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 – 15
Simple Random Sampling
Probabilities of Selection of Samples

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 to learn with the basic fundamentals of Simple Random Sampling.

So, you may recall that in the earlier lecture, we had considered how to compute the probability of selection of a unit. And that was 1/N that is the 1 upon total number of units in the population. And an important point to be noticed was that that either you are using SRSWR or SRSWOR the probability of selection of the unit remains the same, 1/N and that was supplemented with the mathematical proof also, so that there should not be any confusion.

Now, in this lecture, I am going to compute the probability of selection of a sample. What is the difference between a sampling unit and a sample? Sample unit means one unit. And sample mean this is a collection of n number of sampling units.

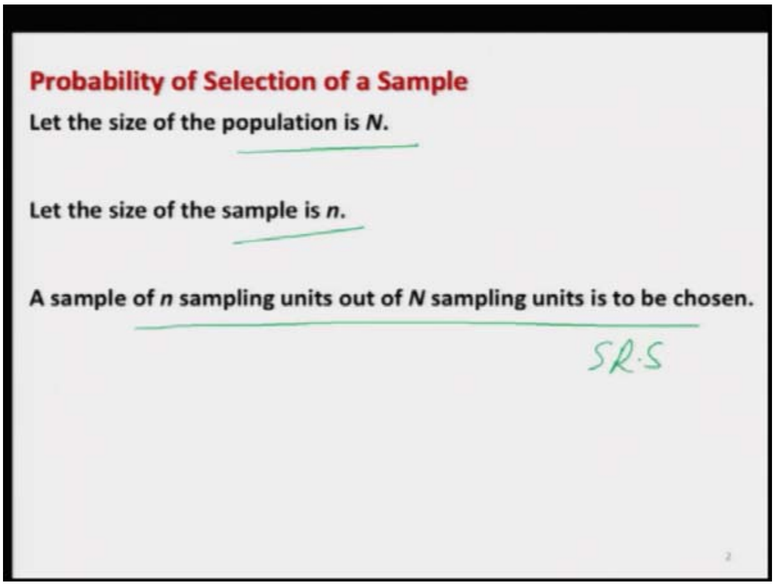
Well, I will try to make here one thing clear that when we are trying to draw a n number of units from a population of size N number of units, then I have two options. First option is this I try to draw one unit at a time means I draw – record the value, then I draw – record the value and so on.

And second option is I try to draw all the n number of units in a single shot. For example, you must have noticed people catching the fishes in the pond or in the river. So, there are two ways. There are some people who will hold a stick something like this, they will put it inside in the river or in the pond, and there is some thread with some food for the fish, and they will catch one fish at a time.

And second option is this sometimes you have noticed that people put a sort of net on the river. And after some time they will take out the net. And then you will see that inside that net there are many, many fishes. So, that means, in the case of net they are trying to catch more than one fish at the same time. But when they are trying to use that stick, then they are trying to catch only one fish at a time.

So, the probabilities of selection of units when they are trying to collect a small n number of fishes one by one, and when they are trying to collect n number of fishes in a single shot, they are computed by different ways. For the second case, when they are trying to use a net, then the probabilities are computed using the hypergeometric distribution which we are not considering here, but because we are considering here the simple random sampling. This is what you have to keep in mind, ok.

(Refer Slide Time: 03:49)



Probability of Selection of a Sample

Let the size of the population is *N*.

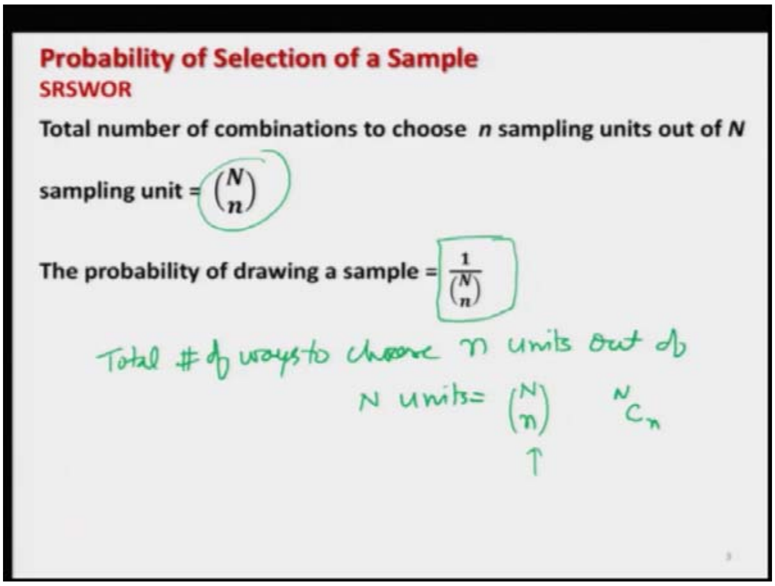
Let the size of the sample is *n*.

A sample of *n* sampling units out of *N* sampling units is to be chosen.

SR·S

So, now, we consider our setup. Suppose, the population size here is capital N, and we want to draw a sample of size n. So, n number of sampling units out of N sampling unit is to be drawn by SRS ok.

(Refer Slide Time: 04:04)



Probability of Selection of a Sample

SRSWOR

Total number of combinations to choose *n* sampling units out of *N* sampling unit = $\binom{N}{n}$

The probability of drawing a sample = $\boxed{\frac{1}{\binom{N}{n}}}$

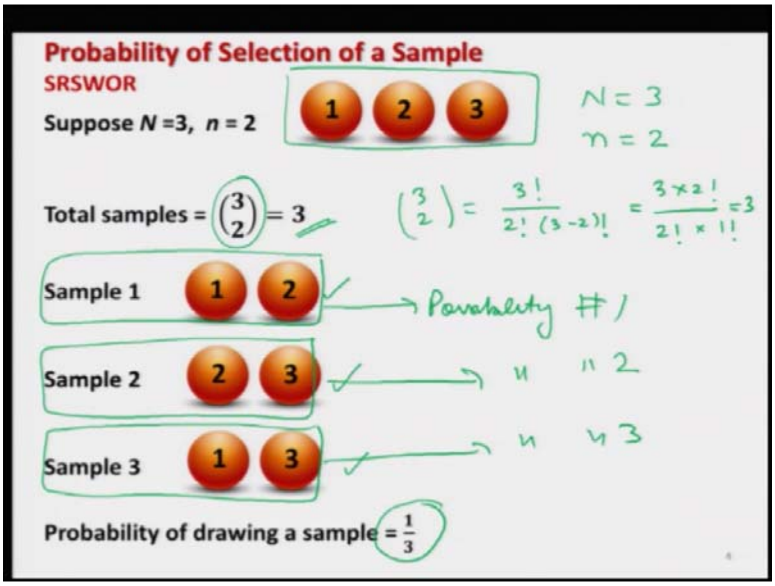
Total # of ways to choose n units out of N units = $\binom{N}{n}$ $\quad {}^{N}C_{n}$

$\uparrow$

So, now I have here two options; WOR – without replacement, and WR – with replacement. So, first I consider without replacement case. So, you know from the theory of combinatorics that the total number of ways to choose n units out of N units is equal to ${}^{N}C_{n}$ or say $\binom{N}{n}$, sometime in the olden style this is also written as ${}^{N}C_{n}$.

But in the, but nowadays, this symbol is more popular $\binom{N}{n}$, ok. So, the total number of combinations of n number of sampling units which can be drawn from a collection of N unit is $\binom{N}{n}$. So, obviously, the probability of drawing one unit that is one sample that is the collection of n number of units will be 1 upon total number of possible cases $1/\binom{N}{n}$, ok.

(Refer Slide Time: 05:23)



Probability of Selection of a Sample

SRSWOR

Suppose N =3, n = 2

$N = 3$

$n = 2$

Total samples = $\binom{3}{2} = 3$

$\binom{3}{2} = \frac{3!}{2!\,(3-2)!} = \frac{3 \times 2!}{2! \times 1!} = 3$

Sample 1 → Possibility #1

Sample 2 → " " 2

Sample 3 → " " 3

Probability of drawing a sample = $\frac{1}{3}$

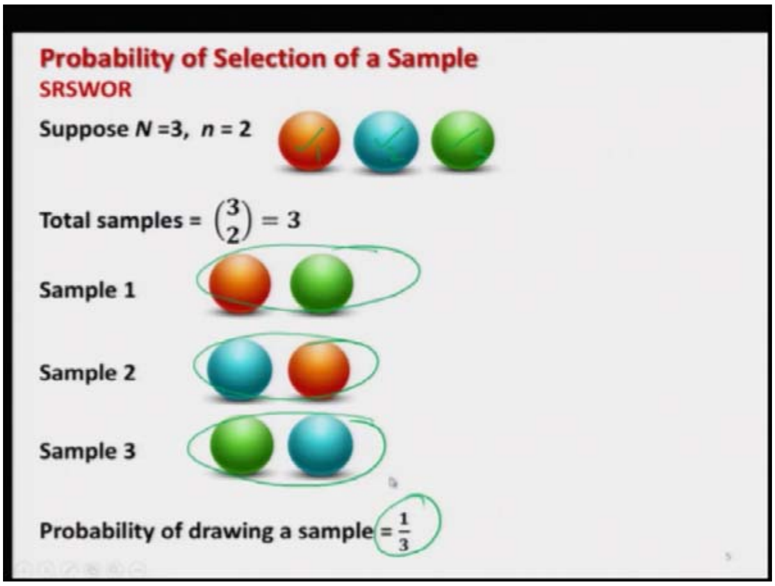
So, now what does this mean? Suppose, an example, because the hypothetical example, I try to consider here a population of size 3. I have got here 3 red balls which are numbered as 1, 2 and 3. Now, in this case, N will become 3. And suppose I want to draw here a sample of size n.

So, n will be equal to 2. So, now, the total number of ways to draw 2 balls from the population of size 3 is $\binom{3}{2}$ which is equal to here $\binom{3}{2}$ is equal to here 3! 2! and 3! - 2 which is equal to here 3 into 2!/2!1!. So, this becomes here 3, right. So, this is here 3.

Now, what are those samples? There are 3 possibilities. I can choose ball number 1 and 2, I can choose ball number 2 and 3, and I can choose ball number 1 and 3. So, this will constitute my one sample; this is another sample of ball 2, and 3; and this is another sample of balls 1 and 3.

And out of this is possibility number 1, possibility number 1; this sample is the possibility number 2, and this is the possibility number 3 and that is why the probability of drawing any of this sample 1, 2 or 3, this is 1/3. So, this is what I meant when I say that probability of choosing a sample, ok.

(Refer Slide Time: 07:11)



Probability of Selection of a Sample

SRSWOR

Suppose $N = 3$, $n = 2$

Total samples = $\binom{3}{2} = 3$

Sample 1

Sample 2

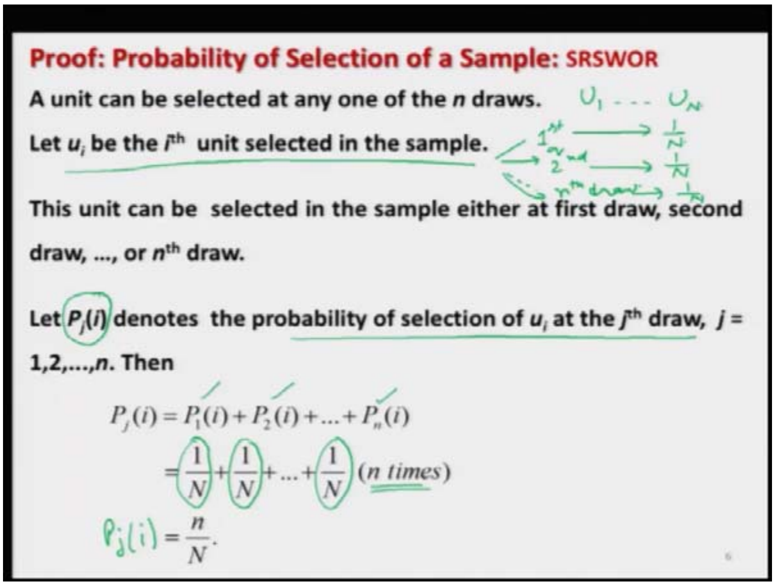
Sample 3

Probability of drawing a sample = $\frac{1}{3}$

Similarly, if I try to take here the similar example with a different color ball, suppose I have got here 3 balls of red color, blue color, green color, and yeah means since there is only one ball of each color. So, there is no numbering required, but if you wish you can do it, right; otherwise, the color of the ball itself is creating the sampling frame.

So, I try to choose here 3 possible samples; and those samples can be red ball, green ball; blue ball, red ball; and green ball, blue ball. So, there are 3 samples. So, the probability of choosing one sample out of 3 sample is simply 1/3, ok.

(Refer Slide Time: 07:57)



Proof: Probability of Selection of a Sample: SRSWOR

A unit can be selected at any one of the n draws. $U_1 \; - - - \; U_N$

Let u_i be the i^{th} unit selected in the sample.

$1^{st} \longrightarrow \frac{1}{N}$

$2^{nd} \longrightarrow \frac{1}{N}$

n^{th} draw $\longrightarrow \frac{1}{N}$

This unit can be selected in the sample either at first draw, second draw, ..., or n^{th} draw.

Let $P_j(i)$ denotes the probability of selection of u_i at the j^{th} draw, $j = 1,2,\ldots,n$. Then

$$P_j(i) = P_1(i) + P_2(i) + \ldots + P_n(i)$$

$$= \frac{1}{N} + \frac{1}{N} + \ldots + \frac{1}{N} \; (n\ times)$$

$$P_j(i) = \frac{n}{N}.$$

Now, I try to give you here first here a proof that how this is happening. So, that will convince you also more. So, first I try to go by SRSWOR. In the case of SRSWR, it is straightforward that you will see later on. So, now I have here population of size n. And whatever are the units they are denoted by here U_1, U_2, capital U_N, right. So, units are simply the units and there is no random variable which is attached to these units, ok. So, now, we denote by u_i as the ith unit which is to be selected in the sample.

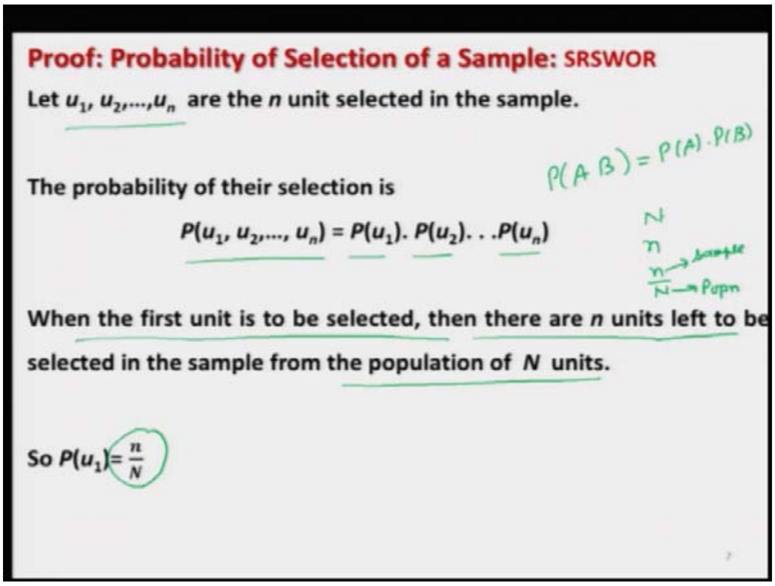
So, I am saying that ok, I want to select this u_i in my sample. Now, what are the different possibilities? This particular unit u_i can be selected in my sample in the first draw or in the second draw or in the nth draw. And we have seen that the probability of selecting the u_i in the first draw is 1 upon the population size, 1/N; probability of drawing u_i in the second draw is again the same, because this is SRS; and probability of drawing u_i in the nth draw is also same 1/N.

So, now I have a n number of possibilities in which I can draw the u_i in any of the given draws from 1 to n. Now, based on that I have to compute the probability of selection u_i at any given draw say jth. So, I try to compute here or first we try to denote let this quantity $P_j(i)$ in the subscript, and i in the parenthesis. Suppose this probability is indicating the probability of selection of the ith unit at the jth draw, right.

So, now we have seen that the ith unit can be selected at the jth draw, where j is equal to 1/N, this probability is that the unit is selected in the first draw, in the second draw, or in the nth draw.

So, this probability here is 1/N probability selection of u_i in the first draw or in the second draw, it is 1/N, and in the nth draw also this is 1/N. So, these are n number of terms. So, when this sum, the $P_j(i)$ will become the probability of selection of u_i that is the ith unit at the jth draw as N/n, right.

(Refer Slide Time: 10:58)



Proof: Probability of Selection of a Sample: SRSWOR

Let $u_1, u_2, \ldots, u_n$ are the n unit selected in the sample.

The probability of their selection is

$$P(u_1, u_2, \ldots, u_n) = P(u_1).\, P(u_2). \ldots .P(u_n)$$

$P(AB) = P(A) \cdot P(B)$

N

n

$\frac{n \rightarrow \text{sample}}{N \rightarrow \text{Popn}}$

When the first unit is to be selected, then there are n units left to be selected in the sample from the population of N units.

So $P(u_1) = \frac{n}{N}$

So, now, if I try to extend it, suppose there are a n number of units u_1, u_2, …, u_n, which I want to select in my sample. So, since we are considering the simple random sampling, such that all the units are independently drawn, such that the probability of selection of any of the unit at any stage is 1/N, so the probability of selection of u_1, u_2, …, u_n will be governed by the probability laws for the independence.

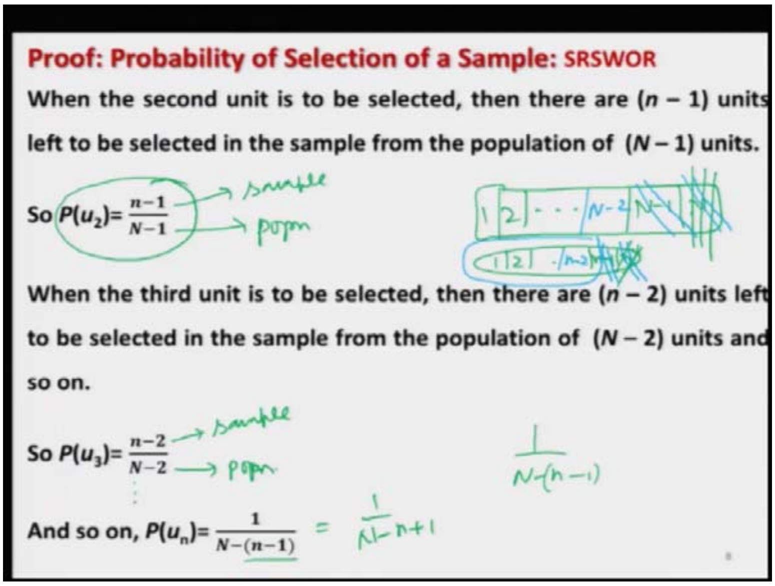
The rule is that if I have got here 2 events A and B, then the probability of the joint occurrence of A and B, if A and B are independent, then this is given by the product of probability of occurrence of A and the probability of occurrence of B.

So, extending this rule to selection of u_1, u_2, …, u_n, I can write down this probability of u 1 into probability of u 2 into probability of u n. So, now, I try to compute here probability of u_1, u_2, …, u_n. And for that, we already have seen that if there are a n number of units in the population.

And if I want to select n number of units in the sample, then the probability of selection of any units say ith unit in the jth draw is n/N. So, remember one thing this is the sample size, and this is here the population size. So, this probability is n/N.

So, now I can extend this rule. How? So, first we try to observe when the first unit is to be selected then there are n number of units left to be selected, which I want to selected out of the population of size capital N, right. So, the probability becomes here n/N in the probability of selection of the first unit.

(Refer Slide Time: 13:01)



Proof: Probability of Selection of a Sample: SRSWOR

When the second unit is to be selected, then there are ($n-1$) units left to be selected in the sample from the population of ($N-1$) units.

So $P(u_2)=\frac{n-1}{N-1}$ $\rightarrow$ sample $\rightarrow$ popn

When the third unit is to be selected, then there are ($n-2$) units left to be selected in the sample from the population of ($N-2$) units and so on.

So $P(u_3)=\frac{n-2}{N-2}$ $\rightarrow$ sample $\rightarrow$ popn

$\frac{1}{N-(n-1)}$

And so on, $P(u_n)=\frac{1}{N-\underline{(n-1)}} = \frac{1}{N-n+1}$

Now, when I come to the probability of the second unit, what is happening? Suppose, this was my population 1, 2 up to here say here N, and here N - 1. Now, I already have selected my one unit. So, now, the total population size remaining is here N - 1. And from the sample size also I wanted to select unit 1, 2 up to here n, right.

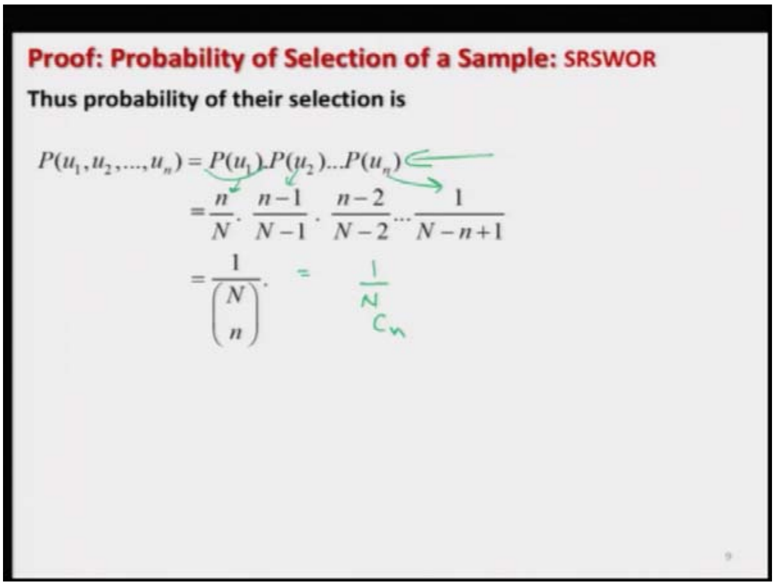
So, one unit I already have selected. So, now, I have to select n - 1 units out of N - 1 units. So, when I am going to select my second unit, the second unit can be selected in the first draw, second draw and so on. And using the same rule, I can write down that the probability of selection of the second unit is n - 1 / N - 1 which is again the same rule the sample size what we want and the population size in the denominator.

And similarly if I want to select the third unit, that means, we already have selected the two units. So, now, in this picture I can say here suppose I have here n - 2, N - 2. So, I already have selected the last two units. So, I have remaining say 1 to n - 2 number of units in the population. And all and I already have selected 1 and 2 sample units in my sample. So, the remaining sampling units are n - 2.

And when I am trying to compute this probability, so the probability of selection of third unit will be the sample size available to us divided by the population size available to us. So, this is n - 2 / N - 2. And if I try to extend by logic this rule, then the probability of selection of the nth unit that is the last unit will be 1/N – (n – 1), or this is equal to 1/N - n + 1.

Yeah, one has to be careful when I am trying to speak right means if I speak here 1/N - n - 1, this will look like this whereas, I wanted to say N - n - 1. So, just be careful ok.

(Refer Slide Time: 15:30)



Proof: Probability of Selection of a Sample: SRSWOR

Thus probability of their selection is

$$P(u_1, u_2, \ldots, u_n) = P(u_1).P(u_2)\ldots P(u_n) \leftarrow$$

$$= \frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \frac{n-2}{N-2} \cdots \frac{1}{N-n+1}$$

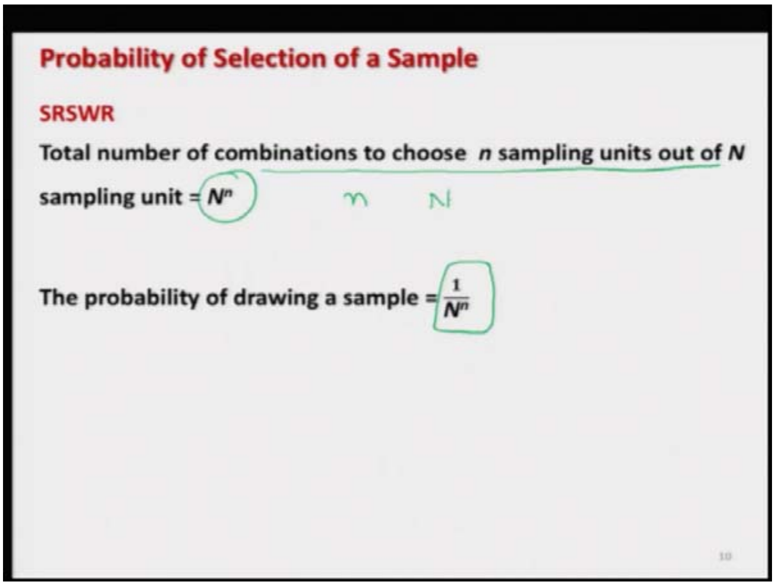
$$= \frac{1}{\binom{N}{n}} \cdot = \frac{1}{{}^{N}C_n}$$

So, now I have selected I have computed all the probabilities probability of $P(u_1)$, $P(u_2)$, …, $P(u_n)$ So, I can substitute them in this expression. And this probability of u_1, we have computed as a n / N; probability of u_2 has been computed as n - 1/N - 1 and so on.

And probability of u n is 1/N - n + 1. So, if you try to just simplify it, this will come out to be $1/{}^{N}C_n$, or 1 / $\binom{N}{n}$. So, this is how you can compute the probability of selection of a sample.

So, you can see now I have given you the mathematical proof. So, you are also happy. I am also confident, you are also confident that whatever we have done this is correct. That is what I will repeat again that even in the data science if your results what you are going to use are not supported by the statistical proof, mathematical proof, statistical logic, mathematical logic, nobody is going to believe on them ok.

(Refer Slide Time: 16:35)



Probability of Selection of a Sample

SRSWR

Total number of combinations to choose *n* sampling units out of *N* sampling unit = N^n n N

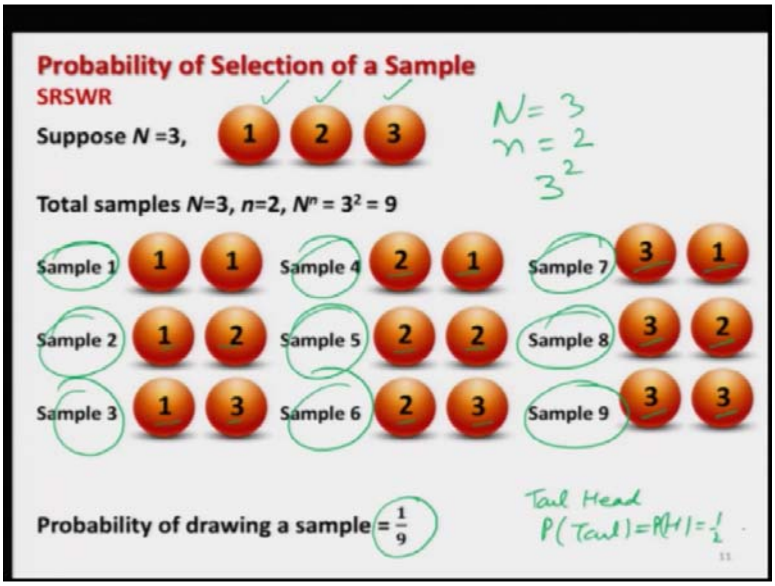
The probability of drawing a sample = $\frac{1}{N^n}$

10

Now, I come to simple random sampling with replacement, and I try to find out the probability of selection of a sample. So, in SRSWR the things are pretty straight forward. We know that when I want to select a n number of units out of say population of N number of units, and at every stage the population size is going to remain the same.

So, the total number of combination to choose a n sampling units out of N sampling units will be N^n. So, obviously, all the samples are equally probable. So, choosing one of the sample out of N^n probable points it is simply $1/N^n$. So, this is the probability of drawing a sample in case of SRSWR.

(Refer Slide Time: 17:35)



Probability of Selection of a Sample

SRSWR

Suppose N =3, 1 2 3

$N = 3$
$n = 2$
3^2

Total samples N=3, n=2, $N^n = 3^2 = 9$

Sample 1: 1 1
Sample 2: 1 2
Sample 3: 1 3
Sample 4: 2 1
Sample 5: 2 2
Sample 6: 2 3
Sample 7: 3 1
Sample 8: 3 2
Sample 9: 3 3

Probability of drawing a sample = $\frac{1}{9}$

Tail Head
$P(\text{Tail}) = P(H) = \frac{1}{2}$.

11

Now, let me take the same example and I try to illustrate it that what does this mean. So, now, once again I try to take the same example, I have here 3 balls – ball number 1, ball number 2 and ball number 3. And I want to draw here a sample of size say 2. So, in this case, N is equal to 3, n is equal to 2. So, there are all together 3^2 say number of total samples.

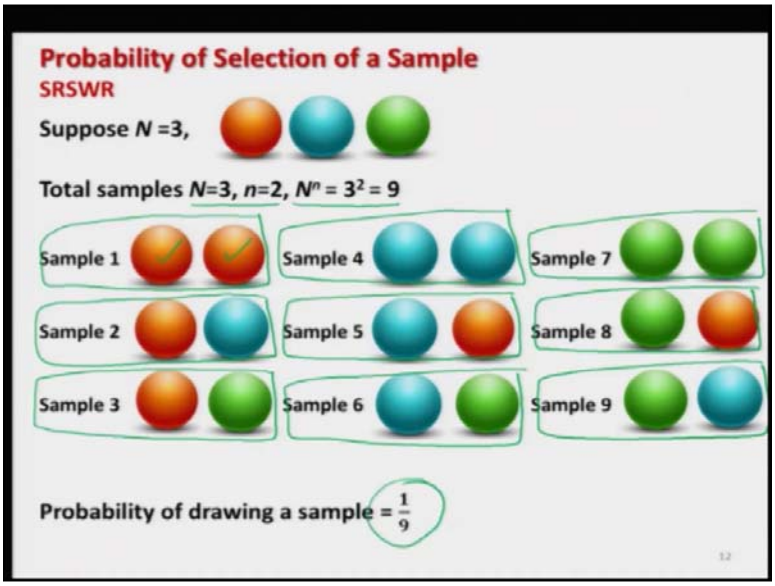
Now, let me try to create those 9 samples, and see what are those samples. So, one option is this, I can choose ball number 1 and 1, and so that can be in my sample. Next option is this; I can choose ball number 1 and 2, so that will create my sample number 2.

Similarly, the third possibilities I can create ball number 1 and 3 to create my third sample. The 4th sample can have ball number 2, ball number 1. 5th sample can have ball number 2 and 2 repeated. Sample number 6 may have ball number 2 and ball number 3. Sample, sample number 7 may have ball number 3 and ball number 1. Sample 8 may have ball number 3 and ball number 2; sample 9 may have ball number 3 and ball number 3 repeated once again.

So, you can see here, I have here sample 1, 2, 3, 4, 5, 6, 7, 8, 9 – 9 samples. And I have to choose one sample out of this 9 sample. So, the probability of selecting one sample out of 9 sample when every sample is equally probable that when the probability of selection of every unit is the same, it is 1/9, right.

This rule is something like if you toss a coin, the coin have 2 possibilities; one is here tail, and another is a head. So, so there are only two possibilities. So, the probability of getting a tail or say probability of getting a head, this is 1 by 2 that is the simple rule of probability which I have used here.

(Refer Slide Time: 19:33)



Probability of Selection of a Sample

SRSWR

Suppose N =3,

Total samples N=3, n=2, $N^n = 3^2 = 9$

Sample 1 Sample 4 Sample 7

Sample 2 Sample 5 Sample 8

Sample 3 Sample 6 Sample 9

Probability of drawing a sample = $\frac{1}{9}$

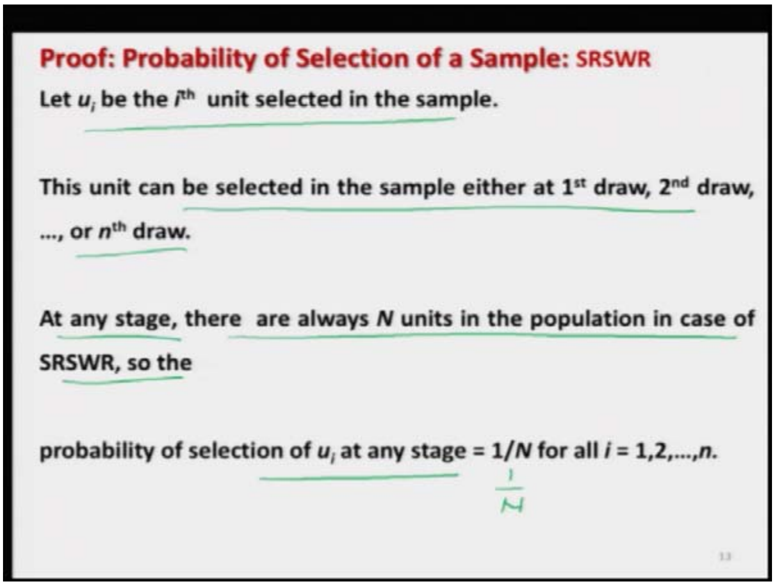
12

Now, similarly if I try to take colored balls as in the earlier case, so once again I can create 9 samples. So, once again my population size here is 3, and I want to draw here a sample of size 2. So, the total number of samples in the case of SRSWR will be 3 square which is equal to 9. So, I can create here 9 possible sample which is here red ball-red ball that will create my sample number 1.

Red ball - blue ball which will create my sample number 2; red ball - green ball which will create my sample number 3; blue ball blue ball both repeated – this is my sample number 4; blue ball - red ball – sample number 5; blue ball - green ball – sample number 6; green ball repeated, so 2 green balls, sample number 7. One green ball, one red ball, sample number 8. And lastly one green ball, one blue ball that will constitute sample number 9.

So, now, I simply have to compute the probability of selection of any of the samples. So, there are 9 possible samples which are equally probable. So, the probability of drawing a sample will simply be 1 upon 9.

(Refer Slide Time: 20:53)



Proof: Probability of Selection of a Sample: SRSWR

Let u_i be the i^{th} unit selected in the sample.

This unit can be selected in the sample either at 1st draw, 2nd draw, ..., or n^{th} draw.

At any stage, there are always N units in the population in case of SRSWR, so the

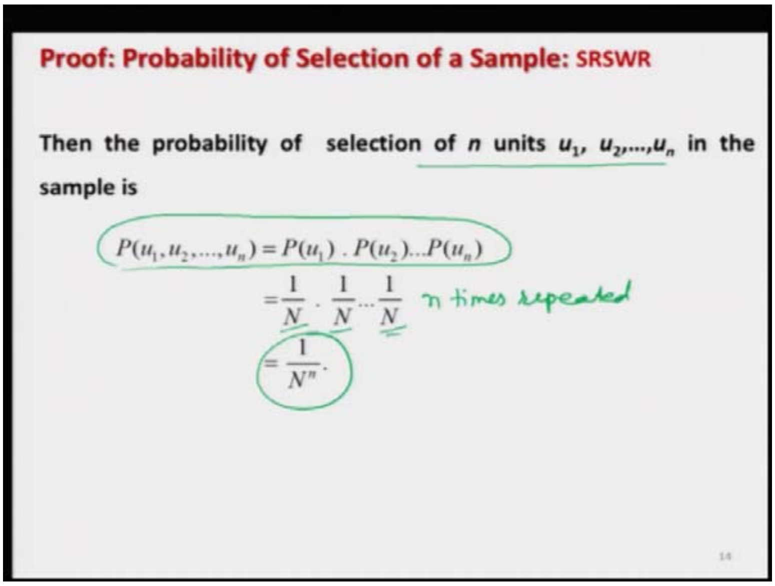
probability of selection of u_i at any stage = $1/N$ for all $i = 1,2,\ldots,n$.

$\frac{1}{N}$

And now in case if you try to have the proof of this probability on the similar lines as we did in the case of SRSWOR, I can briefly give you here the proof. So, let u_i be the ith unit which is to be selected in the sample. Now, this ith unit can be drawn in the first draw second draw or in the nth draw that we already have discussed exactly on the same line.

So, and one difference between SRSWR and WOR is that in case of SRSWR at any stage there will always be N number of units in the population, so that means, the probability of selection of u_i at any stage will remain as 1/N that is now different from the case of SRSWOR, right. So, now that is why the computational probability in case of SRSWR become more simpler.

(Refer Slide Time: 21:52)



Proof: Probability of Selection of a Sample: SRSWR

Then the probability of selection of n units $\underline{u_1, u_2, \ldots, u_n}$ in the sample is

$$\boxed{P(u_1, u_2, \ldots, u_n) = P(u_1)\,.\,P(u_2) \ldots P(u_n)}$$

$$= \underline{\frac{1}{N}} \cdot \underline{\frac{1}{N}} \cdots \underline{\frac{1}{N}} \quad \text{n times repeated}$$

$$\boxed{= \frac{1}{N^n}}.$$

14

So, now, based on that, I can compute the probabilities of selection of n units in the sample which are my $u_1, u_2, \ldots, u_n$. So, the expression that we had discussed earlier was probability of $u_1, u_2, \ldots, u_n$. Since, they are obtained independently, so the joint probability can be expressed as the product of the individual probabilities. And individual probabilities here are 1/ N.

So, this will be 1/N into 1/N up to here 1/N, and this is going to be repeated n times. So, this is n times repeated. And thus the probability of collection of sample of size n that is the selection of units $u_1, u_2, \ldots, u_n$ is simply here $1/N^n$.

So, remember one thing please keep an eye on the slides and my language because both the symbols are here n. So, what I mean that 1/N that is the population size raise to the power of a n which is the sample size. So, the probability is one upon population size raised to the power of sample size. So, this is what you have to keep in mind.

Now, here is the time to stop in this lecture. So, this lecture was pretty simple straight forward. And I have explained you that when you are trying to draw the sample by SRSWR and WOR, then how the things are going to happen and how you are going to compute different probabilities.

So, I think this lecture and the earlier lecture will surely clarify your concept that what is the difference between the between the selection of a sampling unit and the selection of a sample, and also what is the difference between the probability of selection of a unit and the probability of selection of a sample. So, you try to revise this lecture, try to settle down these concepts in your mind, and I will see you in the next lecture.

Till then take care and good bye.