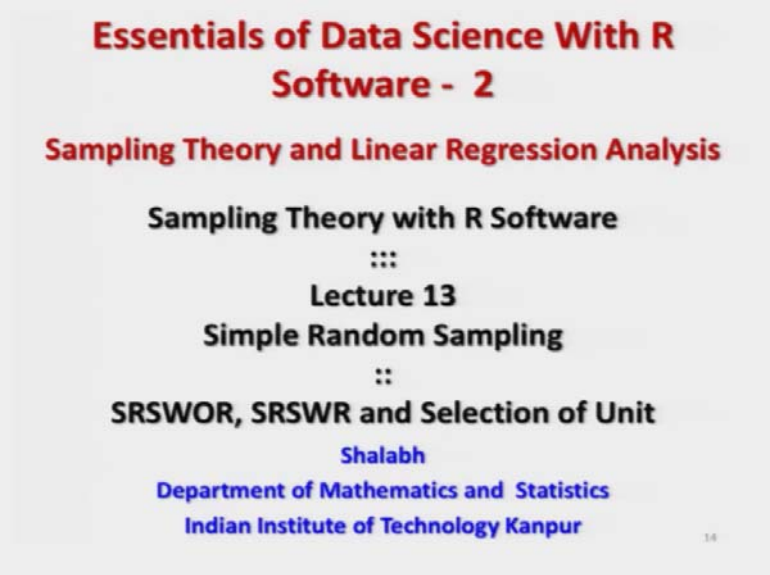
Essentials of Data Science with R Software – 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Sampling Theory with R Software
Lecture – 13
Simple Random Sampling
SRSWOR, SRSWR and Selection of Unit

(Refer Slide Time: 00:13)



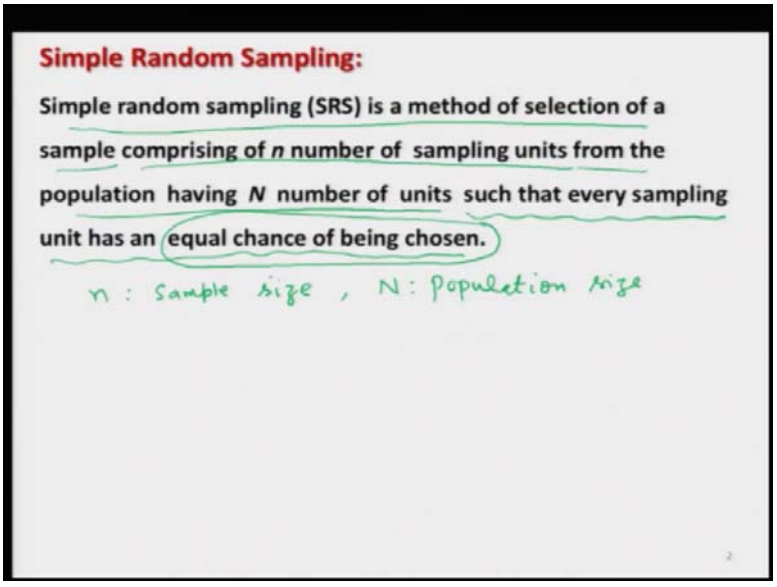
Hello, welcome to the course Essentials of Data Science with R Software- 2; where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. And from this lecture we are going to begin with the new module on the topics of simple random sampling. So, as we had discussed in the earlier lectures that there are various types of sampling methodologies, sampling techniques which help us in drawing a representative sample.

So, there are couple of things that the first question comes that under what type of condition you have to choose what type of sampling scheme. Second point is this; you have to choose a sampling scheme in such a way such that the drawn samples are satisfying the basic assumptions of the statistical tool which are going to be used at a later stage.

Thirdly, whatever samples you are obtaining which are essentially the numerical values of some population, from some population they have to be representative. All the salient features of the population should be present in the sample also.

So, keeping in mind all such characteristics of a sampling scheme, we will start first with the most-simple sampling technique sampling methodology simple random sampling ok. So, let us now begin our lecture and first we try to understand; what is simple random sampling and how it is executed, ok.

(Refer Slide Time: 02:09)



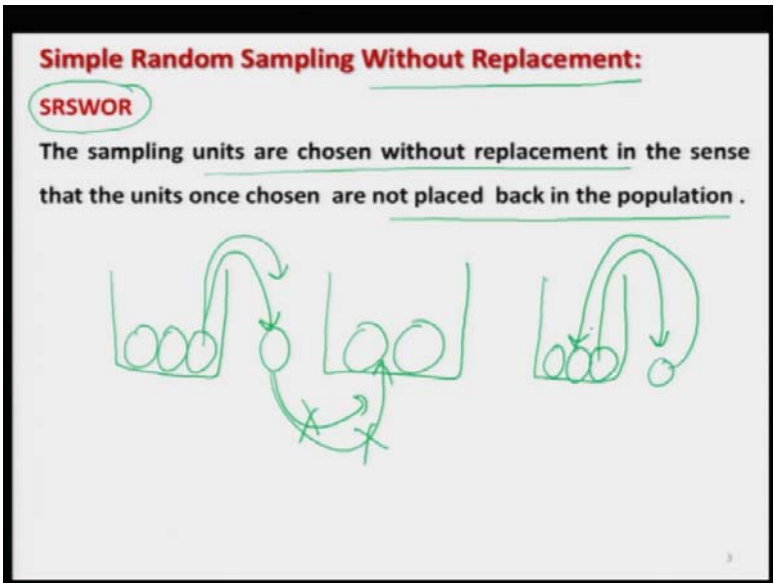
So, simple random sampling is a method of selection of a sample comprising of a small n number of sampling units from the population of having N number of units and thus collection is made in such a way such that every sampling unit has an equal chance of being chosen. So, now, you can see here one thing that I have introduced here a symbol small n. So, this is going to be our sample size and N this is going to be our population size.

What are these sample size and population size? These are the total number of sampling units present for example, in a sample of size n there are small n number of sampling units and in the population of size N there are N number of sampling units.

The most important aspect of simple random sampling is that; that all the sampling units they have been chosen in such a way such that the probability of selection of any

sampling unit is the same and the question is, you have to execute it in such a way such that the chances of selection of any sampling unit remains the same. Now under this simple random sampling I have two options.

(Refer Slide Time: 03:50)



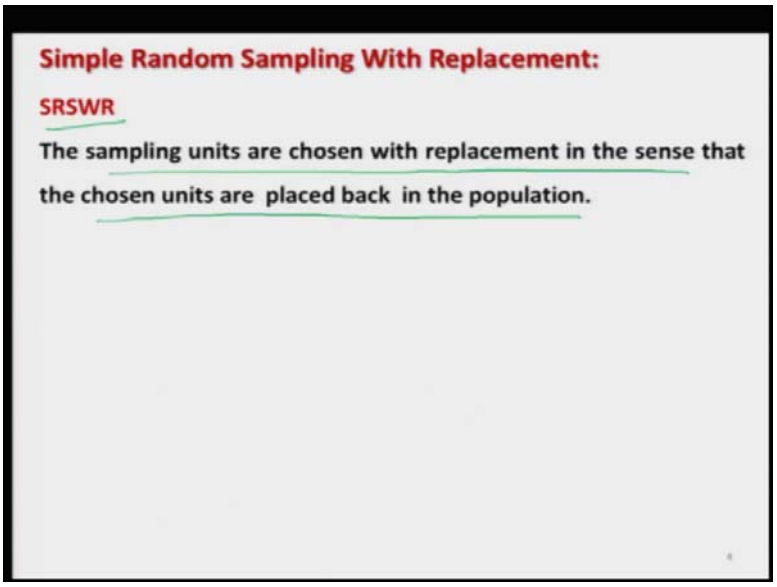
First option is this, we can have simple random sampling without replacement and next choice is simple random sampling with replacement; what does this mean? It simply means that when we are trying to choose the sampling units from a population, once the sampling unit has been chosen, then in case it is not replaced back in the population. So, the sampling units are chosen without replacement, that in case if they are chosen they are not going to be placed back in the population. So, this method is called simple random sampling without replacement.

For example, if I say here suppose if I have here a container which has got suppose here three balls, right. Now suppose I choose one ball in my sample now in the container only two balls will be remaining whatever ball I have chosen here, this is not going to go the container again.

So, that is simple random sampling without replacement and this is also suggested by the name, this is with-out replacement; that means, there should not be any replacement and symbol to represent the simple random sampling without replacement is, what we are going to use here is the SRSWOR briefly.

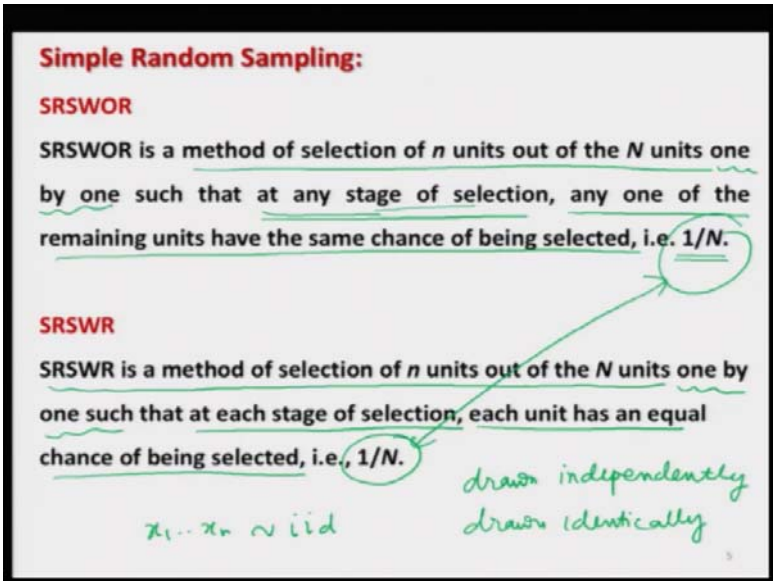
So, that will be our symbolic notation. Now in this case you have one more option. You are saying that once you have withdrawn a ball from the container, then you are not putting it back in the container. So, the next option is this I can replace it back. So, suppose I have here a container which is again suppose I have three balls, I take it out, record the value and then I replace it back inside the container. So, this is going to be your simple random sampling with replacement.

(Refer Slide Time: 06:14)



And in case of simple random sampling with replacement which is briefly denoted as SRSWR, the sampling units are chosen in such a way such that the chosen units are placed back in the population and in both the case simple random sampling and simple random sampling with and without replacement the probability of selection of a unit will always remain the same.

(Refer Slide Time: 06:50)



So, now we try to have an expression about relating the probability of drawing a unit with the simple random sampling. So, in case of SRSWOR this is Simple Random Sampling Without Replacement, this is a method of selection of a small n number of sampling units out of N number of sampling units and these units are drawn one by one such that at any stage of selection any one of the remaining units have the same chance of being selected.

So, since there are N number of sampling units in the population, this probability will become 1/N. So, 1/N is the probability of selection of any sampling unit at any stage of selection under simple random sampling without replacement.

Well at this statement creates a sort of confusion in the minds of students because once you are trying to do the simple random sampling without replacement, you are saying on one hand that there are N number of units in the population.

So, once you draw the first unit; obviously, the probability will remain 1/N. But once you have taken out one unit out of the sample then there will be N - 1 number of units in the population.

So, a natural question comes in most of the students mind is that why not this probability is now becoming 1/N - 1 or why this probability remains the same 1/N. This probability should be changing because the population size is changing? Yes, now that is the point

where I say you need some statistical methodologies mathematical methodologies statistical concept which can sort out these types of confusion.

I promise you after some slide, after some time I will try to prove it mathematically, that in case of simple random sampling without replacement the probability of selection of a unit at any stage will always remain 1/N it will not be changing like 1/N - 1. 1/N - 2 and so, on.

How it happens? We need some more basic concepts to understand this thing, but please take it here as such that the probability will remain 1/N at every stage at any stage ok. Now the similar story is with the simple random sampling with replacement.

So, simple random sampling with replacement is a method of selection of a small n number of sampling units out of a population of N number of sampling units and in this case also the units are drawn one by one and they are drawn in such a way so, that at each stage of selection each unit has an equal chance of being selected which is one 1/N. So, you can see here that in case of SRSWR as well as SRSWOR the probabilities of selection of units they remain the same.

Although, in case of SRSWR there is no confusion that the probability will remain the same because at every stage once you draw the unit outside the population you are replacing it back. So, at any stage the number of units in the population will always remain capital N.

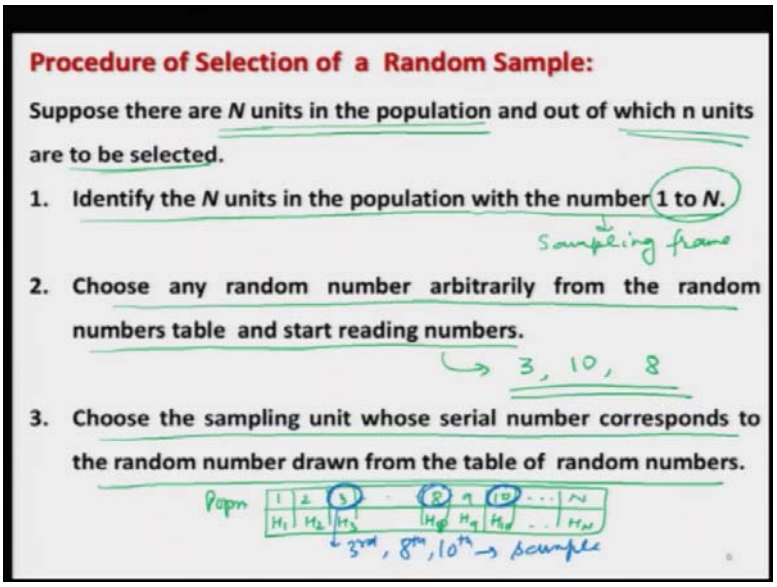
So, there is no confusion in this stage. Now you see the outcome of these two methods will be that the sampling units which you have drawn, they are drawn independently and since you are trying to draw all the sampling units from the same population because the population is not changing in any draw.

So, all the units which are drawn, they are also identically distributed and you may recall that in statistics whenever we use a tool, one of the basic fundamental assumption in most of the statistical tool or for the applicability of most of the statistical tool is that the observations x 1, x 2, x n they are i i d Identically and Independently Distributed.

Which means all the observation have been drawn from the same population and they have been drawn independently; that means, the occurrence of drawing of one sampling unit does not affect the occurrence of drawing of another sampling unit in the population.

So, now the question is how it works and what is the methodology? So, what I do first I try to explain you the methodology and then I will try to show you sort of demonstration that how you can draw the sample.

(Refer Slide Time: 12:31)



So, first we try to understand what is the procedure of selection of a random sample. Suppose there are N number of units in the population and our objective is this we want to draw a small n number of units from this population that will be our sample.

So, first step is this you have to create a sampling frame; that means, you need to identify the N units in the population and you have to give them a number say 1 to capital N. So, now, you can correlate it directly to the concept of sampling frame. Why? Because sampling in sampling frame what you are doing? You are attaching a tag to every unit in the population.

So, suppose there are 50 students in a class suppose and you want to draw a sample of size 10. So, in the first step what you have to do? You have to assign role numbers to all the 50 students which are unique and non-overlapping. So, similar to the story when I say please try to number all the sampling units in the population like 1, 2, 3, 4 up to

capital N. Now in the second step you have to choose any random number arbitrarily from the random numbers table and start reading the numbers what does this mean?

Well before I try to explain you that what does this mean let me give you the concept of random number tables. You see we want to draw a number randomly. So, the question is how to get it done? Means, if you ask any person to choose any particular number, there will always be some hidden personal bias those biases may be sometime the people may try to choose a number which they feel somehow that is lucky for them or they may choose some number which is very complicated or very easy.

So, any choice of number will always will always attach some hidden human biases. So, if you ask me or I ask you to suppose select say 20 numbers between 1 to 100, there is a very high chance that you will try to select some prime numbers, some numbers which are not evenly distributed or say evenly distributed and it is possible that all of the numbers are between only between 1 to 10 or say they say 1 to 50 and so on. So, there are various types of problems that may occur.

So, in order to order to avoid such problems, one simple option is this that you try to prepare some chits say 0 to 9 digits and try to put them inside a bottle jar or some box, try to shake it and then try to take out one number out of that. In this case whether there is bias or not; I do not know, but I cannot question you and I expect that in this process there will be no bias, no hidden human bias whatever number is coming out that is my random number because before choosing that number I had not thought about it, ok.

So, based on this idea people have created the random numbers tables that mean they have drawn those numbers by say lottery method and they have compiled those numbers. So, at least the sequence of those number is completely random. So, earlier when software was not there, we used to draw the random numbers from random number tables and still when I come to the data science, still we try to do the same thing, but the same process has been has been replaced by a software or by a programming language.

For example, in the software whenever we are trying to draw a number usually that is associated with the discrete uniform number or discrete uniform distribution, a number which is generated from the discrete uniform distribution ranging between 1 to N and

when you are trying to generate the random numbers from a software, there are some more issues.

Although, this is not a platform where I can discuss the details of those things, but I can tell you the random numbers which we are trying to generate from the software they are based on some algorithm and whenever you are trying to generate the random numbers from an algorithm, they cannot remain 100 percent independent there will certainly be some correlation among those or between those numbers.

I agree that correlation will be very close to 0, but definitely that cannot be 0. So, in computer science, people are trying to come up though with those algorithms which can produce the random numbers in such a way such that the correlation among the number is extremely small and its ideally it should be very close to 0.

For example, if the correlation coefficient between two numbers is say 0.0000002 for all practical purposes, I can take it to be 0 and very important part, the random numbers which are which are generated from the software from the computer programs, they are essentially called as pseudo random numbers. Why they are called pseudo random numbers? Pseudo is p s e u d o pseudo, why they are called as pseudo because they look like independent, but they are not actually independent.

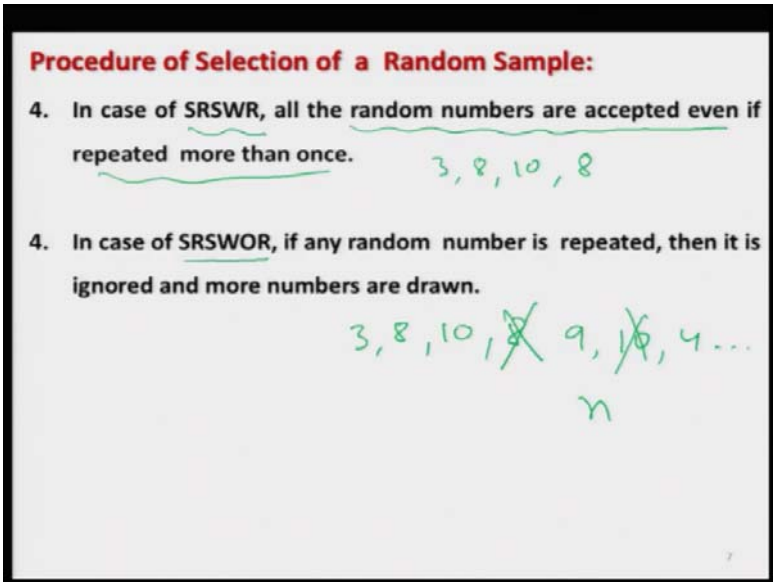
So, that is the now the difference between the drawing of random numbers in the classical sampling theory and now in the random of random number generation in the data science, right. So, this is what you have to keep in mind and this is what I told you that I will try to interlink the concept of classical sampling theory with the data science and as I go further in the topics, I will try to do such more analysis ok.

So, now, let us come back to our procedure. So, now, what I am trying to say that you try to choose any number arbitrarily from the random number tables and start reading the numbers. Now whatever number you have drawn, try to choose the sampling units whose serial number correspond to the random number drawn from the table of random numbers, for example, in the second step suppose you try to draw the random number say here 3, 10 and say here 8.

Now, what you could do? You come to your here population, right in the population what could be there? It will be something like you try to write down here the you have assigned some numbers to the population. So, somewhere there will be here 3, somewhere there it will be here 8 and some somewhere it will be here say here 10 and so, on and there will be some values here, suppose if I say height. So, this is the height of the person number 1, this is the height of person number 2, this is the height of person number 3, this is the height of person number 8, it is the height of person number 9, this is the height of person number 10 and so, on this is the height of nth sampling unit.

So, now I have obtained here three numbers 3 10 and 8. So, what I do? I come here 3, 8 and 10 and I try to choose here the 3^{rd} person, 8^{th} person and 10^{th} person in my sample and whatever are the values of their height that means the heights of the 3^{rd} 8^{th} and 10^{th} persons they will constitute my random sample, ok.

(Refer Slide Time: 21:40)



Now the only difference between SRSWR and SRSWOR will be that in case of simple random sampling with replacement, if a random number is repeated then even then it is accepted. For example, I have drawn here a number 3, 8 and 10 and suppose the next number comes out to be here 8. So, well, that will be will be a part of my sample because here the repetition is allowed.

On the other hand, when you are trying to consider the simple random sampling without replacement in this case what are you doing? Once the unit has been drawn, it is not replaced back in the population.

So, any random number if that is repeated, it is discarded, it is ignored and more random numbers are generated till we reach the sample size n. For example, if in this case suppose I have drawn 3, 8 and 10 number and suppose again 8 comes. So, I will ignore it and I will say take another number suppose this is 9.

Now, next number comes here 10; 10 is again repeated. So, I try to ignore it and I try to keep on say this another number suppose 4 and so, on I will continue till I get a small n number of numbers, right. So, this is about simple random sampling with and without replacement that how they are executed. Now before I go further let me try to show you a demonstration, right that will try to explain you several other aspects ok.

(Refer Slide Time: 23:27)

So, now, you see what I have brought in the class today. I have brought here a jar in which I have some cashews and some almonds and I would like to give you a demonstration that how really a simple random sample is drawn.

So, if you try to see here in this bottle, we have some cashews and some almonds and my random variable here is dry fruits, my random variable is not cashew or almonds. So, now this cashew and almonds both comes under the category of dry fruits, ok.

So, I would now I would like to draw here a sample of say some given size say 5 from this population of cashew and almonds, I can count them also and now just for the sake of understanding, you can believe that whatever cashew and almonds are there in this bottle, I have given them a name 1, 2, 3, 4 up to say this 40, right although I cannot write these names on this almonds and cashews, but, but you can imagine you can assume that those names are written on every almond and every cashew, ok.

So, now, you see if you ask me to draw here a sample of size 5 from this population. So, what I will do?

(Refer Slide Time: 25:09)

I will simply open it and you see I am looking before you, I put my hand and I try to take out here. See here some number of dry fruits. So, you see I have got here one cashew, one almond, one cashew 3 one almond and 4, 5 this cashew and I have got here this sample do you think that this is a random sample? Well, you can say because I was looking at you and I was putting my hand inside the jar and I am and I was trying to take out the sample, it is not a random sample, why?

Because you see what I did I just put my hand and whatever was there on the surface of those dry fruits, I picked them only from that place or I picked those dry fruits only from the surface. So, all those dry fruits which are in the bottom, they cannot come into my hand.

So, this is not a representative sample. So, now, what I try to do? I try to put them back and now I try to I am looking at you and you see I am putting my hand here and I try to take the thing from the bottom and I take it here and then I try to put them here; so 1 2 3 4 and 5, right.

(Refer Slide Time: 26:28)

But do you think that these dry fruits which I have taken now are they really a random sample? You can say yes why because means in the earlier case, I removed the means I made a mistake that I took the dry fruits only from the surface, now I am taking from the bottom. So, is it really a random sample? No, it is again not a not really a random sample.

Why? Because you have drawn the sample only from the bottom. So, now, I will raise one more question that all the dry fruits which are on the sides of the jar, they are also, they have no chance of being selected in the sample. So, that is again not a random sample, right.

So, now, what is the reason? Why this is not a random sample? One reason I can diagnose is that in this jar, all the dry fruits do not have an equal probability of selection. In the simple random sampling every dry fruit should have the same chance of being selected in the sample that is not happened.

So, what I try to do, I close it and I shake it, now you see I am shaking it and then now means all the almonds and cashew they are here and there and I have absolutely no idea which almond or which cashew is there and I can repeat this and I can continue with this process for some time, right. So, that you are hundred percent sure that I cannot manage or control the location of the almonds of the cashew, ok.

Now, I am you see looking at you and I am taking the sample from there. So, you see I have taken here some samples and I try to show you what I have got here. So, this is one cashew, one almond, one cashew, one cashew and one cashew. So, this is again I have drawn a random sample of size 5, but do you think that is it really a random sample?

I will say once again, no, this is not a really a random sample. Why? Because although you have randomized it, you have shaken it, but then you have taken the sample only from one place. So, it is possible that all the units in the population; that means, my cashew and almonds they do not have the same chance of being selected in the sample.

So, now I have given you several options, but every time I am saying that this is not the correct way then the question is; what is the correct way? Right. So, now, I am showing the correct way. So, you see first I will try to consider the simple random sampling without replacement that how are you going to draw a sample by this scheme.

So, I try to take it and you can see my eyes are always with you and my jar is there in the bottom. So, I cannot really see what is happening inside the jar. So, that will ensure that I that my personal bias or my personal intentions to choose any particular almond or cashew is not being reflected in the sample. So, that will ensure at least that that from my side, I am not trying to put any effort so that at the particular cashew or almond is being selected, ok.

So, now you see I am just shaking it, but I am looking at you right now whatever I am going to do that is the correct way of drawing a simple random sample with the out replacement. So, I try. So, I have shaken it now I am looking at you and I am trying to draw here one piece. Now it comes you see, this comes out to be a to an almond I am putting it here as my first unit which is selected in the sample now this is without replacement.

So, I am not putting it back I am not putting this almond back, but I am putting it here and I close my jar and I try to shake it again right and then I open it, I am looking at you and I do not know where my hand is going and I am trying to put here I am trying to take out one more unit and this turns out to be you see a cashew.

(Refer Slide Time: 30:53)

Now, this is my second unit. So, I try to keep it here and then I try to close my jar and I shake it here once again. Now you see I am means every time I am trying to randomize it. So, all the locations of all the almonds and cashews they are getting changed and I don't know where is what.

(Refer Slide Time: 31:30)

So, now I try to I am looking at you, but I am trying to draw here my third unit and this comes out to be an almond right. So, this is my unit number 3 similarly I can close it I can shake it as much as you want and then I try to put it take out another unit which is here almond right. So, now, I have got here a sample of size 4.

Now, I shake it again I randomize it and I try to open it, but I am looking at you. So, I do not know which of the thing whether almond or cashew or which almond or which cashew is going to come into my sample and you see this comes out to be an cashew a cashew.

So, now, you can see here I have this sample, right. Now this can be treated as a sample which is drawn by simple random sampling without replacement, ok. Now I try to put all the all of the dry fruit back and I try to show you how are you going to draw a sample using simple random sampling with replacement. So, once again I try to do the same thing, I try to shake it, I try to randomize it I can do as much as randomization I want more you do better is your sample, but yeah certainly there is a limit.

So, I try I am looking at you, I am putting my hand into it and I am trying to draw here one unit this comes out to be a cashew. Now this is my unit number 1, but since I am going for simple random sampling with replacement.

So, I am putting it back into my population and I try to randomize them and I try to draw one more sample, ok. So, I try to shake it here and yeah as much as you want you randomize it and then again I try to take out one more unit this comes out to be a cashew once again, but this is my unit number 2.

So, now, I am putting it back and then I again try to repeat this process for three more times. So, I can show you here I randomize it and I am not looking in my jar, but I am looking at you and I try to get here one more unit which is cashew. This is unit number 3, I put it back and then I again try to randomize it shake it as much as you can and then I try to open the jar and try to take out one more unit this again comes out to be cashew, right.

So, up to now I have drawn four units and all are cashew right and then I try to shake it again and I try to try to put my finger and I try to put my hand into this one and I this time I get here an almond. So, this is my unit number 5 and now I have drawn a sample of size 5 by simple random sampling with replacement, ok.

So, I put it here and now my population is the same which I from where I started. So, this is how we try to draw a sample using simple random sampling with and without replacement. Although, I am not going to discuss the topic of varying probability scheme, but still I would try to show you something here. So, that at least you know what is really what really happens and I would like to give you one more say this concept, but that will be only the concept.

(Refer Slide Time: 35:03)

So, you see I have got here some walnuts. So, I am putting this walnut into my bottle now my and if you try to see here what is the difference between cashew, almonds and walnuts?

(Refer Slide Time: 35:23)

So, you first you can see cashew and say this almond they are almost of the same size.

(Refer Slide Time: 35:33)

But if you try to compare them with walnut, you can see the size of the walnut is multiple times more than the size of either cashew or almond. So, I want to show you although you are taking here the variable as dry fruit and almond is also a dry fruit, cashew is also a dry fruit and walnut is also a dry fruit. So, all are uniform, but the size of the unit how it affects your sampling scheme.

So, I try to do, I try to put all the things inside the jar and I try to shake it now and I would try to take a simple random sample, you can say without replacement or with replacement whatever you want. So, suppose I try to take here simple random sampling with replacement.

So, I try to shake it here you see as much as I can and I try to open the jar and you see I am looking at you, but I put my hand and water come whatever comes; obviously, whatever will come in my hand from the surface I will try to put I will try to take it out.

(Refer Slide Time: 36:42)

So, I try to put my hand and I get here a walnut, right. So, now, I try to record its value whatever is my variable weight, size, whatever you want and I put it back so, but this is my sampling unit number 1. Now I want to take out a sample of size suppose 3.

So, I try to shake it again and again, once again I try to take out one more unit out of the population and you can see again it comes out to be a walnut. I try to record its value and I put it in the jar and then again I try to shake it and I try to take out another unit, but this time again it is coming out to be a walnut.

Now, if you try to see I have drawn three units after adding walnut into my cashew and almonds, but every time the walnut is coming out as a sampling unit in my sample. So, what is really happening? Can you think about it? Because the size of the walnut is so, big in comparison to the cashew and almonds. So, what is happening that whenever I try to put my hand inside my bottle, without any intention I am telling you, the walnut is coming into my picture into my hand.

So, this means the size of the sampling unit the size does not mean the volume or the surface area or the weight, the size of the unit at least in this case the size of the unit is the surface area or the volume that is playing an playing a major role in choosing the units and those sampling unit who have a bigger size they are getting more chance of being collected.

So, now, if these bigger units come into my sample, they are going to disturb the homogeneity and if I try to find out say this sample mean, the sample mean will shoot up to some other value.

So, the only thing what you have to observe here that the size of the sampling unit is going to make an impact on the sampling scheme and if the sizes are not homogeneous, sizes are not similar, then we have to do something so, that after getting the sample whatever parameters I want to estimate they are estimated properly.

So, actually this type of sampling in which the size of the sampling unit is going to affect the choice of the sampling unit they are called as Probability Proportional to Size sampling scheme, PPS sampling scheme and they come under the broader purview of varying probability scheme.

So, what happened that in the case of simple random sampling you assume that all the units have got equal chance of being collected, but in the case of varying probability scheme or particularly in the PPS sampling what we assume that the size of the unit is proportional to the probability of its selection, right.

So, now this gives you an idea that how are you going to choose an appropriate sampling scheme in real life. In case if the homogeneity in the sampling units; that means, the units are not varying much with respect to the character under study to use simple random sampling. For example, when my variable was dry fruit, then I can choose any of the this almonds or cashew, but when I have a bigger sampling unit like almond in my population, then it is not really appropriate to go for the simple random sampling.

So, for example, if you try to see the sample which I have just drawn, if you try to look it is showing you only the walnuts. So, once you try to look into your sample and there are only walnuts, it will show that the population consists of only the walnuts which are of bigger size and there are no units which are of smaller size like say almonds or cashews.

So, that may not really give you a representative sample. So, in this situation when our variable is not homogeneous or the characteristic that we want to study is not really homogeneous with the population, then we try to choose another type of sampling scheme.

Well, giving the details of varying probability scheme or probability proportional to size it is difficult here and it is really the out of purview of this course, right, but, but surely I wanted to inform you about this thing that these concepts do exist and in case if we are not covering them in this course, it does not mean that they are not useful, right.

So, now, I have given you a fair idea that under what type of conditions you are going to use simple random sampling and definitely I will try to tell you here one thing beforehand that I took a population which has two different components say cashew and almonds. So, it is also possible that you can divide the entire population into two parts one for cashew and one for almonds and yes, then my variable may become say size of the of cashew and almond.

So, you can see that even if I try to take all cashew or all almond, they are not of exactly the same size, there will be some variation in their sizes and then so, now, what you have done? You have a population of cashew and walnuts and then you are trying to divide them into two population and then you can draw sample from both the population separately.

So, this is actually the concept of stratified sampling or also if you want to add here walnuts also then you can divide the entire population into three groups one for walnut, one for almonds and one for cashew. So, the advantage is this now these groups become within homogeneous with respect to the characteristic understudy.

Now, you can draw sample from these three population separately. So, that is actually the concept of stratified random sampling and similarly we have many types of sampling scheme and depending on the experimental condition real life condition we try to choose appropriate sampling scheme, right.

So, now we stop in this lecture and in this lecture, I have given you a demonstration that that what are the common mistakes which people do make when they are trying to draw a simple random sample and I have shown you, I have given you a demonstration. So, this was pretty simple lecture, very easy to understand.

So, now, in the next lecture, I will try to continue with some more concepts of simple random sampling. So, you please try to revise this concept, try to see; what are the other

possibilities where one can make mistakes and make sure that you do not at least make those mistakes. So, you practice and I will see you in the next lecture once again. Till then; good bye.